International Journal of Geo-Information

MDPI

Article

Small Manhole Cover Detection in Remote Sensing Imagery with Deep Convolutional Neural Networks

Wei Liu [1,2,*], Dayu Cheng [2,3,*], Pengcheng Yin [4], Mengyuan Yang [1], Erzhu Li [1], Meng Xie [1] and Lianpeng Zhang [1]

1 School of Geography, Geomatics and Planning, Jiangsu Normal University, Xu Zhou 221116, Jiangsu, China; myfengxi@126.com (M.Y.); liezrs2018@jsnu.edu.cn (E.L.); 15262036926@163.com (M.X.); zhanglp2000@126.com (L.Z.)
2 State Key Laboratory of Resources and Environmental Information System, Beijing 100101, China
3 School of Mining and Geomatics, Hebei University of Engineering, Handan 056038, Hebei, China
4 Bureau of Land and Resources of Xuzhou, Xuzhou 221006, Jiangsu, China; cumtyingpc@163.com
* Correspondence: liuw@jsnu.edu.cn (W.L.); chengdy@lreis.ac.cn (D.C.); Tel.: +86-130-1393-9855 (W.L.); Fax: +86-158-1137-1850 (D.C.)

Received: 28 December 2018; Accepted: 16 January 2019; Published: 19 January 2019

Abstract: With the development of remote sensing technology and the advent of high-resolution images, obtaining data has become increasingly convenient. However, the acquisition of small manhole cover information still has shortcomings including low efficiency of manual surveying and high leakage rate. Recently, deep learning models, especially deep convolutional neural networks (DCNNs), have proven to be effective at object detection. However, several challenges limit the applications of DCNN in manhole cover object detection using remote sensing imagery: (1) Manhole cover objects often appear at different scales in remotely sensed images and DCNNs' fixed receptive field cannot match the scale variability of such objects; (2) Manhole cover objects in large-scale remotely-sensed images are relatively small in size and densely packed, while DCNNs have poor localization performance when applied to such objects. To address these problems, we propose an effective method for detecting manhole cover objects in remotely-sensed images. First, we redesign the feature extractor by adopting the visual geometry group (VGG), which can increase the variety of receptive field size. Then, detection is performed using two sub-networks: a multi-scale output network (MON) for manhole cover object-like edge generation from several intermediate layers whose receptive fields match different object scales and a multi-level convolution matching network (M-CMN) for object detection based on fused feature maps, which combines several feature maps that enable small and densely packed manhole cover objects to produce a stronger response. The results show that our method is more accurate than existing methods at detecting manhole covers in remotely-sensed images.

Keywords: manhole cover; remote sensing images; object detection; deep convolutional neural networks

1. Introduction

Manhole cover surveys are a complex issue in engineering. Digital city management makes it necessary to obtain the spatial position and attribute information of a given manhole cover quickly and accurately. The most commonly used manhole cover measurement method is survey-based manual acquisition using a total station, digital camera and other equipment. To reduce labor and time costs, researchers have proposed a variety of manhole cover detection methods, such as unmanned aerial vehicle (UAV) surveys, mobile measurement acquisition vehicles and on-board laser scanning [1–3]. The above methods are more efficient to some extent but they cannot eliminate the disadvantages of

heavy workload involved in field collection and the complexities of internal processing. Furthermore, vehicle-mounted equipment can only determine information for manhole covers near roadside curbs.

With improvements in the spatial resolution of remotely-sensed images, an increasing number of scholars have studied object detection in remotely-sensed images [4–10]. Presently, this approach is mainly divided into three categories of methods: object-based image analysis, fusion of spatial information and machine learning. The mainstream object detection algorithms are mainly based on deep learning models, which can be divided into two categories. First, two-stage detection algorithms can produce region proposals, based on the classification scores of these proposal uses non-maximum suppression (NMS) to eliminate redundant proposal, after NMS screening to obtain detected objects. Representative algorithms include R-CNN series algorithms based on region proposals, such as R-CNN [6], Fast R-CNN [11] and Faster R-CNN [12]. Second, one-stage detection algorithms that do not need the region proposal phase directly generate the class probability and position coordinates of objects; typical examples are YOLO [13] and SSD [14].

Although Faster R-CNN, YOLO and SSD have proven to be successful for detecting objects such as cats, cars, ships or people in nature-based images, they have not been specially designed to detect small manhole cover objects in remotely-sensed images and several challenges limit their applications in this manner. Manhole covers are relatively small and appear in densely distributed groups; even in a high-resolution optical remote sensing image with 0.1 m $\times$ 0.1 m resolution, a manhole cover is only 5 to 8 pixels wide. Faster R-CNN, YOLO and SSD struggle with small objects because the CNN features used for object detection are pooled from the topmost convolutional feature map with lower resolution. After being down-sampled multiple times, the manhole cover object size in the topmost convolutional feature map is 1/16 or 1/32 of the original size in the input remotely-sensed images. This drop in resolution may result in important features being lost and thereby lead to poor detection performance.

To address these issues, in this paper, we demonstrate a method for improving the HED network [15] and propose an effective deep CNN-based approach for detecting small manhole covers in remotely-sensed images. Similar to Faster R-CNN, our method consists of two stages: a multi-scale output network (MON) and a multi-level convolution matching network (M-CMN). First, we redesign the architecture of the feature extractor adopting the visual geometry group (VGG) method [16], which can increase the variety of receptive field sizes. For detecting small manhole cover objects, MON combines several side-outputs of intermediate layers to increase the resolution of feature maps; thereby, enabling small and densely packed manhole cover objects to produce larger regions of strong responses. The object proposals from various intermediate feature maps are combined together to form the outputs of MON. Then, these object proposals are sent to the M-CMN for accurate object detection.

The main contributions of this paper are as follows:

1. We redesigned the CNN architecture by adopting the powerful HED module to increase the variety of receptive field sizes that can be used to capture small manhole cover objects more effectively. Although HED has been tested for scene classification and edge detection, to our knowledge this is the first time it has been used to verify the effectiveness of small object detection tasks within remotely-sensed images.

2. We combined multiple intermediate feature maps so that multiple levels of details can be considered simultaneously, thus increasing the resolution and improving the detection accuracy of small and densely concentrated manhole cover objects.

The rest of this paper is organized as follows. Section 2 describes the framework of manhole covers object detection in detail. Section 3 presents the comparative experimental results for manhole covers object detection. Section 4 contains a discussion of these results and the conclusions are presented in Section 5.

2. Multi-scale CNN for Manhole Cover Object Detection

First, we redesigned the HED network structure. Figure 1 illustrates the detailed architecture of our proposed method, which consists of a VGG16 and a multi-scale output network (MON) that lead to

a multi-level convolution matching network (M-CMN). The VGG16 is used as a feature extractor while the MON aims to generate multi-scale side outputs with different filter receptive fields using a series of intermediate layers. After concatenating the output value of the side-outputs' de-convolution and conducting convolution calculations to obtain the fusion layer, these multiple side outputs and their fusion layer are sent to the M-CMN for accurate detection. With reference to deep supervision [15], the network parameters ℓ_{side} and L_{fuse} are automatically updated using a back-propagation algorithm to complete the network training.

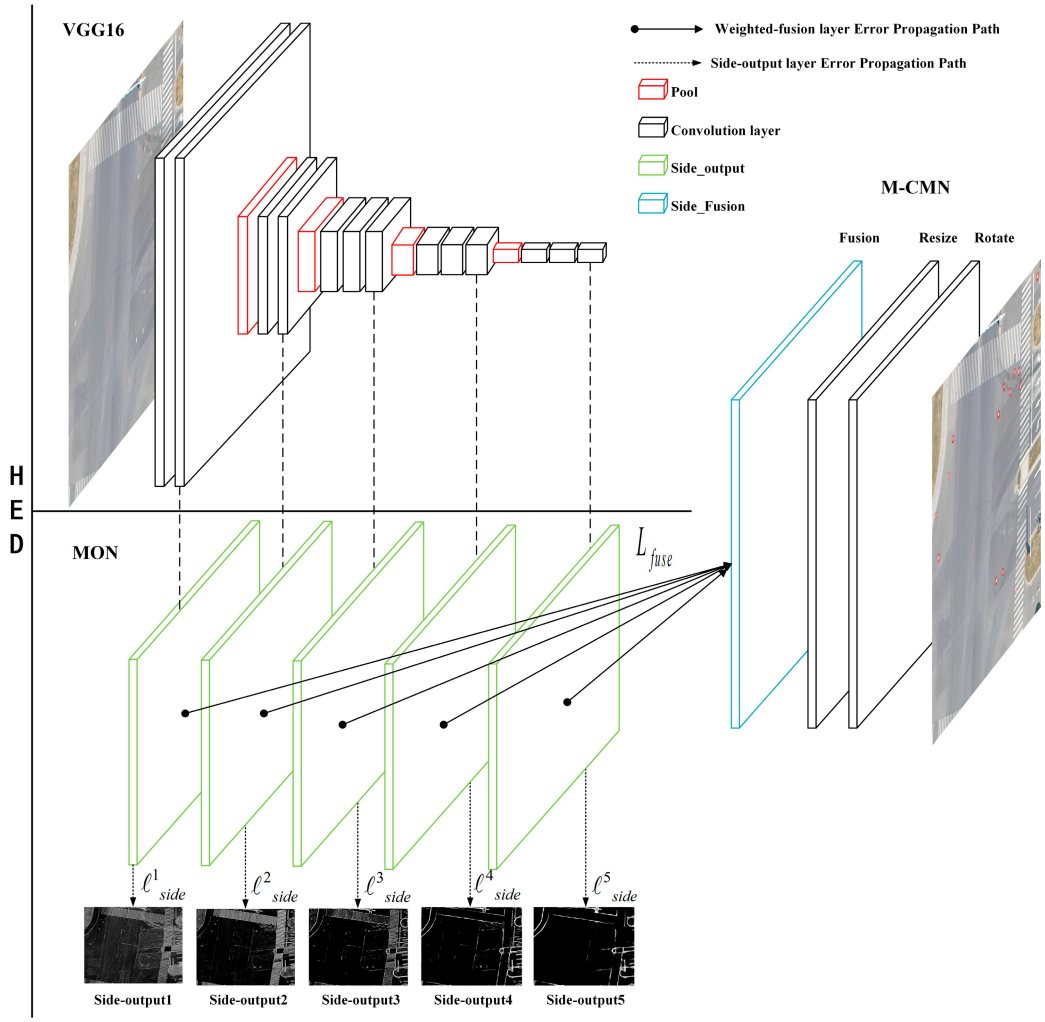

Figure 1. Architecture of our proposed method. Deep supervision is imposed at each side-output layer, guiding the side-outputs to obtain multi-level outputs. The subsequent fusion layer aids in learning how to combine outputs from multiple scales.

2.1. Details on VGG16 Architecture

The feature convolution extractor takes a remote sensing image of any size as input and outputs multiple-level feature maps. The design of this element is of crucial importance as the types of layers and number of parameters directly affects its efficiency, accuracy and performance. Studies [17] have shown that using a deeper convolution model with a depth of several hundred layers can significantly improve the performance of many visual recognition tasks, such as object detection, image classification and semantic segmentation. However, it is difficult to directly use the very deep object detection model in remote sensing image recognition because these very deep models can incur high computational costs, since remote sensing images are large (usually several hundred megapixels). Furthermore, a very deep computational model requires a large number of training samples but there is a relative dearth of

labeled remote sensing images that can be used as training data. In order to fulfill these requirements, we adopted VGG16 as the backbone network, which has been widely used in feature extractors due to its good generalization performance [18,19].

Because the detection task in this study involves very small objects, we adopted the tail-cutting, small convolution kernel and "keep input size" techniques for the VGG16 network to increase the network depth. This ensures that the input size of each layer does not decrease sharply with an increase in the depth and that it is better adapted to small object detection. For this reason, the network only uses the first five groups of VGG16; the fully connected layers and Soft-Max layers are clipped out. After tail-cutting, the network structure shown in Figure 2 is obtained, with the network configuration given in Table 1. The modified network is a multi-scale and multi-fusion feature learning network structure that, as shown in Figure 1, extracts the output of the last convolution layer in each set of VGG16 because the size of each image set is different. Therefore, it is also necessary to use transposed convolution/de-convolution to extend the image from each group, which is, in effect, equivalent to 2–16 times the size of the second to fifth groups of images, respectively. In this way, the image at each scale (each set of VGG16 is a scale) is the same size and these images are then fused together.

There are three advantages to the above process.

1. Cutting the fully connected layers and Soft-Max layers can significantly reduce the memory and time cost during both training and testing. Furthermore, because there is no restriction on fully connected layers and Soft-Max layers, remote sensing images of any size can be input for training and object detection.

2. Using 1×1 convolution kernels can help reduce the convolution parameters. Furthermore, the network capacity and model complexity can be enhanced effectively, which improves small manhole cover object detection.

3. Transposed convolution/de-convolution allows each group of images to be extended, which can make the feature output "keep input size" and better adapt it to small manhole cover detection.

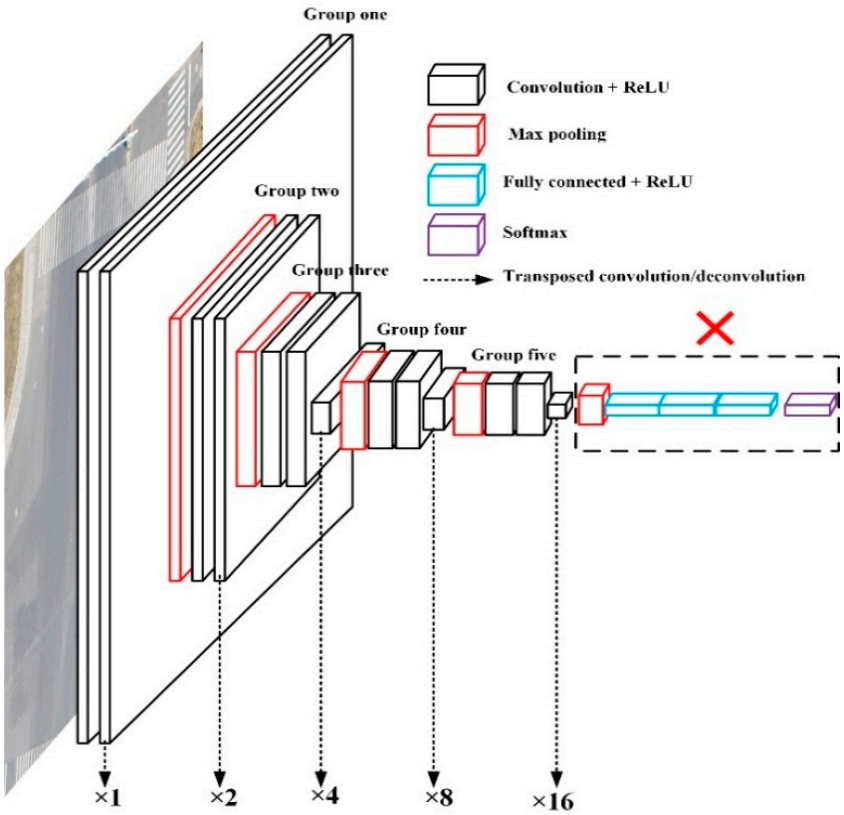

Figure 2. Details of the improved VGG16 architecture.

Table 1. VGG16 network configuration.

Layers	C1	C2	C3	C4	C5	Fully Connected
16	Conv3-64	Conv3-128	Conv3-256	Conv3-512	Conv3-512	FC-4096
	Conv3-64	Conv3-128	Conv3-256	Conv3-512	Conv3-512	FC-4096
	Max pool	Max pool	Conv1-256	Conv1-512	Conv1-512	FC-1000
			Max pool	Max pool	Max pool	Soft-max

Note: the convolutional layer parameters are denoted as "convolution kernel size-number of channels." For brevity, the ReLU activation function is not shown.

2.2. Multi-scale Output Network (MON)

HED is a new multi-scale and multi-level feature learning algorithm using edge detection to achieve end-to-end prediction with a deep learning model based on fully convolutional neural networks and deeply supervised networks. HED automatically learns rich hierarchical expressions and is very important for solving challenging fuzziness in edge detection and object boundary detection. In order to handle the very small size of manhole covers, we improved the HED model to better conduct boundary extraction and generate edge regions through several intermediate layers with different receptive fields (Table 2), inspired by SSD. This is named the multi-scale output network (MON). Specifically, we added smaller size filters (1×1) to capture densely packed manhole cover objects in remote sensing images.

Table 2. The receptive field (RF) and stride size of the VGG16 used in MON.

Layer	C1_2	P1	C2_2	P2	C3_3	P3	C4_3	P4	C5_3	P5
RF	5	6	14	16	32	36	68	76	140	156
Stride	1	2	2	4	4	8	8	16	16	32

As shown in Figure 1 and 2, we connect our side output layer to the last convolutional layer in each group (conv1_2, conv2_2, conv3_3, conv4_3 and conv5_3). The receptive field size of each of these convolutional layers is identical to the corresponding side-output layer. We cut the last group of VGG16, including the 5th pooling layer and all the fully connected layers, because the layer with stride 32 yields an output manhole cover that is too small, with the consequence that the interpolated prediction map will be too fuzzy to utilize.

During training, more than 90% of the pixels in ground-truthing are non-edge, which is extremely biased. In response to this biased sampling, Hwang and Liu [8] introduced a cost-sensitive loss function and included additional tradeoff parameters. MON uses the same method as HED to avoid the loss of balance between positive and negative samples, which is to introduce a pixel level of class-balancing weight β. The class-balanced cross-entropy loss function is defined as follows:

$$\ell_{side}^{(m)}(W, w^{(m)}) = -\beta * \log(sigmoid(y)) - (1-\beta) * \log(1 - sigmoid(y)) \quad (1)$$

where W refers to the standard network layer parameters, $w = (w\ (1), \ldots, w(m))$ refers to the corresponding weights of each side-output layer, m is the network having m side-output layers and $\beta = |Y\_|/|Y|$ and $1 - \beta = |Y+|/|Y|$ and $|Y\_|$ and $|Y+|$ are the edge and non-edge ground truth label sets, respectively.

In order to use the side-output prediction results, the loss function named L_{fuse} (W, w, h) for adding a weighted-fusion layer is:

$$L_{fuse}(W, w, h) = Dist(Y, \hat{Y_{fuse}}) \quad (2)$$

where $h = (h_1, \ldots, h_m)$ is the fusion weight and $Dist(\ldots)$ is the distance between the fused predictions and the ground truth label map. Putting every loss function together, MON minimizes the following objective function via back propagation stochastic gradient descent (SGD):

$$(W, w, h)^* = \text{argmin}(\ell_{side}^{(m)}(W, w^{(m)}) + L_{fuse}(W, w, h)) \quad (3)$$

where the optimal parameters W are optimized by SGD [20]. To prevent over-fitting, we adopt the pre-trained HED model [15] for PASCAL VOC-2012 segmentation to initialize the convolutional layers. When the training of the MON is completed, we obtain the predicted results from the side-output layer and the weighted-fusion layer at the same time and merge them together to get a better image.

2.3. Multi-level Convolution Matching Network (M-CMN)

Traditional object detection methods rescale the input remote sensing image multiple times (Figure 3a) or apply multiple filters to a single input image (Figure 3b) to match all possible target objects. This makes it difficult for these methods to use multiple feature map layers. In order to increase detection accuracy, we propose a new method, named M-CMN, to obtain the manhole cover detection through five side-output layers and some fusion layers with different filter sizes (Figure 3c). M-CMN takes a remote sense image with its predicted edges (generated by MON) as the input and outputs the refined manhole cover detection. Inspired by the success of combining multi-level representation in SSD, M-CMN combines multi-level layers with different resolutions to achieve more informative feature maps for accurate manhole cover detection (Figure 3c). Since manhole cover objects in large-scale remote sensing images are relatively small in size and appear in densely distributed groups, we specifically chose the conv1_2 layer as a reference layer and concatenated the conv2_2 and conv3_3 layers and the conv4_3 and conv5_3 layers with up-scaling (using transposed convolution/de-convolution). This is because the conv1_2 layer with higher resolution is better suited for detecting small, densely distributed manhole cover objects. As shallower layers are more suitable for reference and deeper layers are more suitable for matching, the concatenated feature maps are complementary for small size manhole cover detection, as shown in our experiments.

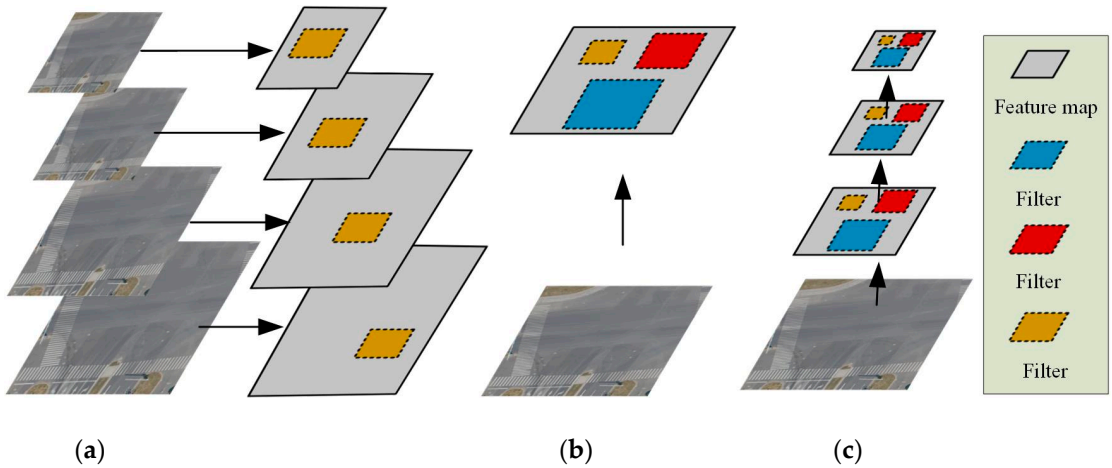

Figure 3. Different strategies for multi-level detection: (**a**) Prediction using multiple image scales with a single filter size; (**b**) Prediction using a single feature map with multiple filter sizes; (**c**) Prediction using multiple feature maps with multiple filter sizes.

3. Experimental Results

In this section, we evaluate our method for small manhole cover detection from remote sensing images. Experiments are implemented based on our deep learning framework and executed on

a server with E5-2697V4 * 2 CPU, NVIDIA K80*20 GPU, 256 GB memory and Ubuntu 16.04 as the server operating system.

3.1. Dataset

Aerial photographs of 0.05 m spatial resolution of Zhenjiang City, Jiangsu Province, taken in 2017, were used in this study. There is a survey result document (vector points) for manhole covers in this area, which can be easily labeled by ground truth batch in the image; the tagged ground truth data can be seen in Figure 4. Considering GPU memory and process speed, each original aerial image is cropped into several adjacent image blocks with resolutions of 512 × 512 pixels, making them easier to augment and consuming less GPU memory, which could improve training efficiency. Considering the small size of the manhole covers, we set the adjacent image block overlap ratio as 0.05 and removed the annotation of targets that cross image block boundaries. Then, the image blocks without manhole cover targets are discarded. Of the 2382 images (23252 manhole cover objects) after batch processing, 1500 are used as training data sets, 500 as validation datasets and 382 as test data sets.

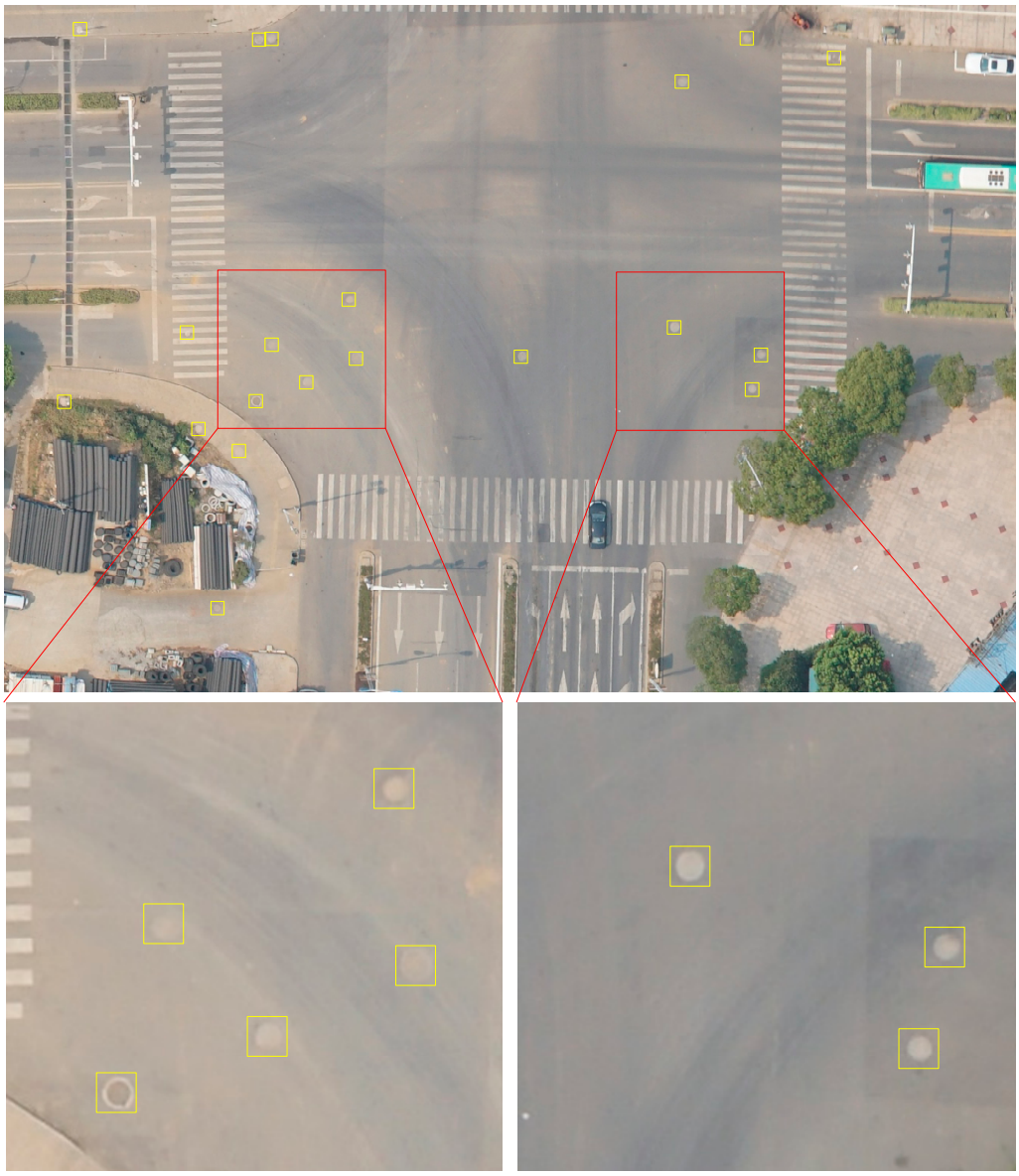

Figure 4. Ground truth boxes.

3.2. Model and Parameters

Our method requires relatively little engineering hacking as our framework is implemented using the publicly available Caffe and TensorFlow Library and the network is fine-tuned from the VGG16 model and pre-trained HED model.

Following the strategies outlined in Dollár and Zitnick [21], we evaluate various network modifications as well as training hyper-parameters on a validation set. Through experimentation, we chose the following hyper-parameters: mini-batch size (12), learning rate (1e-6), loss-weight αm for each side-output layer (1), momentum (0.9), nested filter initialization weights (0), fusion layer initialization weights (1/5), weight decay (0.0002) and training iterations (10,000; divide learning rate by 10 after 5,000).

3.3. Results

In order to better utilize convolution networks to detect manhole cover objects, we perform convolution detection on multiple output layers of MON. As shown in Figure 5, information on remote sensing images detected by the six output layers of the MON network shows that the detail of the side output layers 1–5 tends to degrade, while the edge of the fusion layer is obvious and some details are retained. This is because the MON network itself is a multi-scale fusion network. As the receptive fields of the side-output layer become larger, the local details gradually degrade, while the fusion layer retains the local details by obtaining multi-scale information.

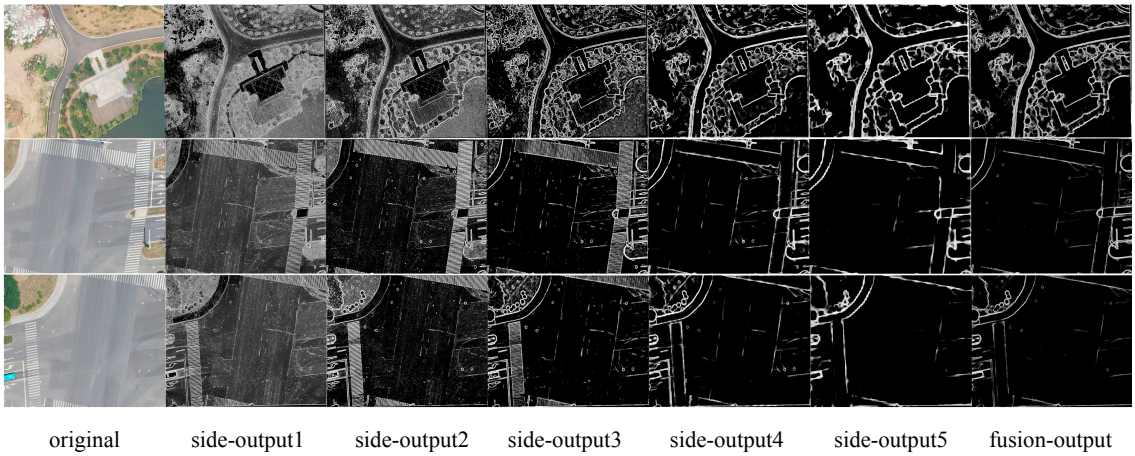

Figure 5. Outputs of MON detection.

We adopt three widely used indicators to evaluate the performance of manhole cover detection; namely, Precision, Recall and F1-score. Precision is the measure of the fraction of detections that are true positives and Recall is the measure of the fraction of positives that are correctly identified. The F1-score is the combined precision and recall metrics in a single measure for comprehensively evaluating the quality of an object detection method [4].

Table 3 shows the quantitative comparison results of nine different methods measured by Precision, Recall and F1-score; the best performances are italicized and underlined. The following observations were noted: (1) Compared with Faster R-CNN, YOLOv3 achieves similar detection performance measured in Precision and YOLOv3 obtains performance gains in Recall and F1-score. SSD obtains the highest Recall but the Precision is slightly inferior. Among the three comparison methods, YOLOv3 has the best performance when recall rates and accuracy are considered and its processing speed is also the fastest. (2) Compared with the SSD, the DSSD [22] algorithm boosts the FI-score to 0.8108. However, as DSSD uses resnet-101 as a backbone network, the training speed is much lower; FSSD [23] draws lessons from FPN and uses a small convolution kernel, so the accuracy of the algorithm is clearly improved and the training speed is not reduced. In the small manhole

cover test, FSSD performance is slightly better, with an FI-score reaching 0.8266. (3) Our fusion methods achieve optimal or suboptimal Precision, Recall and F1-score values for manhole cover objects. Compared with all nine methods, Ours-fusion 4 obtained the best performance in terms of Precision and Ours-fusion 2 obtained the best performance in terms of Recall, which shows that the manhole cover object appearing 4 times can effectively improve object detection accuracy but has a great influence on the Recall. The manhole cover object appearing twice can effectively raise the object detection Recall but has a great influence on object detection Precision. The result for three times fusion of manhole cover objects is the best in terms of appropriate Precision, Recall and F1-score. This is due to most manhole cover objects appearing three times in five side-output layers and one fusion layer (Figure 5).

Table 3. Performance comparisons of nine different methods in terms of Precision, Recall and F1-score. The italic underlined numbers denote the optimal values in each row. The bold numbers denote the suboptimal values in each row. Ours-fusion is the result of object detection using only the fusion layer (as illustrated in Figure 1) and Ours-fusion2 is the fusion result of the manhole cover appearing twice in the six layers of side-output1, side-output2, side-output3, side-output4, side-output5 and fusion; Ours-fusion3 and Ours-fusion4 can be deduced in turn.

Method	**Faster R-CNN**	**YOLOv3**	**SSD**	**DSSD**	**FSSD**
Precision	0.7108	0.7034	0.6720	0.8058	0.8039
Recall	0.7481	0.7858	0.8369	0.8159	0.8506
F1-score	0.7289	0.7395	0.7454	0.8108	**0.8266**
Speed (*fps*)	6.2	*76*	52	7.8	**69**
Method	Ours-fusion	Ours-fusion 2	Ours-fusion 3	Ours-fusion 4	
Precision	0.7626	0.6928	**0.8640**	*0.9486*	
Recall	0.8008	*0.9658*	**0.9272**	0.6949	
F1-score	0.7812	0.8068	*0.8946*	0.8022	
Speed (*fps*)	18	16	14.2	11.6	

Figure 6 shows a number of manhole cover object detection results with the proposed approach. Red dots represent the detected manhole cover objects. Some objects are densely peaked and small with complex backgrounds. This shows that our method can successfully detect most manhole cover objects.

Figure 6. Manhole cover detection results with the proposed approach on three images.

4. Discussion

4.1. Can Increasing the Depth of the Backbone Network Improve Small Object Detection Performance?

In this study, YOLOv3 used Darknet-53 as a backbone network and DSSD used Resnet-101as a backbone network. However, the true positive of YOLOv3 and DSSD object detection was not been significantly improved (Table 4). Darknet-53 and Resnet-101 are deeper than VGG16 and the extracted features have higher semantic information, so the detection effect of Darknet-53 and Resnet-101 should

be higher than that of VGG16. Some studies show that replacing VGG16 with Resnet-101 directly under the input image of 300 × 300 results in decreased rather than increased accuracy [22].

The image blocks with resolutions of 512 × 512 pixels used in this article were passed through increasing backbone network depth without significant improvements in the performance of small manhole cover detection (Tables 3 and 4). This is because manhole covers in remotely sensed images are relatively small and appear in densely distributed groups, while deep convolution networks like Darknet-53 and Resnet-101 used for object detection are pooled from the topmost convolutional feature map with lower resolution. After being down-sampled multiple times, the small manhole covers disappear in the topmost convolutional feature map. The side-output5 in Figure 5 shows that after only five down-sampling times, the manhole cover features disappear, leading to limited improvement of the detection effect using the Darknet-53 and Resnet-101 deep network. For the above reasons, we chose VGG16 as the backbone networks, adopted tail-cutting and small convolution kernel processing and gave up the large receptive field. The detection result was thus better for the small size, single structure and dense appearance of manhole cover objects.

4.2. Can Multi-scale and Multi-level Feature Fusion Improve the Performance of Small Object Detection?

SSD uses multi-scale feature maps to predict targets, uses high-level features with a larger receptive field to predict large objects and uses low-level features with a smaller receptive field to predict small targets. This raises a question: When using the features of a low-level network to predict small targets, the classification results of SSD for small objects are poor due to the lack of high-level semantic features. The point of using DSSD to solve this problem is to fuse high-level and low-level semantic information, enriching the prediction's regression bounding boxes and multi-scale feature maps of the classification task input so as to improve the detection effect. However, due to the model's complexity, its speed is much slower. FSSD uses FPN for reference and reconstructs a set of pyramid feature maps to clearly improve the detection efficiency of the model with superior (i.e., not as slow) speed. Overall, DSSD and FSSD do improve detection performance by fusing the high-level context feature map with the low-level feature map (Table 3).

The results in Table 3 show that the detection effects of DSSD and FSSD are significantly higher than SSD, with F1-scores can reach 0.8108, 0.8266 and 0.7454, respectively. However, as shown in Table 4, their true positive is not high; that of DSSD is even lower than that of SSD (3551 and 3641, respectively). DSSD and FSSD can reach such a high F-1-score because the false negative and false positive miscalculation rates are effectively reduced by multi-scale and multi-level feature fusion. That is, the probability of misjudging other objects as manhole covers has been greatly reduced through multi-scale and multi-level feature fusion. Our model also proves that the detection effect of small manhole covers can be improved by multi-scale and multi-level feature fusion, especially the true positive of Ours-fusion3 (4035), while the false negative and false positive can be reduced to 317 and 635, respectively.

Table 4. Performance comparisons of seven different methods in terms of True Positive, False Negative and False Positive.

Method	**Faster R-CNN**	**YOLOv3**	**SSD**	**DSSD**	**FSSD**
True Positive	3256	3420	3642	3551	3702
False Negative	1096	932	710	801	650
False Positive	1325	1442	1778	856	903
Method	Ours-fusion	Ours-fusion2	Ours-fusion3	Ours-fusion4	
True Positive	3485	*4203*	4035	3024	
False Negative	767	149	317	1328	
False Positive	1085	1864	635	164	

5. Conclusions

In this paper, we propose an effective DCNN-based approach for detecting small manhole cover objects in remote sensing images. The detection is performed using a redesigned DCNN feature extractor and is followed by two sub-networks: a MON for manhole cover object edge generation from several intermediate layers, whose receptive fields match different manhole cover object scales and can produce multi-scale, multi-level feature responses and outputs and an M-CMN for manhole cover object detection based on fused feature maps. Compared with Faster R-CNN, YOLOv3, SSD, DSSD and FSSD, our DCNN network model can effectively improve the Precision and Recall rate in small manhole cover object detection. In future studies, we intend to focus on learning rotation invariant deep features for more types of object detection [24–26]. Furthermore, we will incorporate multi-GPU cluster computation to further reduce deep network model computation time.

Author Contributions: W.L. conceived and designed the experiments, D.C. and E.L. performed the experiments; P.Y. and L.Z. analyzed the results, M.Y. and M.X. contributed reagents/materials/analysis tools and W.L. and D.C. wrote the paper.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 41601405, in part by a grant from State Key Laboratory of Resources and Environmental Information System, in part by the Fund of Jiangsu Provincial Land and Resources Science and Technology Project (No. 2018054), in part by the Fund of Xuzhou Land and Resources Bureau Science and Technology Project (No. XZGTKJ2018001) and in part by the Fund of Xuzhou Science and Technology Key R & D Program (Social Development) Project (No. KC18139).

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

References

1. Yu, J.; Lin, W. The Application of High Precision Mobile Measurement System in Urban Component Survey. *Sci. Surv. Mapp.* **2016**, *8*, 147–148.
2. Li, X.; Tang, J.; Li, H. The Application of Mobile Mapping Technology in Digital Urban Component Census. *Lang Resour. Her.* **2017**, *14*, 53–57.
3. Song, Y.; Zeng, F.; Gao, Z. Application of vehicle panoramic photogrammetry in the investigation of urban parts. *Sci. Surv. Mapp.* **2016**, *11*, 40–43.
4. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
5. Ding, P.; Zhang, Y.; Deng, W.J.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv* **2014**, arXiv:1311.2524v4.
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef]
8. Hwang, J.J.; Liu, T.L. Pixel-wise Deep Learning for Contour Detection. *arXiv* **2015**, arXiv:1504.01989.
9. Zhao, F.; Xia, L.; Kylling, A.; Li, R.Q.; Shang, H.; Xu, M. Detection flying aircraft from Landsat 8 OLI data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 176–184. [CrossRef]
10. Zhong, Y.; Han, X.; Zhang, L. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 281–294. [CrossRef]
11. Girshick, R. Fast R-CNN. *arXiv* **2015**, arXiv:1504.08083.
12. 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. [CrossRef] [PubMed]
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June2016; pp. 779–788.

14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
15. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. *Int. J. Comput. Vis* **2015**, *125*, 3–18. [CrossRef]
16. Yan, Z.; Zhang, H.; Piramuthu, R.; Jagadeesh, V. HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2740–2748.
17. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1636. [CrossRef]
19. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. [CrossRef]
20. Lecun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **2014**, *1*, 541–551. [CrossRef]
21. Dollar, P.; Tu, Z.; Belongie, S. Supervised Learning of Edges and Object Boundaries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 1964–1971.
22. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
23. Li, Z.; Zhou, F. FSSD: Feature Fusion Single Shot Multibox Detector. *arXiv* **2017**, arXiv:1712.00960.
24. Cheng, G.; Han, J.; Zhou, P. Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. *IEEE Trans. Image Process.* **2018**, *28*, 265–278. [CrossRef]
25. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
26. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]