Fusion Object Detection with Convolutional Neural Network

Ying Ya, Han Pan*, Zhongliang Jing, Xuanguang Ren, Lingfeng Qiao

School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, 200240 China
E-mail: {yaying, hanpan, zljing, lightness, qiaolf927}@sjtu.edu.cn

Abstract: The availability of multi-source image by different sensors poses a serious challenge for object detection. However, the varied spatial structure by multi-source images make object detection difficult. Multi-model image fusion provides a possibility to improve the performance of object detection. In this paper, we propose a fusion object detection scheme with convolutional neural network. First, nine kinds of image fusion methods are adopt to fuse multi-source images. Second, a novel object detection frameworks with Faster RCNN (Region-based Convolutional Neural Network) structure is utilized, which suitable for large-scale satellite images. We use the Region Proposal Network (RPN) to generate axially aligned bounding boxes in different orientations, then extract features by pooling layers with different sizes. The features are used to classify the proposals, adjust the bounding boxes and predict the score. Smaller anchor for small objects is added. Finally, inclined non-maximum suppression method is utilized to get the detection results. Experimental results showed that the object detection method performs better than YOLO-v2, YOLO-v3 frameworks on satellite imagery and the proposed fusion object detection method has a significant improvement over object detection method with single image. Some numerical tests are reported to illustrate the efficiency of the proposed method.

Keywords: object detection, image fusion, satellite imagery, CNN

1 Introduction

Object detection is an important and challenging research hotspot in the field of computer vision and digital image processing. It is widely used in many fields, such as robot navigation, intelligent video surveillance, industrial detection, aerospace, etc. It is an important branch of image processing and computer vision, and also the core part of intelligent monitoring system. At the same time, object detection is also a basic algorithm of object recognition, which plays a vital role in subsequent recognition tasks. Because of the extensive application of deep learning, object detection algorithm has been developed rapidly. It has made great strides in the past few years since the convolutional neural networks (CNN)[1] method was used and won in the ImageNet competition[2] in 2012. Satellite images are important information resource of great significance to national security and economic and social development. Because of its practicability and timeliness, it is widely used in military reconnaissance, disaster monitoring, environmental monitoring, resource investigation, land use assessment, agricultural output estimation, urban construction planning, etc. It has important significance for national defense security, economic and social development. Although the deep learning methods perform well in the task of object detection of ground images, it is not easy to transfer this technology to satellite images. There are four main problems that the algorithm needs to satisfy. Firstly, the resolution of satellite images are ultra high, usually up to megapixels. Secondly, the objects such as ships, small vehicles, planes are extremely small and dense in satellite images, rather than the obvious large target objects in typical and common datasets such as PASCAL VOC[3] and ImageNet. Thirdly, there is a relative lack of public training datasets. And the problem of complete rotation invariance. Many target objects such as cars, ships and planes have lots of orientation when viewed from overhead. Among these problems, the primary problem and the biggest challenge is that the input images are enormous, while the objects are very small, which is a quite complex task for traditional computer vision technology.

There are many kinds of observation methods in satellite missions, such as RGB image, infrared image, hyperspectral image, multispectral image and SAR image. Multi-source images by different sensors are widely used in military and civilian fields. Detection and recognition of important objects can monitor the distribution of targets in key areas, analyze the enemy's operational strength, master operational intelligence at sea, and conduct precise guidance. In recent years, with the rapid development of earth observation technology, many optical remote sensing imaging satellites with high spatial resolution have emerged. Panchromatic images with sub-meter resolution can be obtained, which provides a very rich data source for space multi-source image object detection. However, because of the disadvantageous factors such as multi-view imaging, long shooting distance, cloud, haze occlusion, uneven illumination, brightness and colour differences, multiplicative noise, etc. It is easy to cause false alarm and missed detection. How to detect and extract objects accurately, quickly and steadily and gain more response and processing time based on multi-source image information on satellite has become an urgent problem to be solved.

In this paper, we propose a fusion object detection scheme with convolutional neural network. Multi-model image fusion provides a possibility to improve the performance of object detection. First, nine kinds of image fusion methods are adopt to fuse multi-source images. Second, a novel object detection frameworks with Faster RCNN structure is utilized, which suitable for large-scale satellite images. To detect objects in any orientation, the bounding boxes are in different orientations, and then extract features by pooling layers with different sizes. The features are used to classify the proposals, adjust the bounding boxes and predict the text score. Smaller anchor for small objects is added. Finally, inclined non-maximum suppression method is utilized to get the detection results.

Section 2 introduces related work about object detection and image fusion. Section 3 part 1 describes the image fusion methods we utilized and part 2 details our object detection method for satellite imagery. Section 4 describes the datasets we used in our experiments. Finally, in Section 5, the evaluation indicators are introduced and the experimental results of our algorithm are showed and discussed in detail.

2 Related Work

2.1 Deep Learning

Hinton et al. first proposed Deep Neural Network[4] as the representative of deep learning technology in 2006, which attracted the attention of academia. Bengio, LeCun et al. followed up the relevant research, which opened the upsurge of deep learning research. Convolutional Neural Network (CNN) is a deep neural network with convolution structure, sparse connection and weight sharing. Its characteristics can reduce the scale of parameters of the neural network, reduce the complexity of model training, and avoid the cumbersome feature extraction and data reconstruction in traditional algorithms. At the same time, convolution preserves the spatial information of image pixels, and has the invariance of translation, rotation and scale. When multi-dimensional images are directly input into the network, this advantage is more obvious. In 1989, LeNet-5[5], a CNN model, was proposed by LeCun et al. for handwritten characters recognition. This method achieved satisfactory results. In 2012, AlexNet[6], a CNN model constructed by Krizhevsky et al. greatly reduced the error rate in image classification of ImageNet large-scale visual recognition challenge competition, refreshed the record of image classification, and established the position of deep learning in computer vision. Deep learning uses multi-layer computing model to learn abstract data representation from complex structures in numerous data. This technology has been successfully applied to many pattern classification problems, including computer vision.

2.2 Object Detection

The analysis of object motion in computer vision can be roughly divided into three levels: motion segmentation, object detection, object tracking, action recognition and behaviour description[7]. Object detection is not only the basic tasks in the field of computer vision, but also the basic task of video surveillance technology. Because the objects in video have different attitudes and often occlude, and their motion is irregular, and considering the conditions of depth of field, resolution, weather, illumination and the diversity of scene, the results of object detection algorithm will directly affect the follow-up tracking, action recognition and behaviour description. Therefore, even in today's technological development, the basic task of object detection is still a very challenging subject, which has great potential and space for improvement.

Traditional object detection methods usually use shallow trainable architectures and handcrafted features. Object detection including two sub-tasks: object location to determine where objects are located in given images and object classification to determine which category the objects belong to. The pipeline of traditional object detection are generally divided into three stages: informative region selection, feature extraction and classification. But the performance of traditional methods is not good when constructing complex ensembles which combine multiple low-level image features with high-level context from object detectors and scene classifiers[8]. With the rapid development in deep learning, an obvious gain is achieved. Girshick R et al. proposed Regions with CNN (R-CNN)[9] features. Deep learning methods have the capacity to learn more complex features than the traditional methods and learn informative object representations rather than design features manually[10].

In the domain of deep learning-based object detection, three of the best rapid object detection schemes are: Faster R-CNN[11], SSD[12], and YOLO[13][14][15]. Faster R-CNN uses 1000 × 600 pixel input images, SSD runs on 300 × 300 or 512 × 512 pixel inputs, and YOLO ingests 416 × 416 or 544 × 544 pixel images. These frameworks have good performance, but it is difficult to process satellite imagery[16].

Faster R-CNN is a typical model. It inspired numerous detection and segmentation models that came after it. R-CNN first proposes regions, then extracts features, and then classifies those regions based on their features. It was intuitive, but the speed is very slow. R-CNN’s immediate descendant was Fast R-CNN[17]. Fast R-CNN performed much better in terms of speed. But the selective search algorithm for generating region proposals was a big bottleneck. Faster R-CNN’s main insight was to replace the slow selective search algorithm with a fast neural net and introduced the Region Proposal Network (RPN).

Most object detection methods are detecting horizontal bounding boxes, J. Ma et al. proposed Rotation Region Proposal Network (RRPN)[18] to detect arbitrary-oriented scene text which based on Faster R-CNN. An algorithm called Rotational Region CNN[19] based on RRPN can detect arbitrary-oriented texts in natural scene images, our goal is to detect small objects in any orientation, therefore our work used the Rotational Region CNN framework and added smaller anchor for small objects in satellite imagery.

2.3 Image Fusion

Fusing multi-band images has became a thriving area of research in a number of different fields, such as space robotics, remote sensing, etc. Multi-band image fusion[20][21][22][23][24] aims to combine spatial and spectral information from one or multiple observations and other image sources, such as panchromatic images, multispectral images or hyper-spectral images. Pansharpening aims at fusing a multispectral and a panchromatic image, featuring the result of the processing with the spectral resolution of the former and the spatial resolution of the latter[25].

Panchromatic (PAN) and multispectral (MS) image fusion method originated in 1980s[26][27]. Since SPOT-1 satellite provided panchromatic and multispectral images simultaneously in 1986, fusion methods have developed rapidly. Generally, fusion methods can be classified into three categories[28]: component replacement fusion methods, multi-resolution analysis fusion methods and model-based fusion methods. Among them, the component replacement method is the simplest and most popular fusion method, which has been widely used in professional remote sensing software such as ENVI and ERDAS. First, the luminance component is obtained based on spectral transformation, and then the spatial information of multispectral image is enhanced by replacing the luminance component with panchromatic image. Typical methods include Principal Component Analysis (PCA) fusion[29], Gram-Schmidt (GS) fusion[30], Intensity-Hue-Saturation (IHS) fusion[31] etc. Multi-resolution analysis fusion method extracts high spatial structure information of panchromatic image based on wavelet transform or Laplacian pyramid, and injects the extracted spatial structure information into multispectral image to obtain high spatial resolution fusion image with a certain injection model[32], such as multi-hole wavelet fusion method[33], Laplace pyramid fusion method[34] and Contourlet wavelet fusion method[35]. For component substitution fusion method and multi-resolution analysis fusion method, Tu et al[36] further extended them to the same fusion framework, which greatly promoted the development of panchromatic/multi-spectral fusion method.

3 Proposed Approach

3.1 Object Detection Framework

In the field of deep learning-based object detection, Faster R-CNN is a canonical model. The original model of Faster R-CNN is R-CNN, it worked as:

- Selective Search: scan the input image for possible objects, and generate about 2000 region proposals.
- Run a CNN on top of each of these region proposals.
- Feed the output took from each CNN into an SVM to classify the region and a linear regressor to tighten the bounding box of the object if exists.

The speed of R-CNN is very slow. Fast R-CNN improved the detection speed through performing feature extraction over the image before proposing regions. Instead of running 2000 CNN's over 2000 overlapping regions, it running only one CNN over the entire image. And using softmax layer to replace the SVM. But the selective search algorithm for generating region proposals is a bottleneck problem. The main contribution of Faster R-CNN was to replace the slow selective search algorithm with a fast neural net. Specifically, it introduced the Region Proposal Network (RPN). It worked as follows:

- At the last layer of an initial CNN, a 3x3 sliding window moves across the feature map and maps it to a lower dimension.
- For each sliding-window location, it generates multiple possible regions based on k fixed-ratio anchor boxes.
- Each region proposal consists of an "objectness" score for that region and 4 coordinates representing the bounding box of the region.

In a sense, Faster R-CNN = RPN + Fast R-CNN[37]. SSD and YOLO etc. object detection frameworks do not rely on region proposal, they estimate object candidates directly. Our object detection architecture in this paper named Rotational Region CNN [19] is based on the Faster R-CNN, utilizing the object candidates proposed by RPN to predict the orientation information, its network architecture shows in Figure 1. The RPN is used for proposing axis-aligned bounding boxes that enclose the arbitrary-oriented objects. For each box generated by RPN, 3 different pooled sizes ROI poolings are performed and the pooled features are concatenated for predicting the objects scores, axis-aligned box and inclined minimum area box. Then an inclined non-maximum suppression is conducted on the inclined boxes to get the final results.

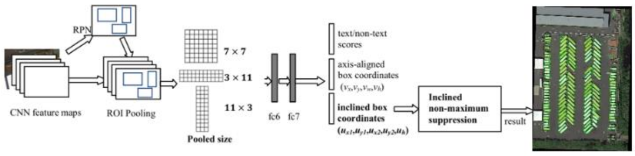


Fig.1 The network architecture of Rotational Region CNN.

The objects in satellite imagery are very small, therefore smaller anchors are added in RPN. The anchor aspect ratios and other settings of RPN are same as Faster R-CNN. The 3

ROI Poolings' pooled sizes are: 7 × 7, 3 × 11, 11 × 3, which can obtain more horizontal and vertical features on small scale objects. We estimate both the axis-aligned bounding box and the inclined bounding box, therefore we not only do normal NMS on axis-aligned bounding boxes and but also do inclined NMS on inclined bounding boxes [19], the method of calculating Intersection-over-Union (IoU) refer to [38]. The loss function in the training process is same as Faster R-CNN, while the loss defined on each proposal is different, it includes the object/non-object classification loss and the box regression loss as:

$$\begin{aligned} L(p,t,v,v^*,u,u^*) = L_{cls}(p,t) \\ +\lambda_1 t \sum\nolimits_{i\in\{x,y,w,h\}} L_{reg}(v_i,{v_i}^*) \\ +\ \lambda_2 t \sum\nolimits_{i\in\{x1,y1,x2,y2,h\}} L_{reg}(u_i,{u_i}^*)\ \ (1) \end{aligned}$$

in which λ_1, λ_2 parameters balance the trade-off between three subformulas and t is the indicator of the class label (object: $t = 1$, background: $t = 0$). $p = (p_0, p_1)$ is the parameter means the probability over object and background computed by the softmax function. $L_{cls}(p,t) = -log\ p_t$ is the log loss for true class t. v and u are tuples of true axis-aligned and true inclined bounding box regression targets. v^* and u^* are the predicted tuples for the object label. The parameterization for v and v^* is given in [39], in this paper, v and v^*specify a scale-invariant translation and log-space height/width shift relative to an object proposal. We use (w, w^*) indicates $(v_i, {v_i}^*)$ or $(u_i, {u_i}^*)$, $L_{reg}(w, w^*)$ is defined as:

$$L_{reg}(w,w^*) = \mathrm{smooth}_{\mathrm{L1}}(w-w^*) \qquad (2)$$

$$\mathrm{smooth}_{\mathrm{L1}}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \qquad (3)$$

3.2 Pansharpening Algorithms

Multi-model image fusion provides a possibility to improve the performance of object detection. Pansharpening aims at fusing a panchromatic (PAN) and a multispectral (MS) image simultaneously acquired over the same area. MS has fewer spatial details while PAN only has single band. Pansharpening can combine the spatial details resolved by the PAN image and the several spectral bands of the MS in a unique product.

Pansharpening methods usually be divided into two main classes, the component substitution (CS) methods and the multi-resolution analysis (MRA) methods. First, the notation used in this paper are described in Table 1:

Table 1: List of the Main Symbols

Symbol	Description
MS	Multispectral image
$\widetilde{\mathrm{MS}}$	MS image interpolated at the scale of PAN
P	PAN image
$\widehat{\mathrm{MS}}$	Pansharpened image
R	Spatial resolution ratio between MS and PAN
N	Number of MS bands

Vectors in this paper are expressed in bold lowercase (e.g., $\mathbf{x}$), x_i indicates the i th element. 2-dimensional and 3-dimensional arrays are indicated in bold uppercase (e.g., **X**). We use a 3-D array $\mathbf{X} = \{\mathbf{X}_k\}_{k=1,\dots,N}$ to indicate an MS image composed by N bands indexed by the subscript $k = 1,\dots,N$; and $\mathbf{X}_k$ indicates the kth band of **X**. A PAN image is a 2-D matrix and will be expressed as **Y**.

A typical formulation of CS fusion is given by

$$\widehat{MS_k} = \widetilde{MS_k} + g_k(\mathbf{P} - \mathbf{I}_L), \qquad k = 1,\dots,N \qquad (4)$$

in which k indicates the kth spectral band, the vector of the injection gains indicated as $\mathrm{g} = [g_1,\dots,g_k,\dots,g_N]$, and $\mathbf{I}_L$ is defined as

$$\mathbf{I}_L = \sum_{i=1}^{N} w_i\, \widetilde{MS}_\imath \qquad (5)$$

in which the weight vector $\mathbf{w} = [w_1,\dots,w_i,\dots,w_N]$ is the first row of the forward transformation matrix and can measure the degrees of spectral overlap among the MS channels and PAN[40][41].

The pipeline of CS approach: First, to match the scale of PAN, interpolate the MS image; then, calculate the intensity component using formula (5) and match the histograms of the PAN and the intensity component; finally, inject the extracted details by (1).

The CS approaches include many pansharpening methods, in the following, they will be introduced in detail. Table 2 summarize the values of the spectral weights and injection gains by (5) and (4). In $w_{k,i}$, subscripts k and i refer to output and input bands respectively.

Table 2: Spectral Weight in (5) and Injection Gains in (4) for several CS-Based Methods

Method	$\mathrm{w}_{k,i}$	g_k
BT[42]	$^1/_N$	$\frac{\widetilde{MS_k}}{I_L}$
PCA[43]	$\mathbf{X}_{1,i}$	$\mathbf{X}_{1,k}$
GS[43]	$^1/_N$	$\frac{cov(I_L,\widetilde{MS_k})}{var(I_L)}$
GSA[44]	$\widehat{w_\imath}(Eq.6)$	$\frac{cov(I_L,\widetilde{MS_k})}{var(I_L)}$
BDSD[45]	$\widehat{w_{k,\imath}}(Eqs.(8)-(9))$	$\widehat{g_k}(Eqs.(8)-(9))$
PRACS[46]	$\widehat{w_\imath}(Eq.6)$	$(Eq.(11)-(12))$
IHS	$^1/_N\ (N=3)$	1
GIHS[47]	any $w_i \geq 0$	$(\sum\nolimits_{i=1}^{N} w_i)^{-1}$

1) PCA

PCA is achieved through a multidimensional rotation of the original coordinate system of the N-dimensional vector space, i.e., a linear transformation of the data, such that the projection of the original spectral vectors on the new axes, which are the eigenvectors of the covariance matrix along the spectral direction, produces a set of scalar images, called principal components (PCs), that are uncorrelated to each other. PCs are generally sorted for decreasing variance, which quantifies their information content.

2) GS

The GS transformation is a usual technique used in linear algebra and multivariate statistics to orthogonalize a set of vectors. GS orthogonalization proceeds one MS vector at the time, by finding its projection on the (hyper) plane defined by the previously found orthogonal vectors and its orthogonal component, such that the sum of the orthogonal and projection components is equal to the zero-mean version of the original vectorized band. Pansharpening is accomplished by replacing $\mathbf{I}_L$ with the histogram-matched $\mathbf{P}$ before the inverse transformation is performed [25]. B. Aiazzi, et al. proposed adaptive GS (GSA) in [43], in which $\mathbf{I}_L$ is generated by a weighted average of the MS bands, with MSE-minimizing weights by a low-pass-filtered version of PAN:

$$\mathbf{P}_L = \sum_{k=1}^{N} w_k \widetilde{MS_k} \tag{6}$$

3) BDSD

The Band-Dependent Spatial Detail (BDSD) algorithm [45] starts from an extended version of the generic formulation (4) as follows:

$$\widehat{MS_k} = \widetilde{MS_k} + g_k \left(\mathbf{P} - \sum_{i=1}^{N} w_{k.i} \widetilde{MS_\imath} \right), \qquad k = 1, \dots, N. \tag{7}$$

The coefficients is defined as:

$$\gamma_{k,i} = \begin{cases} g_k & if\ i = N+1 \\ -g_k \cdot w_{k.i} & otherwise \end{cases} \tag{8}$$

equation (1) can be rewritten in compact matrix form as:

$$\widehat{MS_k} = \widetilde{MS_k} + \mathbf{H}\gamma_k \tag{9}$$

in which $\mathbf{H} = \left[\widetilde{MS}_1, \dots, \widetilde{MS_N}, \mathbf{P}\right], \gamma_{k,i} = [\gamma_{k,1}, \dots, \gamma_{k,N+!}]^T$.

4) PRACS

The concept of partial replacement of the intensity component is described in [46] named Partial Replacement Adaptive CS (PRACS). This method utilizes $\mathbf{P}^{(k)}$, a weighted sum of PAN and of the k th MS band, to calculate the kth sharpened band in (4). For $k = 1, \dots, N$, the band-dependent high-resolution sharpening image is calculated as:

$$\mathbf{P}^{(k)} = \mathrm{CC}\left(\mathbf{I}_L, \widetilde{MS_k}\right) \cdot \mathbf{P} + \left(1 - \mathrm{CC}\left(\mathbf{I}_L, \widetilde{MS_k}\right)\right) \cdot \widetilde{MS_k}' \tag{10}$$

The injection gains $\{g_k\}$ are obtained by

$$g_k = \beta \cdot \mathrm{CC}\left(\mathbf{P}_L^{(k)}, \widetilde{MS_k}\right) \frac{std\left(\widetilde{MS_k}\right)}{\frac{1}{N}\sum_{i=1}^{N} std\left(\widetilde{MS_\imath}\right)} L_k. \tag{11}$$

L_k is defined as:

$$L_k = 1 - \left| 1 - \mathrm{CC}\left(\mathbf{I}_L, \widetilde{MS_k}\right) \frac{\widetilde{MS_k}}{\mathbf{P}_L^{(k)}} \right| \tag{12}$$

The pipeline of the MRA approach: First, interpolate the MS image to reach the PAN scale; then calculate the low-pass version $\mathbf{P}_L$ of the PAN by means of the equivalent filter for a scale ratio equal to R and compute the band-dependent injection gains $\{g_k\}_{k=1,\dots,N}$; finally, inject the extracted details by (13). A brief summary of the MRA-based approaches is showed in Table 3.

$$\widehat{MS_k} = \widetilde{MS_k} + g_k(\mathbf{P} - \mathbf{P}_L), \mathrm{k} = 1, \dots, \mathrm{N}. \tag{13}$$

Table 3: MRA-Based Pansharpening Methods and Related MRA Schemes With Filters and Injection Gains

Method	Type of MRA and filter	g_k
HPF[29]	ATWT w/ Box Filter	1
HPM[48]/SFIM[49][50]	ATWT w/ Box Filter	$\frac{\widetilde{MS_k}}{p_L}$
Indusion[51]	DWT w/ CDF Bior. Filt.	1
MTF-GLP[52]	GLP w/ MTF Filter	1
MTF-GLP-CBD	GLP w/ MTF Filter	$\frac{cov(P_L, \widetilde{MS_k})}{var(P_L)}$
MTF-GLP-HPM[53]	GLP w/ MTF Filter	$\frac{\widetilde{MS_k}}{p_L}$
MTF-GLP-HPM-PP	GLP w/ MTF Filter	$\frac{\widetilde{MS_k}}{p_L}$

5) Low-Pass Filtering (LPF)

An implementation of applying a single linear time-invariant LPF h_{LP} to the PAN image $\mathbf{P}$ to obtain $\mathbf{P}_L$ is given by (14), the notation * presents the convolution operator.

$$\widehat{MS_k} = \widetilde{MS_k} + g_k(\mathbf{P} - \mathbf{P} * h_{LP}), k = 1, \dots, N \tag{14}$$

6) Pyramidal Decompositions

This method is commonly referred to as pyramidal decomposition utilizing Gaussian LPFs to carry out the analysis steps. The Gaussian filters can be tuned to match the sensor MTF closely, and allow extracting from the PAN those details, which are not seen by the MS sensor because of the coarser spatial resolution [25].

4 Data Sets

4.1 Object Detection Training Data

The training dataset we utilized in our object detection algorithm is DOTA[54]. It is a large-scale dataset for object detection in aerial images. It can be used to train and evaluate object detectors in aerial images. DOTA contains 2806 images from different sensors and platforms. There are 15 categories in total in DOTA dataset, including large vehicle, small vehicle, plane, helicopter, ship, harbour, bridge, baseball diamond, basketball court, soccer ball field, tennis court, ground track field, roundabout, basketball court and storage tank. The size of each image range from about 800 × 800 to 4000 × 4000 pixels. The objects in DOTA have a wide variety of scales, orientations, and shapes.

There are 188, 282 instances of the fully annotated DOTA images annotated by an arbitrary quadrilateral. Figure 2 shows examples of annotated DOTA images.

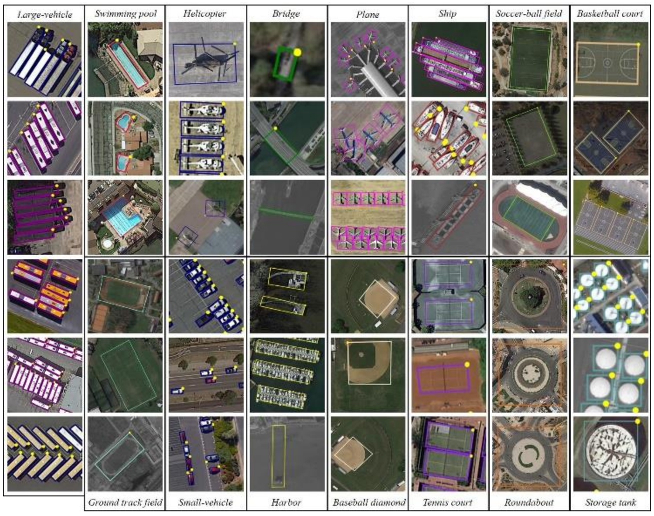


Fig.2 Examples of annotated DOTA images.

4.2 Lateral-directional control architecture

For image fusion, we use three data sets:

1) Pléiades data set

The Pléiades data set was used for the 2006 contest[55], which was collected by an aerial platform and provided by CNES, the French Space Agency. The images are an urban area of Toulouse (France) with the size of 1024 × 1024 pixels. The resolution of the 4 MS bands is 0.6 meter. The PAN data with high-resolution were simulated by the following procedure. The red and green channels were averaged, and the result was filtered with a system characterized by the nominal MTF of the PAN sensor. After the resampling to 0.8 meter, which adding thermal noise. Finally, inverse filtering and wavelet denoising was used to obtain simulated image [25].

2) Kaggle Dstl Satellite Imagery Feature Detection Competition Data

In this competition[56], Dstl provides you with 1000m × 1000m satellite images in both 3-band and 16-band formats. The 3-band images are the traditional RGB natural colour images. The 16-band images contain spectral information by capturing wider wavelength channels. This multi-band imagery is taken from the multispectral (400 – 1040nm) and short-wave infrared (SWIR) (1195-2365nm) range.

3) 2019 IEEE GRSS Data Fusion Contest Data

In the contest[57], they provide Urban Semantic 3D (US3D) data, a large-scale public dataset including multi-view, multi-band satellite images and ground truth geometric and semantic labels for two large cities. The US3D dataset includes incidental satellite images, airborne lidar, and semantic labels covering approximately 100 square kilometres over Jacksonville, Florida and Omaha, Nebraska, United States. WorldView-3 panchromatic and 8-band visible and near infrared (VNIR) images are provided courtesy of Digital Globe. Source data consists of 26 images collected between 2014 and 2016 over Jacksonville, Florida, and 43 images collected between 2014 and 2015 over Omaha, Nebraska, United States. Ground sampling distance (GSD) is approximately 35 cm and 1.3 m for panchromatic and VNIR images, respectively. VNIR images are all pan-sharpened. Satellite images are provided in geographically non-overlapping tiles, where Airborne LiDAR data and semantic labels are projected into the same plane.

5 Experiments

5.1 Object Detection Performance

We evaluated our object detection approach on DOTA and compared with YOLO-v2 and YOLO-v3 framework. The evaluation images are clipped to 800×800 pixels. The evaluation indicators are Precision, Recall, F1-Measure, AP (Average Precision) and mAP (mean average precision) defined as:

TP: True Positives
FP: False Positives
FN: False Negatives
TN: True Negatives

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{F1} - \text{Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

The area enclosed by P-R curve is the value of AP and mAP is the mean value of AP. The results are obtained by the evaluation code provided for VOC dataset from GitHub. Table 4~7 are Precision, Recall, F1-Measure, AP and mAP of YOLO-v2, YOLO-v3 framework and our method in horizontal and rotational orientations.

The test speed of our experiment is in Tabel 8, which is obtained on a NVIDIA GeForce GTX 1080Ti GPU.

Table 4: Precision of YOLO-v2, YOLO-v3 and Our Method.

Precision **Classes**	YOLO -v2	YOLO -v3	horizon of our method	rotation of our method
tennis-court	0.2454	0.2686	**0.9447**	0.9446
harbour	0.1293	0.0845	**0.7694**	0.6977
bridge	0.0803	0.0122	**0.5876**	0.5140
plane	0.3453	0.1752	0.8857	**0.9101**
ship	0.2253	0.2054	**0.7928**	0.6353
ground-track-field	0.2184	0.0105	0.6447	**0.6527**
large-vehicle	0.2003	0.1391	**0.7871**	0.6765
helicopter	0.0283	0.0049	**0.5000**	0.4891
basketball-court	0.0977	0.0164	0.6193	**0.6444**
roundabout	0.1117	0.0433	0.6498	**0.6752**
small-vehicle	0.2601	0.1001	**0.5795**	0.5390
storage-tank	0.2631	0.2137	0.8213	**0.8293**
soccer-ball-field	0.1588	0.0172	0.6000	**0.6063**
swimming-pool	0.0062	0.0201	**0.5849**	0.5654
baseball-diamond	0.2694	0.0179	0.6964	**0.7143**

Table 5: Recall of YOLO-v2, YOLO-v3 and Our Method.

Recall	YOLO	YOLO	horizon of	rotation of
Classes	-v2	-v3	our method	our method
tennis-court	0.7422	0.7427	0.7603	**0.8669**
harbour	0.5761	0.4508	**0.6221**	0.6166
bridge	0.3545	0.1377	**0.3716**	0.3284
plane	0.6585	0.4256	0.7577	**0.8249**
ship	0.6707	**0.7492**	0.5240	0.5247
ground-track-field	0.2387	0.1080	0.5653	**0.5854**
large-vehicle	0.4116	**0.6943**	0.4237	0.5052
helicopter	0.1888	0.0153	**0.4745**	0.4592
basketball-court	0.4278	0.1813	0.3824	**0.4363**
roundabout	**0.6571**	0.1918	0.4940	0.5084
small-vehicle	0.3509	**0.8190**	0.4669	0.4719
storage-tank	**0.5073**	0.2969	0.4593	0.4639
soccer-ball-field	0.2482	0.1106	0.4054	**0.4275**
swimming-pool	0.0091	0.0751	**0.5659**	0.5527
baseball-diamond	0.4566	0.1030	0.6626	**0.6768**

Table 6: F1-Measure of YOLO-v2, YOLO-v3 and Our Method.

F1-M	YOLO	YOLO	horizon of	rotation of
Classes	-v2	-v3	our method	our method
tennis-court	0.3688	0.3945	0.8425	**0.9041**
harbour	0.2112	0.1424	**0.6879**	0.6547
bridge	0.1309	0.0223	**0.4553**	0.4008
plane	0.4531	0.2482	0.8167	**0.8654**
ship	0.3373	0.3224	**0.6310**	0.5747
ground-track-field	0.2281	0.0192	0.6024	**0.6172**
large-vehicle	0.2694	0.2317	0.5509	**0.5784**
helicopter	0.0492	0.0074	**0.4869**	0.4737
basketball-court	0.1591	0.0300	0.4729	**0.5203**
roundabout	0.1909	0.0706	0.5613	**0.5800**
small-vehicle	0.2987	0.1783	**0.5171**	0.5032
storage-tank	0.3465	0.2485	0.5891	**0.5949**
soccer-ball-field	0.1937	0.0297	0.4839	**0.5014**
swimming-pool	0.0074	0.0317	**0.5753**	0.5590
baseball-diamond	0.3388	0.0305	0.6791	**0.6950**

Table 7: AP and mAP of YOLO-v2, YOLO-v3 and Our Method.

AP	YOLO	YOLO	horizon of	rotation of
Classes	-v2	-v3	our method	our method
tennis-court	0.5498	0.6996	0.7591	**0.8590**
harbour	0.3259	0.2297	**0.5779**	0.5310
bridge	0.1840	0.1019	**0.2874**	0.2345
plane	0.5841	0.3443	0.7502	**0.8151**
ship	0.4905	**0.5619**	0.4918	0.4252
ground-track-field	0.2004	0.0694	0.4870	**0.5051**
large-vehicle	0.2241	**0.4181**	0.3941	0.3985
helicopter	0.1591	0.0182	**0.4208**	0.3757
basketball-court	0.2817	0.1483	0.3416	**0.3830**
roundabout	0.4416	0.0492	0.4446	**0.4586**
small-vehicle	0.2235	0.3535	**0.3643**	0.3466
storage-tank	0.4113	0.2002	0.4453	**0.4511**
soccer-ball-field	0.2150	0.0937	0.3534	**0.3751**
swimming-pool	0.0007	0.0027	**0.4551**	0.4416
baseball-diamond	0.3250	0.0081	0.5943	**0.6199**
mAP	0.3078	0.2199	0.4778	**0.4813**

Table 8: Speed of YOLO-v2, YOLO-v3 and Our Method.

	YOLO-v2	YOLO-v3	our method
Speed (fps)	25.8771	5.9878	17.5131

From the tables above, we can easily draw a conclusion that our object framework performs better on satellite imagery than YOLO-v2 and YOLO-v3. It achieved competitive results of Precision, Recall, F1-Measure and AP, the mAP of our method is 17.4% higher than that of YOLO-v2, and 26.1% higher than that of YOLO-v3. The model we used performed extremely well on helicopters, ports and swimming pools.

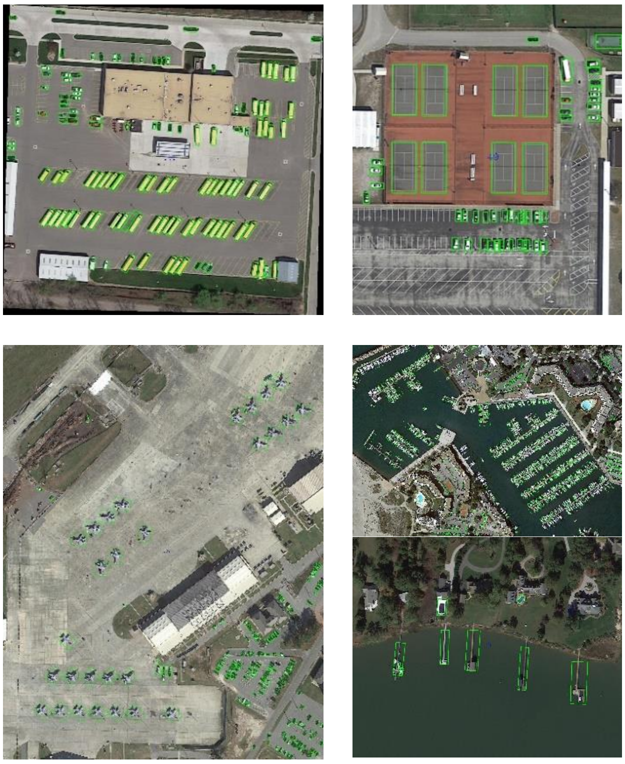
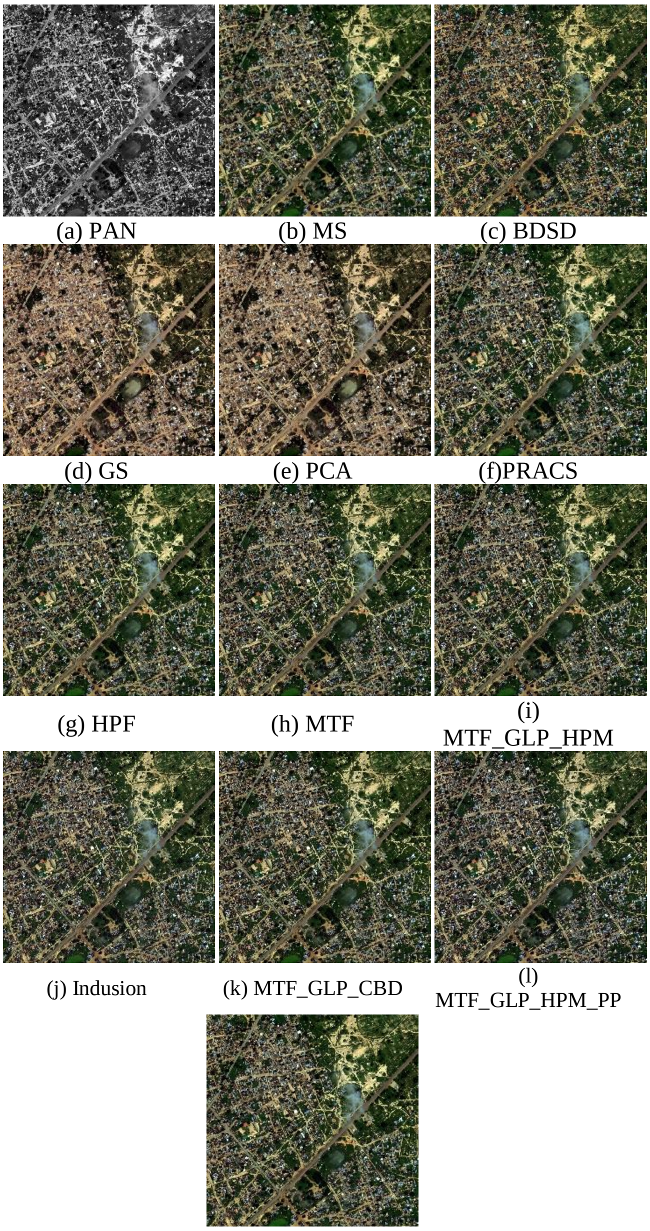


Fig.3 Some detection results of our CNN framework on DOTA dataset.

Our approach can detect arbitrary-oriented small objects on DOTA dataset demonstrated in Figure 3. Even if the image resolution is as high as 3000×4000, small objects with only15 pixels can be detected.

5.2 Image Fusion Results

Three datasets mentioned in 4.2 are utilized to evaluate the results of image fusion method mentioned in Section 3.2. The results of Kaggle Dstl Satellite Imagery Feature Detection Competition Data are shown in Figure 4. (a) is PAN image and (b) is MS image with 16 bands, the fusion method of (c) (d) (e) (f) belong to CS families and others belong to MRA families. From the results of our image fusion experiment, we draw a conclusion that CS methods have a higher spectral distortion but the final products have better visual appearance, while MRA methods have a higher spatial distortion, but the spectral consistency is better.



(a) PAN (b) MS (c) BDSD

(d) GS (e) PCA (f)PRACS

(g) HPF (h) MTF (i) MTF_GLP_HPM

(j) Indusion (k) MTF_GLP_CBD (l) MTF_GLP_HPM_PP

(m) SFIM

Fig.4 The fusion results of Kaggle Dstl Satellite Imagery Feature Detection Competition Data.

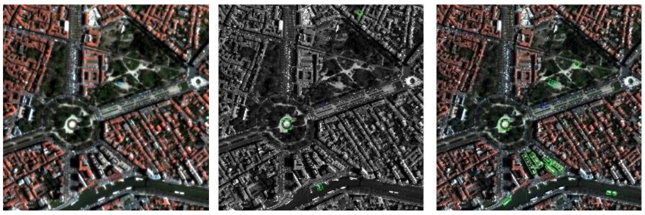
5.3 Fusion Object Detection Performance

We utilize PAN image, MS image and the images fused by the methods have mentioned as the CNN object detection framework's inputs, and statistics results on three datasets. We use the number of detected objects to evaluate the experiment performance showed in Table9.

We utilize 15 images from the three datasets. It is obvious that the number of objects detected by fused images is much more than that of PAN and MS images. And some detected objects in low resolution MS images and single band PAN images are not true objects as Figure 5. None object is detected in MS image, 3 objects are detected in PAN image and 10 objects are detected in fused image by MTF_GLP_HPM_PP method. The performance of object detection task on satellite imagery has improved obviously through proposed fusion object detection method.

Table 9: The Number of Detected Objects.

method	number of detected objects
PAN	13
MS	5
BDSD	85
GS	21
PCA	37
PRACS	145
HPF	85
MTF	114
MTF_GLP_HPM	111
MTF_GLP_HPM_PP	98
MTF_GLP_CBD	108
Indusion	120
SFIM	82



MS PAN MTF_GLP_HPM_PP

Fig.5 The detection performance on MS, PAN and MTF_GLP_HPM_PP images.

6 Conclusion

Object detection algorithm has been developed rapidly and made great strides in the past few years since the popularity of CNN. But it is not easy to transfer this technology to satellite images. A fusion object detection scheme with convolutional neural network we proposed in this paper. Experimental results showed that the object detection method performs better than YOLO-v2, YOLO-v3 frameworks on satellite imagery in Precision, Recall and mAP, etc. And the proposed fusion object detection method has a significant improvement over object detection method with single image. The efficiency of multi-model image fusion on object detection task has been proved in this paper. Next, we will try to utilize the fused images as the training set of the detection scheme and verify the efficiency.

Acknowledgment

This work is jointly supported by National Natural Science Foundation of China (Grant Nos. 61603249, 61673262), and key project of Science and Technology Commission of Shanghai Municipality (Grant No. 16JC1401100).

I would like to show my deepest gratitude to my supervisor, Dr. Zhongliang Jing and Dr. Han Pan who have provided me with valuable guidance in every stage of the research. I shall extend my thanks to Mr. Ren and Mr. Qiao for their kindness and help. I would also like to thank my families and my friend Dan Wu who have helped and supported all the time.

References

[1] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.

[2] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge [J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[3] Everingham M, Winn J. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit [J]. International Journal of Computer Vision, 2006, 111(1):98-136.

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[5] 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 Computation, 1989, 1(4): 541−51.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: MIT Press, 2012, 1097−1105.

[7] Vishwakarma S, Agrawal A. A survey on activity recognition and behavior understanding in video surveillance [J]. The Visual Computer, 2012: 1-27.

[8] Zhao Z Q , Zheng P , Xu S T , et al. Object Detection with Deep Learning: A Review[J]. 2018.

[9] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2014.

[10] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[11] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. 2015.

[12] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2015.

[13] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. 2015.

[14] Redmon, J, Farhadi A. [IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - YOLO9000: Better, Faster, Stronger[J]. 2017:6517-6525.

[15] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [J]. 2018.

[16] Van Etten A. You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery [J]. 2018.

[17] Girshick R. Fast R-CNN [J]. Computer Science, 2015.

[18] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. arXiv preprint arXiv:1703.01086, 2017.

[19] Jiang Y, Zhu X, Wang X, et al. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection [J]. 2017.

[20] S. Chaudhuri and K. Kotwal, Hyperspectral image fusion. Springer, 2013.

[21] E. M. Middleton, S. G. Ungar, D. J. Mandl, L. Ong, S. W. Frye, P. E. Campbell, D. R. Landis, J. P. Young, and N. H. Pollack, “The earth observing one (eo-1) satellite mission: Over a decade in space,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 6, no. 2, pp. 243-256, 2013.

[22] Z. Jing, H. Pan, and G. Xiao, Application to Environmental Surveillance: Dynamic Image Estimation Fusion and Optimal Remote Sensing with Fuzzy Integral. Cham: Springer International Publishing, 2015, pp. 159-189. [Online]. Available: https://doi.org/10.1007/978-3-319-12892-4_7

[23] J. Zhongliang, P. Han, L. Yuankai, and D. Peng, Non-Cooperative Target Tracking, Fusion and Control: Algorithms and Advances. Springer International Publishing, 2018.

[24] H. Pan, Z. Jing, L. Qiao, and M. Li, “Visible and infrared image fusion using l0-generalized total variation model,” Science China Information Sciences, vol. 61, no. 4, p. 049103, 2018.

[25] Vivone G, Alparone L, Chanussot J, et al. A critical comparison among pansharpening algorithms [J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 53(5):2565-2586.

[26] Shen H F, Meng X C, Zhang L P .An integrated framework for the spatio-temporal-spectral fusion of remote sensing images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12):7135-7148.

[27] Zhang L P, Shen H F .Progress and future of remote sensing data fusion [J]. Journal of Remote Sensing, 2016, 20(5):1050-1061.

[28] Aiazzi B, Alparone L, Baronti S , et al. Twenty-five years of pansharpening: A critical review and new developments[M] //Chen C H.Signal and Image Processing for Remote Sensing. 2nd edition. Boca Raton, FL: CRC Press, 2012: 533-548.

[29] Chavez Jr P S, Sides S C, Anderson J A .Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic [J]. Photogrammetric Engineering and Remote Sensing, 1991, 57(3):295-303.

[30] Laben C A, Brower B V .Process for enhancing the spatial resolution of multispectral imagery using pansharpening: United States, 6011875[P]. 2000 -01-04.

[31] W. Carper, T. Lillesand, and R. Kiefer, “The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data,” Photogramm. Eng. Remote Sens., vol. 56, no. 4, pp. 459–467, Apr. 1990.

[32] Meng X C, Li J, Shen H F, et al. Pansharpening with a guided filter based on three-layer decomposition [J]. Sensors, 2016, 16(7):1068.

[33] Ranchin T, Wald L .Fusion of high spatial and spectral resolution images:The ARSIS concept and its implementation [J]. Photogrammetric Engineering and Remote Sensing, 2000, 66(1):49-61.

[34] Alparone L, Aiazzi B .MTF-tailored multiscale fusion of high-resolution MS and Pan imagery [J]. Photogrammetric Engineering and Remote Sensing, 2006, 72(5):591-596.

[35] Li W J, Wen W P, Wang Q H .A study of remote sensing image fusion method based on Contourlet transform [J]. Remote Sensing for Land and Resources, 2015, 27(2):44-50.doi: 10.6046/gtzyyg.2015.02.07.

[36] Tu T M, Su S C, Shyu H C, et al. A new look at IHS-like image fusion methods [J]. Information Fusion, 2001, 2(3):177-186.

[37] Joyce Xu, Deep Learning for Object Detection: A Comprehensive Review [EB/OL]. https://towardsdatascience.com/deep-learning-for-object-detection-a-comprehensive-review-73930816d8d9, 2017-09-12/2019-05-28.

[38] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. arXiv preprint arXiv:1703.01086, 2017.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014.

[40] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, “Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics,” IEEE Trans. Geosci. Remote Sens., vol. 46, no. 5, pp. 1301–1312, May 2008.

[41] T.-M. Tu, P. S. Huang, C.-L. Hung, and C.-P. Chang, “A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery,” IEEE Geosci. Remote Sens. Lett., vol. 1, no. 4, pp. 309–312, Oct. 2004.

[42] A. R. Gillespie, A. B. Kahle, and R. E. Walker, “Color enhancement of highly correlated images—II. Channel ratio and “Chromaticity” transform techniques,” Remote Sens. Environ., vol. 22, no. 3, pp. 343–365, Aug. 1987.

[43] P. S. Chavez, Jr. and A. W. Kwarteng, “Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis” Photogramm. Eng. Remote Sens., vol. 55, no. 3, pp. 339–348, Mar. 1989.

[44] B. Aiazzi, S. Baronti, and M. Selva, “Improving component substitution pansharpening through multivariate regression of MS+Pan data,” IEEE Trans. Geosci. Remote Sens., vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[45] A. Garzelli, F. Nencini, and L. Capobianco, “Optimal MMSE pan sharpening of very high resolution multispectral images,” IEEE Trans. Geosci. Remote Sens., vol. 46, no. 1, pp. 228–236, Jan. 2008.

[46] J. Choi, K. Yu, and Y. Kim, “A new adaptive component-substitution based satellite image fusion by using partial replacement,” IEEE Trans. Geosci. Remote Sens., vol. 49, no. 1, pp. 295–309, Jan. 2011.

[47] W. Dou, Y. Chen, X. Li, and D. Sui, “A general framework for component substitution image fusion: An implementation using fast image fusion method,” Comput. Geosci., vol. 33, no. 2, pp. 219–228, Feb. 2007.

[48] R. A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, 2nd ed. Orlando, FL, USA: Academic, 1997.

[49] J. G. Liu, “Smoothing filter based intensity modulation: A spectral preserve image fusion technique for improving spatial details,” Int. J. Remote Sens., vol. 21, no. 18, pp. 3461–3472, Dec. 2000.

[50] L. Wald and T. Ranchin, “Comment: Liu ‘Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details’,”Int. J. RemoteSens., vol.23, no.3, pp. 593–597, Jan. 2002.

[51] M. M. Khan, J. Chanussot, L. Condat, and A. Montavert, “Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique,” IEEE Geosci. Remote Sens. Lett., vol. 5, no. 1, pp.98–102, Jan. 2008.

[52] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, “MTF-tailored multi-scale fusion of high-resolution MS and Pan imagery,” Photogramm. Eng.RemoteSens, vol. 72, no.5, pp.591–596, May. 2006.

[53] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, “An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas,” in Proc. 2nd GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion URBAN Areas, 2003, pp.90–94.

[54] Xia G S, Bai X, Ding J, et al. DOTA: A Large-scale Dataset for Object Detection in Aerial Images [J]. 2017.

[55] L. Alparone et al., “Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data fusion contest,” IEEE Trans. Geosci. Remote Sens., vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

[56] Dstl Satellite Imagery Feature Detection [EB/OL]. https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection, 2017-03-08/2019-05-28.

[57] 2019 IEEE GRSS Data Fusion Contest [EB/OL]. http://www.grss-ieee.org/community/technical-committees/data-fusion/data-fusion-contest/, 2019-05-28.