remote sensing

MDPI

Article

PCAN—Part-Based Context Attention Network for Thermal Power Plant Detection in Remote Sensing Imagery

Wenxin Yin [1,2,3], Wenhui Diao [1,2], Peijin Wang [1,2], Xin Gao [1,2], Ya Li [1,2] and Xian Sun [1,2,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; yinwenxin16@mails.ucas.ac.cn (W.Y.); whdiao@mail.ie.ac.cn (W.D.); wangpj@aircas.ac.cn (P.W.); gaxi@mail.ie.ac.cn (X.G.); liya@aircas.ac.cn (Y.L.)
[2] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: sunxian@mail.ie.ac.cn; Tel.: +86-10-58887208-8199

Abstract: The detection of Thermal Power Plants (TPPs) is a meaningful task for remote sensing image interpretation. It is a challenging task, because as facility objects TPPs are composed of various distinctive and irregular components. In this paper, we propose a novel end-to-end detection framework for TPPs based on deep convolutional neural networks. Specifically, based on the RetinaNet one-stage detector, a context attention multi-scale feature extraction network is proposed to fuse global spatial attention to strengthen the ability in representing irregular objects. In addition, we design a part-based attention module to adapt to TPPs containing distinctive components. Experiments show that the proposed method outperforms the state-of-the-art methods and can achieve 68.15% mean average precision.

Keywords: remote sensing; facility object detection; thermal power plants; convolution neural network; spatial attention; part-based attention

check for updates

Citation: Yin, W.; Diao, W.; Wang, P.; Gao, X.; Li, Y.; Sun, X. PCAN—Part-Based Context Attention Network for Thermal Power Plant Detection in Remote Sensing Imagery. *Remote Sens.* **2021**, *13*, 1243. https://doi.org/10.3390/rs13071243

Academic Editor: Mohammad Awrangjeb

Received: 4 February 2021
Accepted: 17 March 2021
Published: 24 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fixed industrial facilities are buildings with pieces of equipment for a particular purpose. Specifically, power plants supply electricity to the electrical grid, sewage treatment plants remove contaminants from municipal wastewater and garbage dumps are piled with domestic garbage. These facilities greatly influence regional economic situation and ecological environment. Therefore, monitoring the location of fixed industrial facilities is of great significance for regional economic and environmental situation.

Thermal power plants of optical remote sensing images are investigated in this paper. Current research of spectral image object detection [1–4] mostly focuses on the land cover type such as urban land, agriculture land, forest land and water. Such objects are different from thermal power plants, because thermal power plants are functional fixed facilities, which have diverse spatial patterns with blurred boundaries and contain several non-rigid components with separate locations.

Compared with other facilities, it is more challenging to detect thermal power plants in remote sensing images due to the following characteristics. Thermal power plants generally have typical components including sedimentation tanks, cooling towers, chimneys, coal yards and pools. As shown in Figure 1, unlike sewage treatment plants, the components of Thermal Power Plants (TPPs) are non-rigid irregular objects, such as coal yards and pools, which are difficult to describe with a specific shape and scale. In addition, different from garbage dumps, TPPs have diverse spatial patterns with blurred boundaries, containing several components with separate locations, as illustrated in Figure 2. Consequently, it is more difficult but more valuable to study the detection of TPPs compared with other facility objects.

(**a**) (**b**) (**c**)

Figure 1. Facility objects which composed of several separate components. (**a**) sewage treatment plants including sedimentation tanks (red bounding boxes) and arcuation sedimentation tanks (blue bounding boxes); (**b**) thermal power plants, including chimneys (green bounding boxes), coal yards (red bounding boxes) and pools (blue bounding boxes); (**c**) garbage dumps.

(**a**) (**b**) (**c**)

Figure 2. Samples of thermal power plants with diverse separate irregular components including chimneys (green bounding boxes), coal yards (red bounding boxes), pools (blue bounding boxes) and other processing buildings. (**a**) Bathinda thermal power plant; (**b**) Ukai thermal power plant; (**c**) Korba power plant.

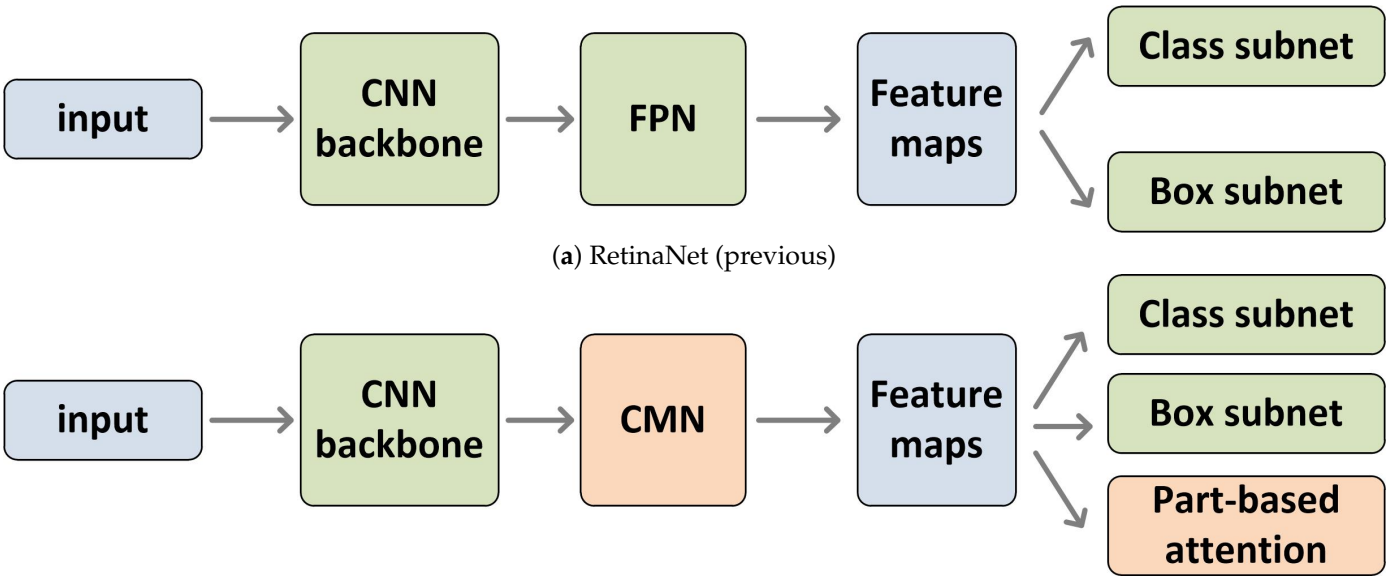
In view of above characteristics, many recent works have already focused on the detection of irregular objects, as well as objects with diverse spatial patterns.

Detection of irregular objects: Zhou et al. [5] construct a fully-convolutional neural network adapted for text detection to predict words of arbitrary orientations and quadrilateral shapes in full images. Wang et al. [6] propose a Progressive Scale Expansion Network (PSENet) to detect text instances with arbitrary shapes, which generates the different scale of kernels for each text instance and gradually expands the minimal scale kernel to the text instance with the complete shape. They propose another arbitrary-shaped text detector, termed Pixel Aggregation Network (PAN) [7] by means of cascadable U-shaped module and feature fusion. However, such arbitrary-shaped texts are irregular but organized objects with clear boundaries rather than TPPs. As a typical method for irregular objects, Deformable Convolutional Networks (DCN) [8] introduce deformable convolution and deformable RoI pooling to enhance the transformation modeling capacity of CNNs. The deformable convolution adds 2D offsets to receptive fields in the standard convolution, which can deform the receptive fields. In DCN, the shape and scale of anchor boxes is predefined, so it is difficult for the generic detector to describe TPPs with a specific shape and scale without adaption.

Detection of objects with diverse spatial patterns: Li et al. [9] divide a pedestrian image into several horizontal stripes for patch matching. Zhao et al. [10] propose Spindle Net for person re-identification, which separately captures semantic features from different body regions for the alignment of macro- and micro-body features. However, such part-based methods for person detection are based on the specific pattern of humans. Han et al. [11] propose a Part-based Convolutional Neural Network (P-CNN) for fine-grained visual categorization. P-CNN contains a part localization network, which learns a bank of

convolutional filters as discriminative part detectors to locate distinctive object parts, and a part classification network, which classify each individual object part as image-level categories and then fuses part features and global feature for the final classification. Although P-CNN has taken discriminative parts into consideration, it is not applicable for TPPs because P-CNN is designed for rigid objects such as aircrafts and cars.

According to related research above, existing methods are mostly designed for rigid objects and organized objects with a specific pattern. These methods have not considered objects like TPPs, which are composed of non-rigid irregular components with separate spatial locations. In order to tackle the above problems, this paper presents an end-to-end detection framework called Part-based Context Attention Networks (PCAN). As illustrated in Figure 3, PCAN is based on a one-stage detector RetinaNet [12] on ResNets [13], using a context attention multi-scale feature extraction network (CMN) with deformable convolution [8] and a part-based attention module for both classification and regression. CMN not only obtains geometric constraint information by deformable convolution, but also enhances the context attention multi-scale feature maps. Part-based attention module is designed for the adaption of the thermal power plants with sparsely distinctive components, which introduce a loss function to focus on certain discriminative regions with high responses. Compared to other generic object detection methods such as RetinaNet, Faster RCNN and Cascade RCNN, our framework is more suitable for the detection of thermal power plants, and has achieved state-of-the-art performance.



(**a**) RetinaNet (previous)

(**b**) PCAN (ours)

Figure 3. The illustration of the detection pipeline. (**a**) RetinaNet [12], a one-stage detection network, extracts deep features by ResNet [13] and Feature Pyramid Networks (FPN) [14] , and then obtains locations and class labels of the anchors by box subnet and class subnet using focal loss. (**b**) part-based context attention networks (PCAN) uses a Context attention Multi-scale feature extraction Network (CMN) to generate multi-scale feature maps containing contextual information for irregular objects and a part-based attention module for the adaption of facility objects composed of distinctive components.

The main contributions of this paper are summarized as follows:

(1) We construct a one-stage end-to-end detection framework called Part-based Context Attention Networks (PCAN). The model adaptively generates multi-scale feature maps containing context and part-based attention, which is more accurate and effective for thermal power plants detection in high-resolution remote sensing imagery.
(2) We propose a Context attention Multi-scale feature extraction Network (CMN) with deformable convolution, which strengthen the feature representations through the combination of context attention and multi-scale feature extraction.

(3) As facility objects generally consist of several components, a part-based attention module is designed for the adaption of such facility objects, which effectively help discover distinctive object components.

Experiments based on remote sensing images obtained from Google Earth show that our PCAN has state-of-art performance for the detection of thermal power plants. The datasets are publicly available in our github repository (https://github.com/wenxinYin/AIR-TPPDD, accessed on 14 March 2021), which can reduce the workload of thermal power plants investigation. The rest of this paper is organized as follows. Section 2 introduces the details of the proposed method. Then Section 3 presents the experiments conducted on a remote sensing dataset to validate the effectiveness of the proposed framework. Section 4 discusses the results of the proposed method. Finally, Section 5 concludes this paper.

2. Methods

2.1. Network Architecture

The proposed PCAN model is an end-to-end framework based on RetinaNet [12]. The overall architecture of PCAN in Figure 4 can be divided into three parts: a deep feature extraction sub-network to extract context-based feature maps for irregular objects, a sub-network for global prediction, and a module proposing a part-based loss function. The deep feature extraction sub-network contains a ResNet [13] backbone and a Context attention Multi-scale Network (CMN). The global prediction sub-network contains a classification subnet and a regression subnet for the bounding box prediction of global object. The part-based attention module sub-network is proposed to focus on discriminative regions with high responses in feature maps. In this sub-network, feature channels are clustered by K-means into certain groups, where a part-based loss function is introduced to highlight the prominent components in the object.

As shown in Figure 3 , in the simple one-stage RetinaNet, only backbone networks and global prediction networks are included. However, due to the non-rigid irregular components of TPPs, context attention multi-scale network has been added into the architecture of this paper to enhance the feature representation capability. In addition, part-based attention module is proposed for detecting thermal power plants with several separate components.

Deep Feature Extraction Sub-network: We use a ResNet-50 [13] architecture pre-trained on ImageNet and our CMN in backbone sub-network. The outputs of the last convolution layer in the last three residual blocks, defined as $\{C3, C4, C5\}$ are activated for feature extraction, whose sizes are $\{1/8, 1/16, 1/32\}$ corresponding to input image. In order to effectively detect multi-scale thermal power plants with irregular components, we design CMN to deal with the set of feature maps and produce global spatial attention features named $\{P3, P4, P5\}$. Deep feature extraction sub-network generates contextual attentioned deep feature maps of input images, which are designed for multi-scale TPPs with irregular components.

Global Prediction Sub-network: This sub-network includes a classification subnet and a regression subnet. These two parallel subnets share a common structure with separate parameters. Specifically, for A anchors and K object classes, the classification subnet predicts the probability of objects in spatial locations, which is a small FCN including three 3×3 conv layers attached to each pyramid level of CMN. Each 3×3 conv layer shares the same parameters, activated by ReLU. Then, the subnet is followed by a 3×3 conv layer with KA filters attached by a sigmoid activations. The difference between two subnets is that regression subnet finally obtains 4 linear outputs for each of the A anchors per spatial location rather than K. Global Prediction Sub-network uses focal loss [12] for classification subnet and smooth L_1 loss [15] for bounding box regression.

Part-based Attention Module: For the adaption of the TPPs with sparsely distributed components, part-based attention module sub-network is proposed to focus on discriminative regions with high responses in feature maps. In this sub-network, K-means method is adopted to cluster feature channels into certain groups, where each group aggregate

spatially-correlated patterns corresponding to each component of TPPs. Part-based loss functions for both classification and regression are proposed to strengthen the influence of prominent components in the object.

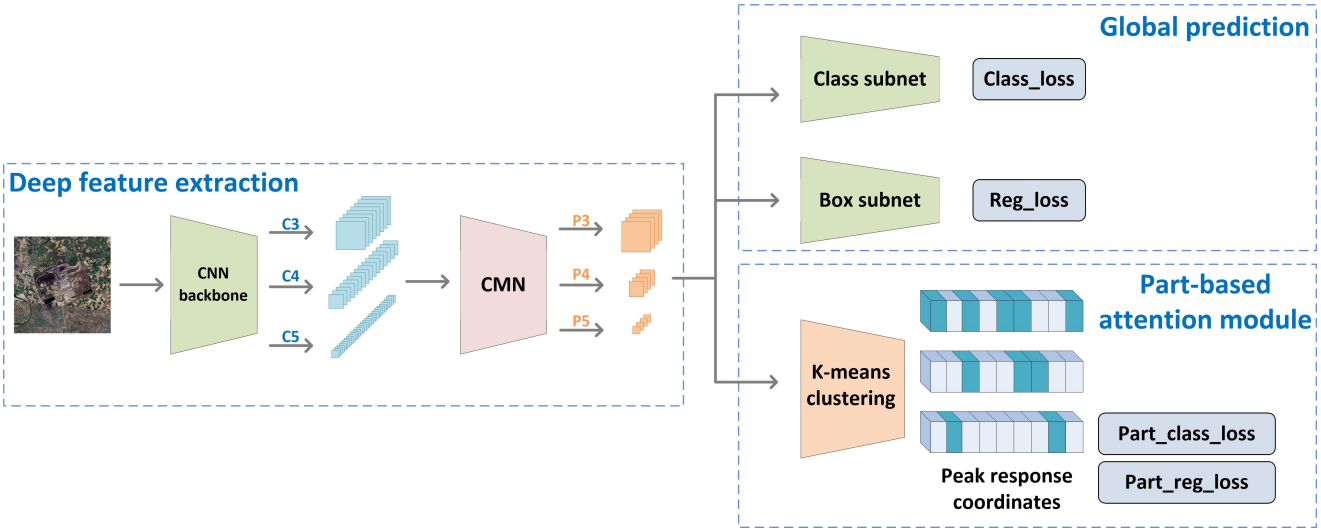


Figure 4. Overall framework of our PCAN, which consists of deep feature extraction sub-network, global prediction sub-network and part-based attention module. In deep feature extraction sub-network, our CMN after the classic Convolutional Neural Network (CNN) produces multi-scale feature maps, which can not only contain contextual information but also model irregular components. Global prediction sub-network includes two subnets, one for predicting the labels for anchors and one for regressing from anchors to ground-truth bounding boxes. Part-based attention module adopts K-means method to cluster feature channels into certain groups, where each group aggregate spatially-correlated patterns , corresponding to one component of Thermal Power Plants (TPPs).

2.2. Context Attention Multi-Scale Feature Extraction Network (CMN)

As previously described in Section 1, thermal power plants contains non-rigid irregular components which are difficult to describe with a specific shape and scale. In order to detect TPPs with irregular components, we design a Context attention Multi-scale feature extraction Network (CMN) with deformable convolution based on FPN [14]. FPN can merge low-level feature maps with higher resolution and high-level semantic information, which is suitable for multi-scale feature representation. To match the component objects in irregular shapes, we use deformable convolutions [8] to obtain geometric constraint information. In addition, global context attention in GCNet [16] is introduced in CMN to aggregate global contextual information for modelling capacity enhancement.

In Figure 4, we use outputs of the last convolution layer in residual blocks of ResNet [13] as $\{C_3, C_4, C_5\}$. Then CMN, which is elaborated in the following, produces unidimensional feature maps with geometric constraint and contextual attention. The resultant set of feature maps, called $\{M_3, M_4, M_5\}$ corresponding to $\{C_3, C_4, C_5\}$, is then laterally connected by up-samling and element-wise addition, generating feature maps for prediction as $\{P_3, P_4, P_5, P_6, P_7\}$. P_6 and P_7 are computed from C_5 as RetinaNet [12].

As shown in Figure 5, CMN is constructed by two modules including deformable convolution and context module (context attention and a transformer). For the input feature map $C_{i=5,4,3}$ in the shape of $\{batch, C_i, H_i, W_i\}$, the 1×1 convolutional layer is firstly used to reduce the dimension as $C_i^{'}$ of $\{batch, C_2, H_i, W_i\}$. In DCN part, in order to transform the receptive field of convolutional kernels, offsets for each point on feature map $C_i^{'}$ are learned by a 3×3 conv, denoted as a tensor of $\{batch, 18, H_i, W_i\}$. As the obtained offsets are usually fractional, offsets are then aggregated to original locations by bilinear interpolation, so as to generate the updated locations. Additionally, a 3×3 deformable conv with stride = 3 is applied to the updated locations, followed by ReLU activation and batch normalization, of which result is denoted as D_i.

In order to acquire global spatial contextual attention efficiently as GCNet [16], the following context part and transform part construct a non-local block to enhance D_i. In context part, the 1×1 conv with a softmax generates a global spatial attention mask which indicates the importance of each pixel in the image. The obtained attention mask is then multiplied to D_i, producing contextual features. In transform part, lightweight bottleneck layers integrate channel-wise dependencies and bottleneck ratio r is set to reduce the computational cost. Batch normalization (BN) can not only reduce the difficulty of optimization and also improve the generalization. The final step is to fuse the transformed contextual features, followed by sigmoid activation, with deformable feature maps D_i.

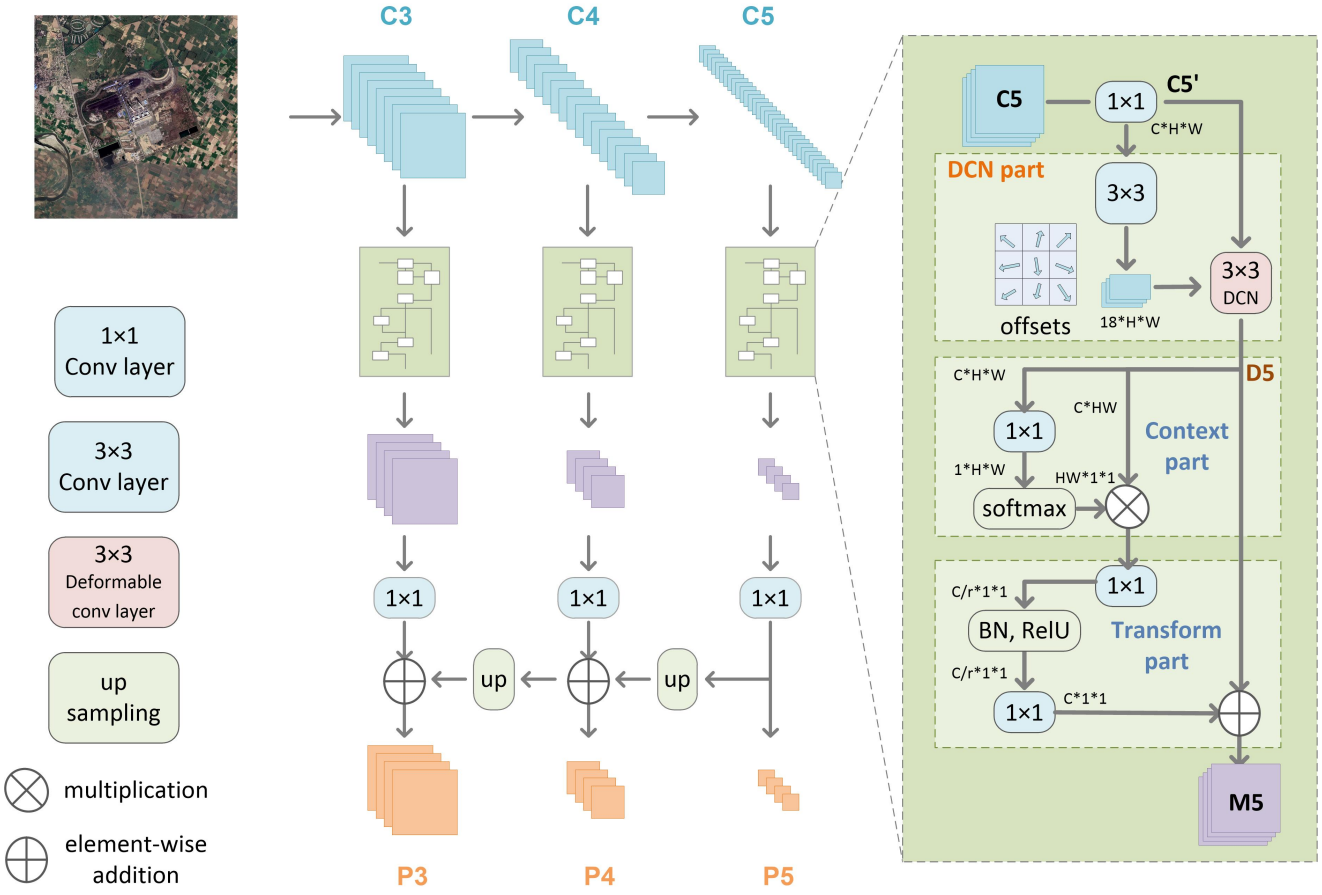


Figure 5. Architecture of our CMN. Based on FPN [14], CMN is constructed by two modules including deformable convolution and context module (context attention and a transformer).

2.3. Part-Based Attention Module

Thermal power plants contain distinctive components with separate locations as illustrated in Figure 2. This paper proposes a part-based loss function during training to strengthen the influence of distinctive components in TPPs. We introduce part-based loss function starting from the loss function in RetinaNet [12].

For an anchor box i, loss function is defined as the sum of classification loss L_{cls} and regression loss L_{reg}.

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

$p_i \in [0,1]$ is the estimated probability for the object class. $p_i^* \in \{0,1\}$ is the ground-truth label, that is, $p_i^* = 1$ for objects and $p_i^* = 0$ otherwise. t_i is a vector representing four parameterized coordinates of predicted anchor box and t_i^* is for the ground-truth box.

N_{cls} and N_{reg} are the numbers of anchor and anchor locations respectively in one batch. λ is used to balance L_{cls} and L_{reg}, which is set to 1 here.

The classification loss L_{cls} is the softmax loss of two classes, that is, object and background. L_{cls} in RetinaNet is the focal loss [12] for binary classification which is designed for class imbalance based on cross entropy (CE) loss [17] during training.

$$L_{cls} = FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t), \tag{2}$$

where $\alpha_t \in [0,1]$ is a weighting factor factor and $\gamma \in [0,5]$ is a tunable focusing parameter for smoothly adjustments of influence of easy examples. p_t is defined as:

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise, \end{cases} \tag{3}$$

where $p \in [0,1]$ is the predicted probability for the class with label $y = 1$ and $y \in \{\pm 1\}$ indicates the ground-truth class, object or background.

The regression loss L_{reg} in RetinaNet is the standard smooth L_1 [15] loss used for box regression. For an anchor box i,

$$L_{reg} = SmoothL_1(t) = \begin{cases} 0.5t^2 & if\ |t| < 1 \\ |t| - 0.5 & otherwise \end{cases} \tag{4}$$

where $t = t_i^* - t_i$.

Loss function in Equation (1) calculates the addition of mean values of both classification loss and regression loss for anchors at all scales of feature maps. However, the influence of distinctive components inside the objects is not taken into consideration. Different from other objects like garbage dumps, TPPs have diverse spatial patterns containing several components with separate locations. As a result of that, we design a part-based loss to strengthen the influence of prominent components in the object in the training stage, where the combined loss function is defined as the sum of L_{global} (Equation (1)) and L_{part} (Equation (6)), balanced by an adjustable parameter α.

$$L = L_{global}(\{p_i\}, \{t_i\}) + \alpha L_{part}(\{p_{i\_k}\}, \{t_{i\_k}\}). \tag{5}$$

In the part-based attention module, a set of multi-scale feature maps are clustered into certain groups by K-means clustering [18], where each group aggregate spatially-correlated patterns corresponding to each component of TPPs. For feature maps $\{P_3, P_4, P_5\}$, K (9, 6, 3) points are respectively extracted by K-means for each channel. Figure 6 visualizes the feature maps effected by part-based attention module, which reflects that the network can pay more attention to these distinctive components by adding part-based loss.

Similar to the loss function in RetinaNet in Equation (1), the part-based loss function is defined as follows.

$$L_{part}(\{p_{i\_k}\}, \{t_{i\_k}\}) = \frac{1}{N_{cls\_k}} \sum_k L_{part\_cls}(p_i, p_i^*) + \lambda_{part} \frac{1}{N_{reg\_k}} \sum_k p_i^* L_{reg}(t_i, t_i^*), \tag{6}$$

where λ_{part} is used to balance L_{part_cls} and L_{part_reg} which is set to 1 here.

In Equation (1), $\{p_i\}$ indicates the probability of object presence at each spatial position for each of the A anchors and N object classes, which can be seen as a set of vectors in the shape of $\{batch, NA, WH\}$. $\{t_i\}$ is a set of four relative offsets between the anchor and the ground-truth box for each of the A anchors per spatial location in the shape of $\{batch, 4A, WH\}$. Thus, part-based loss function in Equation (6) counts $\{p_i\}$ and $\{t_i\}$ at the clustering centers, as $\{p_{i\_k}\}$ and $\{t_{i\_k}\}$ in the shape of $\{batch, NA, K\}$. $\{t_i\}$ and $\{batch, 4A, K\}$ respectively.

For clustering, centers should be mostly positive samples as distinctive components of an object, α-balanced CE loss is used as L_{part_cls}, which can also be seen as $\gamma = 0$ in

focal loss (Equation (2)). Regression loss L_{reg} in part-based loss function (Equation (6)) is identical to smooth L_1 loss (Equation (4)).

$$L_{part\_cls} = CE(p_t) = -\alpha_t log(p_t). \quad (7)$$

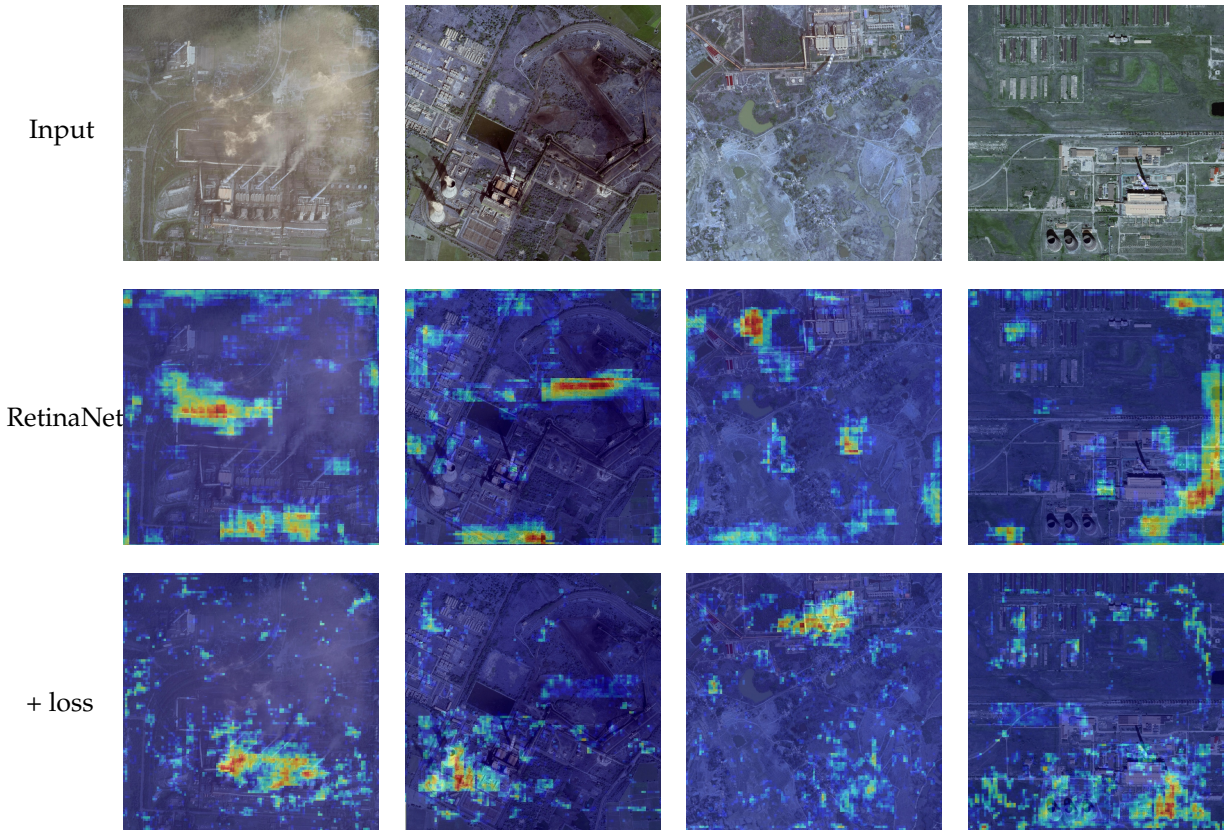


Figure 6. Several input images and corresponding heat map visualization of feature representation in ablation experiments, which proves that part-based attention module can help pay more attention to the distinctive components in the training process. First line: input image examples. Second line: heat map visualization results of RetinaNet for corresponding images. Third line: heat map visualization results of RetinaNet + the loss function defined in part-based attention module.

3. Experiments

3.1. Dataset and Settings

3.1.1. Dataset

Large-scale datasets in remote sensing images such as UCMD [19], EuroSAT [20] and DOTA [21], have contributed to the development of the general object detection of remote sensing images. However, existing publicly available datasets in remote sensing only cover limited categories of objects [22–25]. There is no annotated dataset of fixed industrial facilities including thermal power plants, garbage dumps and sewage treatment plants to the best of our knowledge.

In order to push forward the deep learning based development of the detection of TPPs, we construct a thermal power plant dataset of visible spectrum Google Earth images for object detection, which will be publicly available. We collect 257 potential locations of worldwide power plants from public websites, and then download images of all these locations, examine and check them earnestly. Sites with low credibility are omitted and 230 thermal power plants remain. To increase the diversity of data, we collect historical images of the 230 valid sites from Google Earth, and obtain 487 images ultimately. Each image is 3584 × 3584 pixels, covering the land of 2 km × 2 km with a resolution of

0.60m. All the objects are annotated with horizontal bounding boxes and finally obtain a COCO-style dataset.

In addition, to facilitate the representation of TPPs, we provide annotations including the whole *PLANT* and four components, that is *Coal Yard*, *Chimney*, *Pool* and other processing buildings (*Processing*). *Coal Yard*, *Chimney* and *Pool* are typical components in a thermal power *PLANT*. The study in this paper uses the *PLANT* annotations on a sub-dataset of 300 coal-fired TPP images. The 300 coal-fired TPP images are split into training and testing data with a ratio of 7 to 3. The data in Aerospace Information Research Institute-Thermal Power Plants Dataset for Detection (AIR-TPPDD) are respectively augmented by random cropping and flipping to obtain a dataset of 2000 samples of 900 × 900 pixels to adapt to deep learning based methods.

3.1.2. Evaluation Metrics

To evaluate the practical application of our proposed detection methods for TPPs, we adopt the standard mean average precision (mAP), frame per second (FPS), floating-point operations per second (FLOPs) and the number of trainable parameters (Params) in our experiments.

In a detection task, the predicted bounding boxes can be divided into true positive (TP), true negative (TN), false positive (FP) and false negative (FN). *Precision* and *recall* of detection results are calculated as:

$$p = \frac{TP}{TP + FN} \tag{8}$$

$$r = \frac{TP}{TP + FP}. \tag{9}$$

The F_1 score is the harmonic mean of the precision and recall, which can evaluate the performance comprehensively. Given two bounding boxes B_1 and B_2, intersection over union (IoU) is defined as

$$IoU = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} = \frac{|B_1 \cap B_2|}{|B_1| + |B_2| - |B_1 \cap B_2|}. \tag{10}$$

When IoU varies, precision and recall will change dynamically, constructing the precision-recall (PR) curve. The average precision is viewed as the area under PR curves obtained by setting different IoUs. More specifically, AP@0.5 and AP@0.75 are the areas under the PR curve setting IoU = 0.50 and 0.75 respectively. mAP@[0.5:0.95] is the average AP when IoU ranges from 0.5 to 0.95 in steps of 0.05, which is used as the main evaluation criterion for our task.

$$F_1 = \frac{2pr}{p + r} \tag{11}$$

$$AP = \int_0^1 p(r)\, dr. \tag{12}$$

In addition, average frame per second (FPS) is the number of processing images per second during test stage, which represents the time cost for application. Floating-point operations per second (FLOPs) and the number of trainable parameters (Params) are commonly used to indicate the complexity of deep models. Experiments are all implemented under the same hardware conditions.

3.1.3. Parameter Settings

All experiments are implemented with the PyTorch framework on a NVIDIA TITAN RTX with CUDA11.1. The pre-trained model ResNet-50, which was trained on the ImageNet dataset [26], is used to initialize the network. For the balance between the large-size scene requirements for objects and training efficiency for deep network, all images are processed to 900 × 900 pixels by random cropping and flipping in experiments.

We then utilize stochastic gradient descent [27] to train the network with a momentum of 0.9 and weight decay of 5×10^{-4}. The learning rate is initialized as 0.001 and then dropped by a factor of 0.1 every 10000 steps. In classification loss (Equations (2) and (7)), we set $\alpha_t = 0.25$ and $\gamma = 2$ according to RetinaNet [12]. The ratio of negative and positive samples in training stage is set to 3 in order to suppress negative samples. The balancing factor α in loss function (Equation (5)) is set to 0.25 without specific notice.

3.2. Ablation Study

3.2.1. Effect of CMN

In this section, the proposed CMN is trained for exploring the influence to the generated feature maps. Experiments use the same detection framework and unchanged parameters based on RetinaNet [12]. As shown in Figure 5, MFN is designed for irregular multi-scale feature representation by introducing global spatial attention and deformable convolution.

To prove the effectiveness of CMN we proposed, ablation experiments are designed as Table 1. As Figure 5, CMN can be viewed as the sum of deform module (DCN part) and context module (context attention and a transformer). CMN can be added to feature maps $\{C_3, C_4, C_5\}$ of backbone ResNet50. In Table 1, adding deform module or context module brings a certain improvement to the predicted detection results. It can also be seen that detection results of CMN are mostly obviously improved with the increase in network complexity except RetinaNet+CMN(C_{35}). It could be because that C_5 and C_3 are processed by CMN, so M_5 and M_3 can adapt to non-rigid irregular objects rather than M_4 in Figure 5. As a result of that, M_4 in RetinaNet+CMN(C_{35}) is not consistent with M_5 and M_3, which does not benefit the optimization of networks. In general, the most obvious improvement can reach 4.25% and the mAPs of seven listed ways of CMN addition are enlarged, which can prove the reliability of our proposed CMN.

Table 1. Adding CMN for feature extraction.

Method	C_3	C_4	C_5	mAP	ΔmAP	maxF_1	FLOPs
RetinaNet	-	-	-	0.6309	-	0.665	192.31G
+Deform(C_{345})	✓	✓	✓	0.6486	+1.77%	0.686	202.23G
+Context(C_{345})	✓	✓	✓	0.6564	+2.55%	0.678	192.32G
+CMN(C_3)	✓	-	-	0.6530	+2.21%	0.680	199.84G
+CMN(C_4)	-	✓	-	0.6618	+3.09%	0.697	194.23G
+CMN(C_5)	-	-	✓	0.6494	+1.85%	0.672	192.81G
+CMN(C_{34})	✓	✓	-	0.6729	+4.20%	0.713	201.75G
+CMN(C_{35})	✓	-	✓	0.6449	+1.40%	0.679	200.33G
+CMN(C_{45})	-	✓	✓	0.6632	+3.23%	0.702	194.72G
+CMN(C_{345})	✓	✓	✓	0.6734	+4.25%	0.719	202.25G

3.2.2. Effect of Part-Based Attention Module

As discussed in Section 2.3, part-based attention module is beneficial to the detection of thermal power plants. An adjustable parameter α is used to balance L_{part} (Equation (6)) with L_{global} (Equation (1)) in loss function (5). Same as CMN, we evaluate the effects of part-based loss function based on RetinaNet in ablation experiments. Table 2 shows that by replacing the loss function, the detection result is improved to 65.58% with respect to RetinaNet, delivering a gain of 2.49%. Figure 6 illustrates several input images and corresponding heat maps of feature representation which proves that part-based attention module can help pay more attention to the distinctive components in the training process. Furthermore, part-based attention module with balancing factor α=0.25 is demonstrated in Table 2 to bring an extra improvement of 0.81% to the networks with CMN. Results on our best method RetinaNet+CMN+Part-based-loss (PCAN) are visualized in Figures 7 and 8.

Table 2. Varying α in loss function.

α	mAP	maxF_1
0	0.6309	0.665
0.10	0.6502	0.677
0.25	0.6558	0.681
0.50	0.6550	0.679
0.75	0.6487	0.674
0.99	0.6213	0.622
0.25+CMN	0.6815	0.731

To balance L_{part} with L_{global}, we part-based multiply a modulation factor α to L_{part} in Equation (5). In this section we change α from 0 to 1 to investigate its influence. All models share the same experiment settings based on RetinaNet. When $\alpha = 0$, Equation (5) degenerated into Equation (1) as RetinaNet. If we set α to 1, part-based loss L_{part} weighs the same as L_{global} in overall loss function. Table 2 shows the performance of models with different α, from which we can find that the performance of models with different α approximately obeys normal distribution, achieving best α in [0.25, 0.5]. When α is close to 1, mAP obviously decreases. This result may be because L_{part} roughly focus on the distinctive components which is not reasonable if it has a huge impact on the total loss. For simplicity, modulation factor α is set to 0.25 without specific notice in the part-based attention module to benefit the localisation of TPP targets.

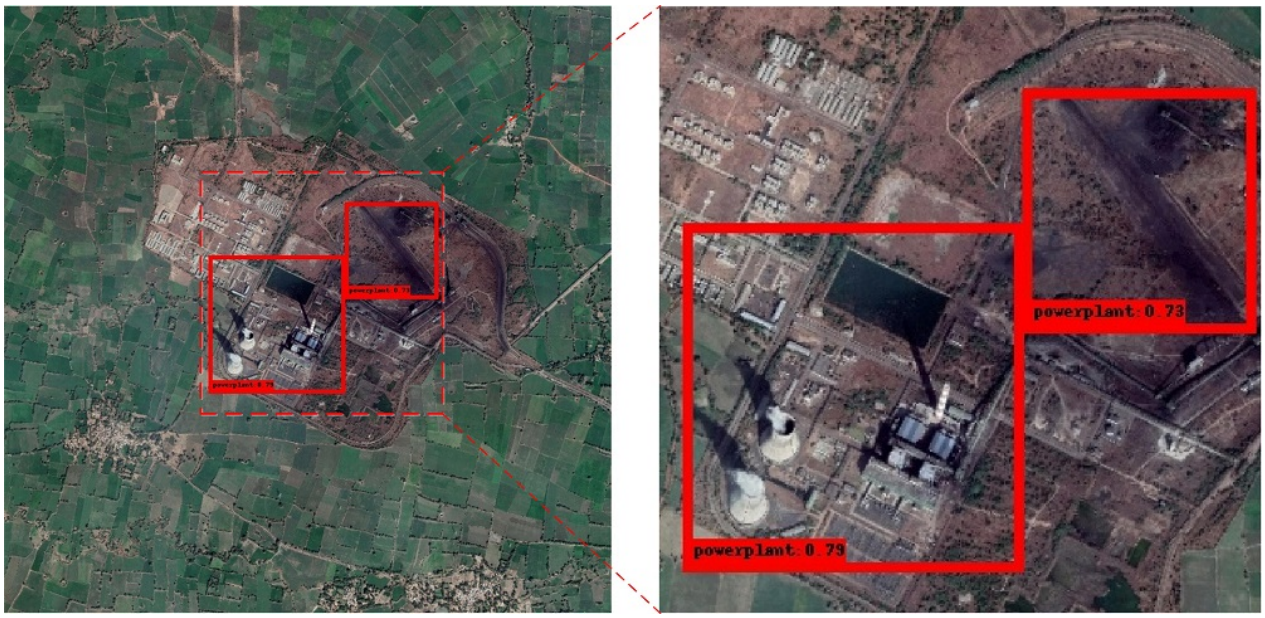


Figure 7. TPP test results on a remote sensing image which covers 2 km $\times$ 2 km. The detected category name and confidence are marked in the lower left corner in predicted bounding boxes.

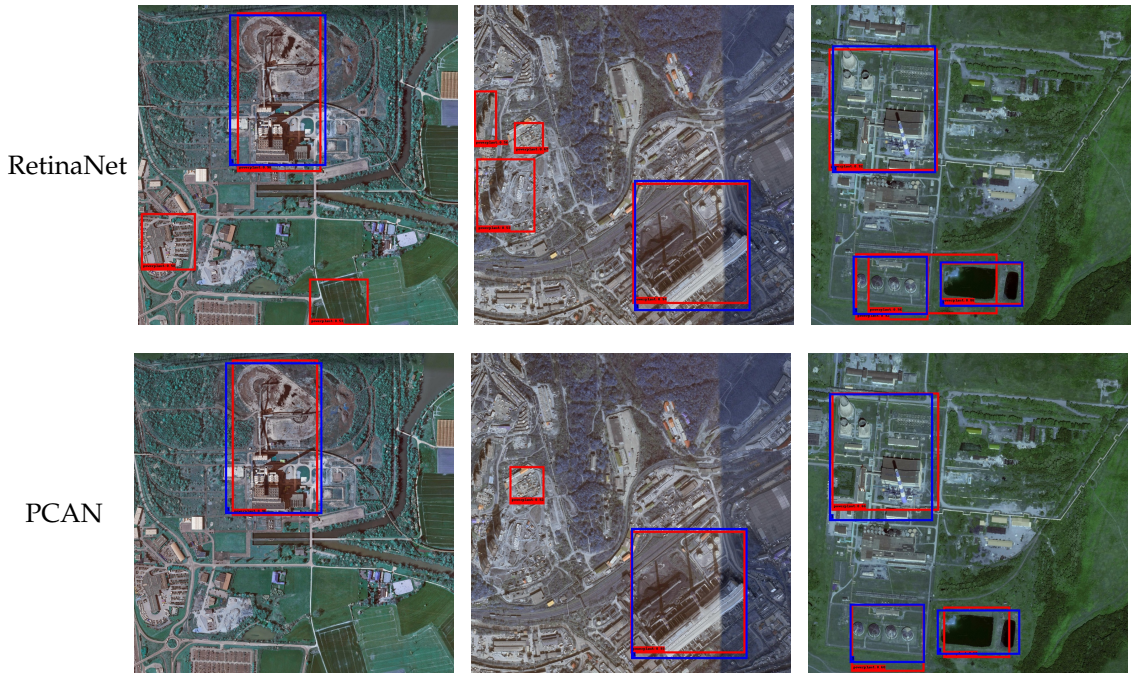


Figure 8. Detection results on Aerospace Information Research Institute-Thermal Power Plants Dataset for Detection (AIR-TPPDD) dataset. The blue box is the ground-truth, and the red box is the test result. Images at the top and the bottom are respectively the results of RetinaNet and our PCAN. The suppression of false alarms indicates that the proposed framework can generate effective deep features for thermal power plants.

3.3. Comparison with State-of-the-Arts

In this paper, classic one-stage RetinaNet detector [12] is used as the baseline method due to its simple structure and wide application in object detection. Two-stage detector Faster-RCNN detector [28] and multi-stage detector CascadeRCNN [29] are also included in contrast experiments.

RetinaNet [12] extracts deep features by ResNet [13] and FPN [14], and then uses a box subnet and a class subnet to obtains locations and class labels of anchors. Focal loss is designed to deal with class imbalance in one-stage detectors, which enlarges the weight of hard examples in cross-entropy loss.

Faster-RCNN [28] is a two-stage framework by integrating the Fast-RCNN [19] with RPN, which also extracts deep feature maps by a CNN backbone. FPN is added in our experiments to the backbone for multi-scale feature extraction. RPN is then trained to generate region proposals and ROI pooling computes proposal feature maps. Lastly, a classifier is used to predict the labels for each proposal and refine proposals.

For better match between the intersection over union (IoU) thresholds for which the detector is optimal and those of the input hypotheses, Cascade-RCNN [29] includes a sequence of detectors trained with increasing IoU thresholds. Compared with Faster-RCNN, Cascade-RCNN consists of at least two more ROI poolings and classifiers which are trained stage by stage.

All experiments are implemented under the same hardware conditions. As shown in Table 3, our PCAN increases mAP by 5.06% compared to RetinaNet. Furthermore, according to hypothetical test principle, statistical tests of the detection results of baseline RetinaNet show that $P(mAP = 0.6309 \pm 0.19\%) = 0.95$ and final results of PCAN show $P(mAP = 0.6815 \pm 0.63\%) = 0.95$, which indicates the enhancement of representation ability of deep feature maps for TPPs. FPS, FLOPs and the number of trainable parameters (Params) are listed in Table 3. It is thus convincing that our method gained better performance than RetinaNet without too much time and memory cost.

Experiments show that mAPs obtained by Faster-RCNN and Cascade-RCNN are close to mAP obtained by RetinaNet, with minor improvements in accuracy between multi-stage and one-stage methods. However, RetinaNet runs much faster than Faster-RCNN and Cascade-RCNN with less number of trainable parameters. This could be because that complicated models are not easy to optimize, especially for the non-rigid irregular TPP object.

Furthermore, experiments of remote sensing ship detection are performed on the AIR-SARShip dataset [30], as shown in Table 4. Results demonstrate a minor improvement of our PCAN for ship detection, which indicates that our proposed method is more suitable for the detection of thermal power plants rather than other objects.

Table 3. Performance of different methods for TPP detection.

Method	mAP	maxF_1	FPS(/s)	FLOPs	Paras(MB)
RetinaNet [12]	0.6309	0.665	19.61	192.31G	34.67
Faster-RCNN [28]	0.6443	0.672	10.3	250.13G	26.97
Cascade-RCNN [29]	0.6518	0.680	5.33	294.06G	93.75
Our PCAN	0.6815	0.731	16.24	246.37G	35.28

Table 4. Experiments on ship dataset [30].

Method	mAP	FPS(/s)
RetinaNet [12]	0.811	58.0
Faster-RCNN [28]	0.793	32.6
Our PCAN	0.824	50.4

4. Discussion

By comparing and analyzing the above experiments, the effectiveness of the proposed method is verified. The proposed PCAN offers superior performance in the TPPs detection task by CMN and part-based attention module based on RetinaNet.

However, through observation of the test results in Figure 9, we can see that not all the detection results are ideal. Figure 9 shows some examples of false alarms and missing alarms, which are mainly caused by hard examples. Hard examples found in our experiments include the following two situations: (1) Disturbances due to similar surfaces. Some background scenes, such as buildings, parking lots and pools, locate near a TPP in Figure 9a–c. In Figure 9a, some residential buildings appear similar to the processing buildings in TPP. In Figure 9b, a parking lot with regularly arranged cars is mistakenly detected. In Figure 9c, it is sometimes difficult to distinguish whether a nearby pool is a part of the TPP object. (2) Missed alarms caused by occlusion and edge location. Objects blocked by clouds and located near the edge make it difficult to recognize in Figure 9d.

In the future, we will explore how to enhance the recognize ability of detectors in order to effectively reduce false alarms and missed alarms. We are particularly planning to split the AIR-TPPDD dataset into easy and hard examples, and then focus on the hard examples during training with strict limitation of ratio of hard and easy samples. In addition, detection and classification of power plants including coal-fired power plants, oil-fired power plants, gas-fired power plants, waste heat power plants may be implemented by constructing more detailed power plants images, which will be carried out in the future.

(**a**) (**b**) (**c**) (**d**)

Figure 9. False alarms and missed alarms on hard examples. (**a**) Disturbances: buildings; (**b**) Disturbances: parking lots; (**c**) Disturbances: pools that locate near a TPP but do not belong to it; (**d**) Missed alarms caused by occlusion and edge location.

5. Conclusions

The detection of thermal power plants is a meaningful but challenging task. The difficulty results from the lack of annotated dataset and highly complex appearances of TPPs. In this paper, an effective TPP detection method, which includes context attention multi-scale feature extraction network (CMN) and part-based attention module, is proposed to solve the problem. CMN enhances the local convolutional features and part-based attention module strengthen the influence of components in TPPs. Experiments demonstrate the effectiveness of our proposed part-based context attention networks (PCAN).

Author Contributions: W.Y., W.D. and P.W. conceived and designed the experiments; W.Y. performed the experiments; W.Y., W.D. and Y.L. analyzed the data; Y.L. and X.S. contributed materials; W.Y. wrote the manuscript, which was revised by W.D. and X.S.; X.G. and X.S. supervised the study and reviewed this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 41701508.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used in this study. This data can be found here: https://github.com/wenxinYin/AIR-TPPDD, accessed on 14 March 2021.

Acknowledgments: The authors are thankful for all the colleagues in the lab, who helped to build the image dataset. The authors would also like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

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

References

1. Zhang, Y.; Zhang, Y.; Du, B.; Zhang, C.; Tu, W. Parallel discriminative subspace for city target detection from high dimension images. *GeoInformatica* **2020**, doi:10.1007/s10707-020-00399-7.
2. Dong, Y.; Du, B.; Zhang, L.; Hu, X. Hyperspectral Target Detection via Adaptive Information—Theoretic Metric Learning with Local Constraints. *Remote Sensing* **2018**, *10*, 1415. doi:10.3390/rs10091415.
3. Nasrabadi, N. Hyperspectral Target Detection : An Overview of Current and Future Challenges. *Signal Process. Mag. IEEE* **2014**, *31*, 34–44. doi:10.1109/MSP.2013.2278992.
4. Sumbul, G.; Cinbis, R.; Aksoy, S. Multisource Region Attention Network for Fine-Grained Object Recognition in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote. Sens.* **2019**, doi:10.1109/TGRS.2019.2894425.
5. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
6. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape Robust Text Detection With Progressive Scale Expansion Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.

7. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and Accurate Arbitrary-Shaped Text Detection With Pixel Aggregation Network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8440–8449
8. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
9. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159. doi:10.1109/CVPR.2014.27.
10. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 907–915. doi:10.1109/CVPR.2017.103.
11. Han, J.; Yao, X.; Cheng, G.; Feng, X.; Xu, D. P-CNN: Part-Based Convolutional Neural Networks for Fine-Grained Visual Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, doi:10.1109/TPAMI.2019.2933510.
12. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. doi:10.1109/ICCV.2017.324.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
14. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
15. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
16. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27 October–2 November 2019.
17. Good, I.J. Some Terminology and Notation in Information Theory. *Proc. IEEE Part Monogr.* **1956**, *103*, 200–204.
18. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. doi:10.1109/TIT.1982.1056489.
19. Yang, Y.; Newsam, S. Bag-of-visual-words and Spatial Extensions for Land-use Classification. In Proceedings of the 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, San Jose, CA, USA, 3–5 November 2010; pp. 270–279. doi:10.1145/1869790.1869829.
20. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226, doi:10.1109/JSTARS.2019.2918242.
21. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 23 June 2018.
22. Fan, Q.; Brown, L.; Smith, J. A closer look at Faster R-CNN for vehicle detection. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV), Gotenburg, Sweden, 19–22 June 2016; pp. 124–129. doi:10.1109/IVS.2016.7535375.
23. Yin, Shoulin, L.H.; Teng, L. Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images. *Sens. Imaging* **2020**, *21*. doi:10.1007/s11220-020-00314-2.
24. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3377–3390.
25. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS^4 Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5398–5413.
26. Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai Li.; Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. doi:10.1109/CVPR.2009.5206848.
27. Robbins, H.; Monro, S. A Stochastic Approximation Method. *Ann. Math. Stat.* **1951**, *22*, 400–407.
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. doi: 10.1109/TPAMI.2016.2577031.
29. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 6154–6162. doi:10.1109/CVPR.2018.00644.
30. Sun, X.; Wang, Z.; Sun, Y.; Diao, W.; Zhang, Y.; Fu, K. AIR-SARShip-1.0: High-resolution SAR ship detection dataset. *J. Radars* **2019**, *8*, 852–862, doi:10.12000/JR19097.