Multidim Syst Sign Process (2016) 27:925–944
DOI 10.1007/s11045-015-0370-3

Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping

Peicheng Zhou[1] · Gong Cheng[1] · Zhenbao Liu[2] · Shuhui Bu[2] · Xintao Hu[1]

Received: 30 June 2015 / Revised: 9 October 2015 / Accepted: 20 November 2015 /
Published online: 28 November 2015

Abstract Target detection in remote sensing images (RSIs) is a fundamental yet challenging problem faced for remote sensing images analysis. More recently, weakly supervised learning, in which training sets require only binary labels indicating whether an image contains the object or not, has attracted considerable attention owing to its obvious advantages such as alleviating the tedious and time consuming work of human annotation. Inspired by its impressive success in computer vision field, in this paper, we propose a novel and effective framework for weakly supervised target detection in RSIs based on transferred deep features and negative bootstrapping. On one hand, to effectively mine information from RSIs and improve the performance of target detection, we develop a transferred deep model to extract high-level features from RSIs, which can be achieved by pre-training a convolutional neural network model on a large-scale annotated dataset (e.g. ImageNet) and then transferring it to our task by domain-specifically fine-tuning it on RSI datasets. On the other hand, we integrate negative bootstrapping scheme into detector training process to make the detector converge more stably and faster by exploiting the most discriminative training samples. Comprehensive evaluations on three RSI datasets and comparisons with state-of-the-art weakly supervised target detection approaches demonstrate the effectiveness and superiority of the proposed method.

Keywords Target detection · Weakly supervised learning · Transferred deep features · Negative bootstrapping · Remote sensing images

1 Introduction

Target detection in remote sensing images (RSIs) is one of the fundamental problems for remote sensing images analysis. Especially, with the rapid development of remote sensing

✉ Gong Cheng
chenggong1119@gmail.com

[1] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

[2] School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China

technology, more and more high-spatial-resolution RSIs are becoming available. They contain richer visual information and make it possible to describe more surface appearance of the earth. However, how to robustly and effectively detect targets in complicated scenes is still a profound challenge faced for RSI analysis.

Recently, target detection in RSIs has been studied extensively. In the early studies, most approaches employed unsupervised methods to detect targets, which largely relied on the features used in their methods and may be effective for detecting specific targets with simple appearance and small variations. For example, Tello et al. (2005) adopted discrete wavelet transform to detect ships in synthetic aperture radar (SAR) images. Sirmacek and Unsalan (2009) utilized scale invariant feature transform (SIFT) keypoints and graph theory to detect urban areas and buildings.

In order to effectively detect targets in complicated scenes of RSIs, more approaches adopted supervised learning (SL)-based classification technique to fulfill this process. In contrast to unsupervised-based methods, SL-based methods can take advantage of the prior knowledge obtained from manually annotated samples to train more robust target detectors. Over the past few years, a number of classifiers have been used for target detection, such as Support Vector Machines (SVM) (Cheng et al. 2013b; Sun et al. 2012), k-nearest neighbour (k-NN) (Cheng et al. 2013a), Sparse Coding (Han et al. 2014; Sun et al. 2012; Zhao et al. 2013), etc. Specifically, Cheng et al. (2013b) developed an object detection framework using a discriminatively trained mixture model. Sun et al. (2012) developed a spatial sparse coding bag-of-words model to represent targets and the detection was achieved by SVM classifier. Cheng et al. (2013a) proposed a landslide detection method based on bag-of-visual-words (BoVW) in combination with probabilistic latent semantic analysis (pLSA) model and k-NN classifier. Han et al. (2014) combined visual saliency and discriminative sparse coding for efficient and simultaneous multi-class targets detection from optical RSIs. However, the above approaches all employed fully supervised learning, good performance could be achieved only when the manually labeled samples are provided. To alleviate the tedious and unreliable manual annotation, some researchers adopted semi-supervised learning (SSL) methods to perform object detection (Capobianco et al. 2009; Cheng et al. 2014; Liu et al. 2008), in which only a few labeled training samples were used to train detectors and then new samples in training set were exploited from unlabeled data. For instance, Cheng et al. (2014) proposed a Collection of Part Detectors (COPD) method to detect multi-class geospatial objects on a publicly available high-spatial-resolution RSIs dataset containing 10-class objects,[1] where each part detector was trained from a representative seed to correspond to a particular viewpoint of one specific object class. However, these methods still require a comparative number of manual labeled positive examples.

To further minimize the manual annotation while not deteriorating the target detection performance significantly, Zhang et al. (2015) developed a novel weakly supervised learning (WSL) framework to detect targets in RSIs efficiently, where the training set only indicated whether an image contains the to-be-detected targets or not, rather than annotating targets with accurate bounding boxes. Experimental results in Zhang et al. (2015) demonstrate the performance of WSL method is comparable with and even surpasses some fully supervised learning methods on some specific datasets.

The typical WSL scheme starts from generating initial training set by some techniques, and then uses them to train target detector and annotate training set iteratively, which results in an optimal detector after convergence is reached. Following the typical WSL scheme, in the last few years, most approaches focus on how to precisely select the initial positive

[1] http://pan.baidu.com/s/1hqwzXeG.

training examples and how to annotate the new positives on each refining iteration. There are few considerations for the selection of negative examples, and randomly sampling is actually a widely adopted technique in the literature. However, this may bring deterioration or fluctuation of the classifier performance during the iterative training process. Although some model drift detection methods have been proposed to evaluate the target detector on each iteration and to stop the iterative learning process when the detector starts to deteriorate (Siva and Xiang 2011; Zhang et al. 2015), they may drop into a local optimization rather than a global optimization.

Since a classifier tends to misclassify negative examples which are visually similar to positive ones, exploiting the informative negatives should be very important for enhancing the effectiveness and robustness of the classifier. Guided by this observation, in this paper, we propose to integrate negative bootstrapping scheme into weakly supervised learning to train more robust target detector. Furthermore, in order to represent targets effectively and further improve the detection accuracy, we develop a transferred deep model to extract deep features from RSIs, which can be obtained by pre-training a deep convolutional neural network (CNN) model on a large-scale dataset (e.g. ImageNet Deng et al. 2012) and then fine-tuning it on a domain-specific RSI dataset. The transferred deep model can carry more semantic meanings and hence yields more effective image representation.

To sum up, the principal contributions of this paper are twofold. First, for better capturing high-level features of remote sensing images, we develop a transferred deep model to extract domain-specific features from RSIs, which carry more semantic meanings than hand-crafted features. Second, we integrate negative bootstrapping scheme into iterative detector training process, which makes the detector converge more stably and faster by selecting high-confidence positives and negatives which tend to be misclassifed and are visually similar to positives rather than randomly sampling. The quantitative and comprehensive experiments on three RSI datasets and comparisons with state-of-the-art weakly supervised target detection approaches demonstrate the effectiveness and superiority of the proposed method.

The rest of the paper is organized as follows. Related work is briefly reviewed in Sect. 2. The transferred deep model training process is described in Sect. 3. Section 4 describes the proposed WSL-based target detection framework and implementation details. Comprehensive experiments are set up in Sect. 5. At last, we conclude the paper and future work in Sect. 6.

2 Related work

2.1 Weakly supervised learning

Weakly supervised learning has attracted much attention as an emerging machine learning technique in recent years. It can alleviate the tedious and time consuming work of human annotation compared with SL and SSL. For weakly supervised object detection, WSL only needs to indicate whether the images in the training dataset contain the targets of interest or not instead of annotating the accurate object locations. There have been extensive works on WSL-based object detection in natural scene images (Shi et al. 2013; Siva et al. 2012; Siva and Xiang 2011). For example, Siva and Xiang (2011) proposed a weakly supervised object detector learning method, in which both inter-class and intra-class information were utilized to initialize annotation and the iterative learning was stopped by a mode drift detection method. Shi et al. (2013) proposed a Bayesian joint topic model to jointly model all object classes and image backgrounds together in a single generative model. Siva et al. (2012) per-

formed initial annotation by negative mining to select exemplars with maximum inter-class variance. Furthermore, there are also a few works using WSL for remote sensing image analysis. For instance, Yang et al. (2012) proposed a weakly supervised hierarchical Markov aspect model (HMAM) for SAR-based terrain classification. Zhang et al. (2015) developed an efficient target detection method by leveraging WSL in RSIs, which can minimize the manual annotation without deteriorating performance significantly. Han et al. (2015a) proposed an improved WSL method by integrating saliency, intra-class compactness, and inter-class separability in a unified Bayesian framework. Although these existing approaches have obtained good results, they mainly focus on positive training data mining while ignore negative training data mining.

2.2 Negative training data mining

How to exploit informative negatives rather than randomly sampling is a critical factor for object detection because randomly sampling may bring deterioration and fluctuation of the target detector performance. To weaken the influence of negatives selection caused by randomly sampling, model averaging strategy was employed in some methods (Natsev et al. 2005; Tao et al. 2006) by combining multiple classifiers that were trained with randomly sampled negatives multiple times, but the effectiveness was still not satisfactory in practice. Under the phenomenon that a classifier tends to misclassify negative examples which resemble positive ones, Li et al. (2011, 2013) proposed a classifier training method going beyond random sampling by negative bootstrapping. It improved the accuracy of classifier by selecting informative negatives and raised the efficiency of classification by model compression.

2.3 High-level image representation

A critical procedure for target detection in RSIs is how to extract effective image features. Conventional methods generally employ handcrafted features, such as low-level features (e.g. the value of original pixels Han et al. 2014 and histogram of oriented gradient (HoG) Cheng et al. 2013b) or mid-level features (e.g. bag-of-words (BoW) Csurka et al. 2004, locality-constrained linear coding (LLC) Wang et al. 2010, sparsely-constructed Gaussian processes (L1-GPs) Liu et al. 2014, and Sparselets Cheng et al. 2015a, b, c), to describe image patches. However, these handcrafted features carry little semantic meanings and are not adaptive for different domains, which severely limit the descriptive power of the image representation. To address the shortcomings of handcrafted features, some high-level image representations (Shao et al. 2014b; Zhang et al. 2014) and domain-adaptive learning methods (Shao et al. 2014a; Zhu and Shao 2014) have been proposed. In addition, some methods employ the information from other domains (such as fMRI from brain image field Han et al. 2013b) to bridge semantic gap between human-centric high-level content and low-level visual features, but the acquisition of these cross domain information is expensive and limited. Nowadays, deep neural network has sprung up as a new feature learning method. It can also narrow the semantic gap between low-level visual features and high-level semantics by computing features hierarchically. As one of the most representative deep models, CNN has achieved significant results in many computer vision fields and beyond. Especially in the work of Krizhevsky et al. (2012), CNN-based method achieved impressive image classification accuracy on ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Deng et al. 2012). Following the structure of AlexNet CNN model in Krizhevsky et al. (2012), there have been a lot of modified CNN frameworks, such as Caffe (Jia et al. 2014), Decaf (Donahue et al. 2013), OverFeat (Sermanet et al. 2013), etc. These frameworks can be seen as image

feature extractors by using their pre-trained network or fine-tuning them to new domains. In this paper, we use AlexNet CNN (Krizhevsky et al. 2012) to perform our transferred deep model training and feature extraction because it has been proven to be effective for image classification (Jia et al. 2014) and object detection (Girshick et al. 2013).

3 Transferred deep model training

For target detection, the detector performance largely depends on the extracted image features. As a feature extraction model, CNN has been proven to be capable of capturing more informative patterns and semantic meanings of images, and hence has substituted hand-designed low-level or mid-level features in many fields. In our work, to effectively mine information from RSIs and improve the performance of target detection, we develop a transferred deep model to extract semantic features from RSIs, which can be achieved by pre-training a CNN model on a large-scale annotated dataset (e.g. ImageNet Deng et al. 2012) and then transferring it to our task by domain-specifically fine-tuning (Girshick et al. 2013; Oquab et al. 2014) it on RSI datasets.

However, in weakly supervised datasets, there are only image-level labels indicating whether an training image contains the to-be-detected targets or not, while the accurate object annotations for CNN model fine-tuning is unavailable. To tackle this problem, we design a strategy to obtain a virtually labeled training dataset from positive RSIs. To be specific, it is generated by using the following steps: (1) Adopt multi-scale sliding window mechanism to collect a large number of image patches in positive RSIs and then randomly sample a portion from the whole image patch set. (2) Perform k-means clustering over these sampled image patches according to a predefined cluster number, and each cluster corresponds to a bag of image patches with the same label. (3) Merge the clusters that are similar to each other and remove the clusters that have few members. Although we do not know the definite class name of each cluster, the image patches from the same cluster have similar visual patterns and the image patches from different clusters are visually different. After these steps, the domain-specific fine-tuning dataset has been constructed and could be used to transfer the pre-trained CNN model to remote sensing image domain.

In our implementation, we employ a pre-trained AlexNet CNN model (Krizhevsky et al. 2012) in Caffe library (Jia et al. 2014), and then use the training data obtained above to fine-tune it. Specifically, the parameters of convolutional layers C1…C5 and the first fully connected layer Fc6 are first trained on ImageNet (Deng et al. 2012), and then are transferred to our RSI dataset and kept fixed. The dimension of the new adapted layer Fc7 is set to 1024, which can be seen as a dimensionality reduction compared with the 4096-dimensional Fc7 layer of the pre-trained AlexNet CNN model (Krizhevsky et al. 2012) to alleviate the curse of dimensionality in detector training process. The dimension of the softmax classification layer Fc8 is set to the number of clusters. Figure 1 shows the flowchart of our transferred deep model training. After fine-tuning, we can use this transferred deep model to extract domain-specific features (i.e., the output of Fc7 layer) for RSIs.

4 Weakly supervised target detection

Figure 2 gives the flowchart of our developed weakly supervised target detection framework. It mainly consists of two stages: target detector training and target detection. In the detector

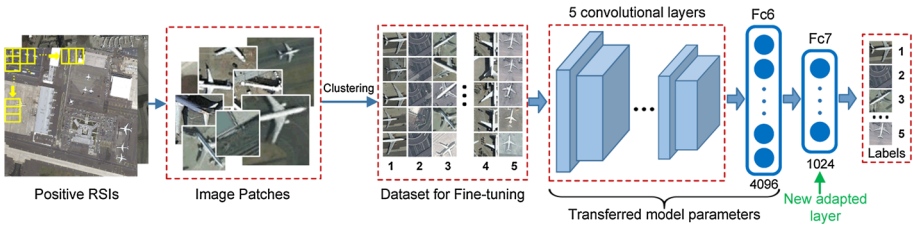

Fig. 1 The flowchart of our transferred deep model training (an example for Google Earth dataset)

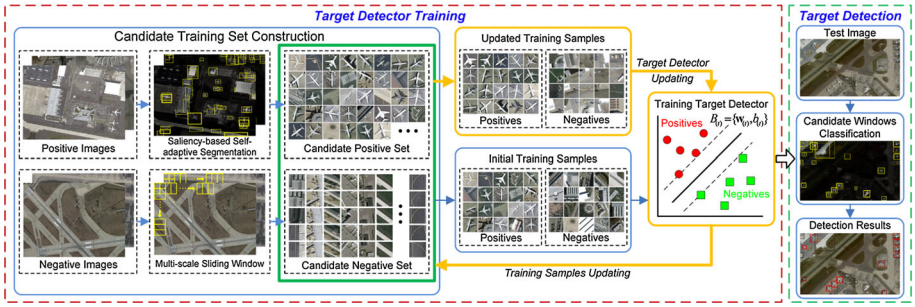

Fig. 2 The framework of the proposed weakly supervised target detection (an example for airplane detection)

training stage, given the image-level labels indicating whether an image contains the to-be-detected targets or not, we firstly initialize training samples by generating the most likely positive samples and the most relevant negative samples. To this end, we collect initial positive samples by a saliency-based self-adaptive segmentation method (Zhang et al. 2015) and refine them by negative mining (Zhang et al. 2015; Zhou et al. 2015). Afterwards we select initial negative samples which are most visually similar to initial positives. Then we use these initialized training samples to train detector iteratively. On each iteration, we exploit the most informative training samples from both positive and negative RSIs based on currently trained classifier. Repeating this until convergence is reached we can obtain an optimal detector. In the target detection stage, given a testing RSI, we first employ a saliency-based self-adaptive segmentation method (Zhang et al. 2015) to predict a small number of candidate windows for accelerating detection speed. Then, we use the target detector trained in the first stage to classify each window and obtain their corresponding responses. Finally, a post-processing scheme is used to eliminate repeated detections via non-maximum suppression (Cheng et al. 2013b, 2014; Han et al. 2015a, 2014; Zhang et al. 2015).

4.1 Candidate training set construction

How to construct a good candidate training set for training samples initialization is very important in WSL scheme. In general, the candidate training set should contain more high-confidence positive samples and more heterogeneous and non-redundant negative samples.

4.1.1 Candidate positive set

Considering there is no prior information about the position, shape, and scale of targets in positive images, in this paper, we employ a saliency-based self-adaptive segmentation method (Zhang et al. 2015) to construct candidate positive set. To be specific, for a positive image, we

first adopt the saliency model of Zhang et al. (2015) to yield an overall saliency map by linearly combining some normalized low-level and mid-level features for each pixel of the positive image. Then a self-adaptive segmentation is performed on the saliency map to obtain candidate positive regions by using multiple thresholds $thresh = \frac{\kappa}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y)$, where W and H are the width and height of the original image, $S(x, y) \in [0, 1]$ is the saliency value of the pixel at position (x, y), and $\kappa = \{1.5, 1.8, 2\}$ is a parameter controlling the segmentation threshold. Finally, the candidate positive set is formed by collecting all the image patches labeled by bounding boxes on the segmented regions (Pandey and Lazebnik 2011). Here, other salient object detection models and segmentation methods can be also used for candidate positive set construction, such as the methods in Feng et al. (2011), Han et al. (2006, 2013a, 2015b, c).

4.1.2 Candidate negative set

In WSL scheme, negative images definitely do not contain any target. Therefore, we can easily collect negative training samples from negative images. However, to accommodate the purpose of mining informative negatives, we need to construct a candidate negative set which contains diverse and non-redundant negative samples as much as possible. The construction of candidate negative set is implemented in terms of the following steps.

(1) Adopt multi-scale window mechanism (Han et al. 2014) to collect a large number of negative samples in negative RSIs. These negative samples form an unrefined negative set U^-.
(2) Although U^- contains substantially diverse negative samples, it also contains a mass of redundant samples which will affect the target detector performance during iterative training process (because with a fixed number of negative examples, if the redundancy is not excluded, the diversity of training examples will decrease). To exclude redundant negative samples and meanwhile maintain the diversity of negative samples, we employ a typical k-means clustering over these samples according to a predefined cluster number. The redundancy can also be removed by using other methods such as adaptive clustering (Ren and Jiang 2009). For a predefined cluster number K, we can obtain cluster centers $\mathbf{D} = \{d_i, i = 1, 2, \ldots, K\}$, where d_i is the i-th cluster.
(3) Combine the top ranked n samples in each cluster to form the candidate negative set. In this way, we can obtain nK samples as the candidate negative set $S^- = \{s_{ij}^-, j = 1, 2, \ldots, n\} \subset U^-$, where s_{ij}^- is the feature representation of the j-th samples in the i-th cluster extracted by our transferred deep model. Figure 3 shows 20 randomly selected clusters from Google Earth dataset, where each column corresponds to a cluster and each cluster has five top ranked negative samples. As can be seen from Fig. 3: (1) Almost all samples within each cluster are visually consistent, so they are redundant and the redundancy should be excluded. (2) Samples between different clusters are visually different, so we should preserve at least one sample in each cluster. After these operations, we can obtain a set of diverse and non-redundant negatives.

4.2 Training samples initialization

4.2.1 Positive training samples

We adopt negative mining strategy (Zhang et al. 2015; Zhou et al. 2015) to obtain initial positive training samples from the candidate positive set S^+, under the observation that

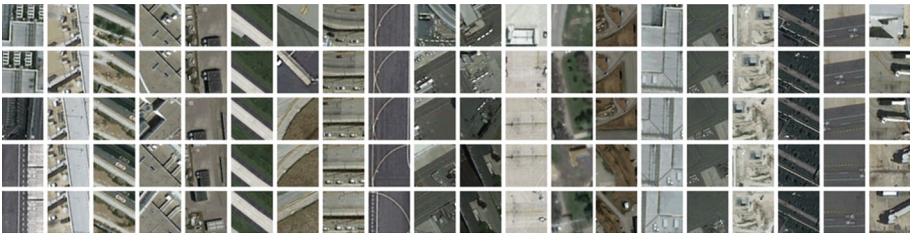

Fig. 3 20 randomly selected clusters from negative candidate set, where each column corresponds to a cluster with its top-five ranked negative samples (an example of Google Earth dataset)

targets are regularly different from negative samples in visual appearance. To be specific, let $S^+_{(1)} = \{s^+_p, p = 1, 2, \ldots, n_p\} \subset S^+$ denote initial positive samples, where s^+_p is the feature representation of p-th positive sample, n_p is the number of initial positive samples. The negative mining algorithm was implemented as follows.

$$S^+_{(1)} = \left\{ s^+_p | dist\left(s^+_p\right) > \tau,\ s^+_p \in S^+ \right\} \tag{1}$$

$$dist\left(s^+_p\right) = \min_{i\in\{1,K\}, j\in\{1,n\}} \left\| s^+_p - s^-_{ij} \right\|_1 \tag{2}$$

where $\|\cdot\|_1$ is the L1-norm and τ is a threshold used for excluding some noisy positive samples that are visually similar to negatives.

4.2.2 Negative training samples

The informative negative samples are considered to be the samples which tend to be misclassified. However, on the first iteration, the pre-trained target detector is not available for predicting which samples are most likely misclassified. To alleviate the effect of randomly sampled negatives caused by traditional way and make the whole training process to be more robust, we initialize the negative samples to be those that are most similar to the initial positive samples by measuring their distances in CNN feature space. The distance between the negative samples in S^- and the initial positive samples $S^+_{(1)}$ can be calculated by

$$Dist\left(S^-, S^+_{(1)}\right) = \left\{ dist\left(s^-_q, s^+_p\right),\ s^-_q \in S^-,\ s^+_p \in S^+_{(1)} \right\} \tag{3}$$

$$dist\left(s^-_q, s^+_p\right) = \min_{p\in\{1,n_p\}} \left\| s^-_q - s^+_p \right\|_1 \tag{4}$$

We then rank all negative samples in S^- by $Dist(S^-, S^+_{(1)})$ in ascending order and select top ranked samples as our initial negative samples. To balance the number of training samples, we select the negative samples with the same number of $S^+_{(1)}$. Let $S^-_{(1)}$ be the initial negative samples, it can be generated by

$$S^-_{(1)} \leftarrow select\ top\ samples\left(S^-, Dist\left(S^-, S^+_{(1)}\right), |S^+_{(1)}|\right) \tag{5}$$

where $|\cdot|$ denotes the cardinality of a given dataset.

4.3 Iterative target detector training

Algorithm 1 gives the procedure of target detector training.

Algorithm 1 Training Procedure of Target Detector

Input: Initial training samples $S_{(1)}^{+}$ and $S_{(1)}^{-}$, candidate training set S^{+} and S^{-}, and the number of learning iteration T.

Output: Target detector $B_{(T)} = \{\mathbf{w}_{(T)}, b_{(T)}\}$

1. Train an initial target detector $B_{(1)} = \{\mathbf{w}_{(1)}, b_{(1)}\}$

2. **For** $t = 2$ to T **do**

(1) **Update training samples**

(a) Calculate the scores of candidate samples by $B_{(t-1)} = \{\mathbf{w}_{(t-1)}, b_{(t-1)}\}$

$$Score_{(t)}(s_p^{+}) = \mathbf{w}_{(t-1)}^{T} s_p^{+} + b_{(t-1)},\ s_p^{+} \in S^{+}$$

$$Score_{(t)}(s_q^{-}) = \frac{\mathbf{w}_{(t-1)}^{T} s_q^{-} + b_{(t-1)}}{1 + \log[1 + Dist(s_q^{-}, S^{+})]},\ s_q^{-} \in S^{-}$$

(b) Exploit new training samples:

$$S_{(t)}^{+} = \{s_p^{+} \mid Score_{(t)}(s_p^{+}) > \sigma, s_p^{+} \in S^{+}\}$$

$$S_{(t)}^{-} \leftarrow select\ top\ samples(S^{-}, Score_{(t)}(S^{-}), |S_{(t)}^{+}|)$$

(2) **Optimize target detector**

Using $S_{(t)}^{+}$ and $S_{(t)}^{-}$ to train a new target detector $B_{(t)} = \{\mathbf{w}_{(t)}, b_{(t)}\}$ by (6)

end

4.3.1 Target detector training

After obtaining the initial training samples $S_{(1)}^{+}$ and $S_{(1)}^{-}$, we can train an initial target detector $B_{(1)}$, and then update training samples and optimize target detector iteratively until the convergence is reached. Let T denote the total number of iterations, $t = 2, \ldots, T$ be the iteration index, $S_{(t)}^{+}$ and $S_{(t)}^{-}$ be the updated training samples on the t-th iteration, $B_{(t)}$ be the target detector trained on t-th iteration. In our work, we adopt a linear SVM to train target detector, which is formulated as

$$\min_{\mathbf{w}_{(t)}, b_{(t)}} \frac{1}{2}\mathbf{w}_{(t)}^{T}\mathbf{w}_{(t)} \quad s.t.\ \ y_m\left(\mathbf{w}_{(t)}^{T} s_m + b_{(t)}\right) - 1 \geq 0 \tag{6}$$

where $s_m \in S_{(t)}^{+} \cup S_{(t)}^{-}$ is the m-th training samples, $y_m \in \{1, -1\}$ is the label of s_m. The target detector can be represented as $B_{(t)} = \{\mathbf{w}_{(t)}, b_{(t)}\}$. For predicting the score of a sample, we regard it as a classification problem, which is formulated as

$$Score_{(t+1)}(s_m) = \mathbf{w}_{(t)}^{T} s_m + b_{(t)} \tag{7}$$

4.3.2 Training samples updating

On the t-th iteration, the informative positive training samples $S^+_{(t)}$ and negative training samples $S^-_{(t)}$ can be updated by $B_{(t-1)}$, which is trained on the $(t\text{-}1)$-th iteration. To this end, we use the target detector $B_{(t-1)}$ to compute the scores of positive and negative candidate samples respectively. For positive candidate samples, the scores can be calculated by (7) and represented by

$$Score_{(t)}(S^+) \leftarrow \left\{ Score_{(t)}\left(s_p^+\right) = \mathbf{w}_{(t-1)}^T s_p^+ + b_{(t-1)},\ s_p^+ \in S^+ \right\} \tag{8}$$

Considering the informative negative samples tend to be easily misclassified and are visually similar to positive samples, so their scores are proportional to $\mathbf{w}_{(t-1)}^T s_q^- + b_{(t-1)}$ and inversely proportional to $Dist(s_q^-, S^+)$, where $\mathbf{w}_{(t-1)}^T s_q^- + b_{(t-1)}$ is the response of s_q^- to the target detector $B_{(t-1)}$ and $Dist(s_q^-, S^+)$ is the distance similarity measurement between s_q^- and positive samples based on their feature representation. Besides, since the values of $\mathbf{w}_{(t-1)}^T s_q^- + b_{(t-1)}$ are generally far smaller than the values of $Dist(s_q^-, S^+)$, we use log function to reduce the scale of $Dist(s_q^-, S^+)$. Thus, the scores of negative candidate samples can be calculated by

$$Score_{(t)}(S^-) \leftarrow \left\{ Score_{(t)}\left(s_q^-\right) = \frac{\mathbf{w}_{(t-1)}^T s_q^- + b_{(t-1)}}{1 + \log\left[1 + Dist\left(s_q^-, S^+\right)\right]},\ s_q^- \in S^- \right\} \tag{9}$$

Then, the updated positive samples can be obtained by selecting the samples with their scores above a given threshold σ in S^+.

$$S^+_{(t)} = \left\{ s_p^+ | Score_{(t)}\left(s_p^+\right) > \sigma,\ s_p^+ \in S^+ \right\} \tag{10}$$

And the updated negative samples $S^-_{(t)}$ can be obtained by selecting the top ranked samples with the same number of $S^+_{(t)}$.

$$S^-_{(t)} \leftarrow select\ top\ samples\left(S^-, Score_{(t)}(S^-), |S^+_{(t)}|\right) \tag{11}$$

4.4 Target detection

To detect targets in RSIs efficiently and accurately, we adopt a candidate-patch-based target detection scheme (Zhang et al. 2015), which results in about 10 times speed gain compared with conventional sliding window-based methods. For a given RSI, the candidate windows can be obtained by saliency-based self-adaptive segmentation method following previous work (Han et al. 2014; Zhang et al. 2015). Then, we use the target detector $B_{(T)} = \left\{\mathbf{w}_{(T)}, b_{(T)}\right\}$ to obtain their responses to determine whether these candidate windows contain targets or not. Finally, a non-maximum suppression scheme (Cheng et al. 2013b, 2014; Han et al. 2015a, 2014; Zhang et al. 2015) is adopted to eliminate repeated detections.

Table 1 Detailed information about three datasets (Zhang et al. 2015)

Datasets	Dimension (pixels)	Spatial resolution	Target area (pixels)
Google Earth	About 1000 × 800	About 0.5 m	700–25,488
ISPRS	About 900 × 700	8–15 cm	1150–11,976
Landsat	400 × 800	30 m	1760–15,570

5 Experiments

5.1 Experimental setup

5.1.1 Dataset description

We quantitatively evaluate the proposed method on three different RSI datasets from Zhang et al. (2015), which come from Google Earth, ISPRS (provided by the German Association of Photogrammetry and Remote Sensing (Cramer 2010)) and Landsat-7 ETM+. These three datasets are used to detect airplanes, vehicles, and airports, respectively. The detailed information about these three datasets is shown in Table 1. To objectively and fairly evaluate this method, we use the same data selection strategy as in Zhang et al. (2015). Specifically, for Google Earth dataset, we separated it into two parts, where 70 RSIs for training and 50 RSIs for testing. For ISPRS dataset, we randomly selected 60 RSIs for training and the remaining 40 RSIs for testing. For Landsat dataset, the training set includes 123 RSIs and the testing set contains 57 RSIs. In addition, there are 50 negative RSIs in Google Earth dataset, 24 negative RSIs in ISPRS dataset, and 37 negative RSIs in Landsat dataset, respectively. All these 111 negative RSIs do not contain any target.

5.1.2 Feature extraction

In our previous work (Zhou et al. 2015), we employed a universal AlexNet CNN model (Krizhevsky et al. 2012) implemented in open source Caffe library (Jia et al. 2014) to extract image features directly. Although it has achieved good performance, it is still insufficient for remote sensing image analysis. In this paper, we develop a transferred deep model built on AlexNet CNN model to extract RSI features. Specifically, we firstly resize each image patch from their original pixel size to a uniform 227 × 227 pixel size because the architecture of the transferred deep model requires inputs of a fixed 227 × 227 pixel size. Then, we feed the raw data into a forward propagating neural network with five convolutional layers and two fully connected layers to output a 1024-dimensional feature vector for each image patch. We implement feature extraction using Matlab platform and run the Caffe toolkit in CPU mode on a PC with Windows 7 and Intel Core2 2.93 GHz CPU and 4 GB memory. The time consuming for each image patch is about 1.1 s.

5.1.3 Implementation details

For transferred deep model training, we randomly sampled 50,000 image patches from the positive RSIs on each dataset and set the initial predefined cluster number to 1000 for all three datasets. After cluster centers merging and removing, the numbers of virtual training data classes (i.e. the dimension of Fc8 layer) are 931, 967, and 909 for Google Earth dataset,

Table 2 Detailed parameter settings for candidate negative set construction

Datasets	The scales for collecting negative samples	Clustering number K	Top-n	τ	σ	T
Google Earth	60×60, 80×80, 100×100	5000	1	0.85	0.95	100
ISPRS	40×40, 50×50, 60×60	5000	1	0.90	0.95	100
Landsat	80×80, 100×100, 120×120	5000	1	0.80	0.85	100

ISPRS dataset, and Landsat-7 dataset, respectively. The learning rate and the number of iteration are set to 0.005 and 20,000 empirically.

To construct candidate negative set, we collected a large number of negative samples with multiple scales and refined them using k-means clustering. Then we selected the top-n samples in each cluster to form our candidate negative set. The parameters of threshold τ in training samples initializing, σ in training samples updating, and the number of iterations T were set empirically according to our experimental results. The binary classifier was trained by using LibSVM toolbox (Chang and Lin 2011) with a linear kernel. The detailed parameter settings for candidate negative set construction for three different datasets is listed in Table 2, where the scales are set empirically.

5.1.4 Evaluation criterion

We adopt Average Precision (AP) to quantitatively evaluate the performance of the proposed method, which is a standard criterion for evaluating target detection (Cheng et al. 2013b; Han et al. 2014; Zhang et al. 2015) and is measured by the area under Precision-Recall curve (PRC). The higher the AP value is, the better the performance and vice versa. By following the works of (Cheng et al. 2013b, 2014; Han et al. 2015a, 2014; Zhang et al. 2015), a detection result is considered as a true positive if the overlap area between a detection window and the ground truth is more than 50 %.

5.2 Experimental results

5.2.1 The influence of informative negatives

One of the goals of this paper is to enhance the robustness and effectiveness of target detector by exploiting informative negatives in WSL scheme. Naturally, we compared our method with conventional WSL methods in which negative training samples were obtained by randomly sampling. For fair comparison, the negative samples used in these methods were all selected from the same candidate negative set constructed in Sect. 3. As shown in Fig. 4, after several iterations, the average precision of our WSL method with informative negatives tends to stable, while the performance of conventional WSL method with randomly sampled negatives is fluctuated through the whole iterative process. Furthermore, we also compared the proposed method with supervised learning strategy in which target detector was trained using manually labeled positive samples and randomly sampled negatives. As can be seen from Fig. 4, our proposed method is more robust than supervised learning method, and even obtains better performance on some datasets (such as Google Earth dataset and ISPRS dataset). The results show that exploiting informative negatives to train target detector is very important for improving its robustness and effectiveness in WSL scheme.

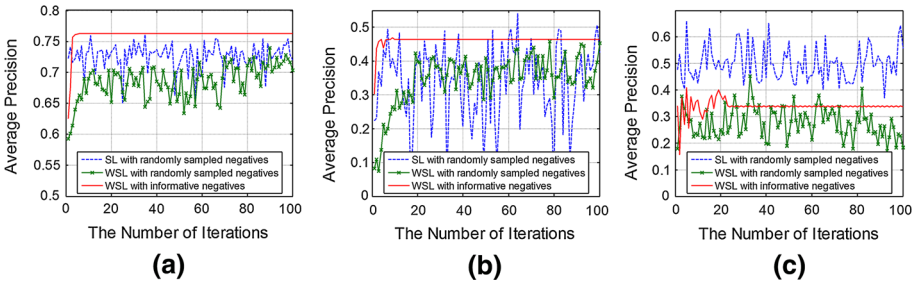

Fig. 4 The performance comparison of two WSL methods and one SL method on three datasets: **a** Google Earth dataset for airplane detection, **b** ISPRS dataset for vehicle detection and **c** Landsat dataset for airport detection

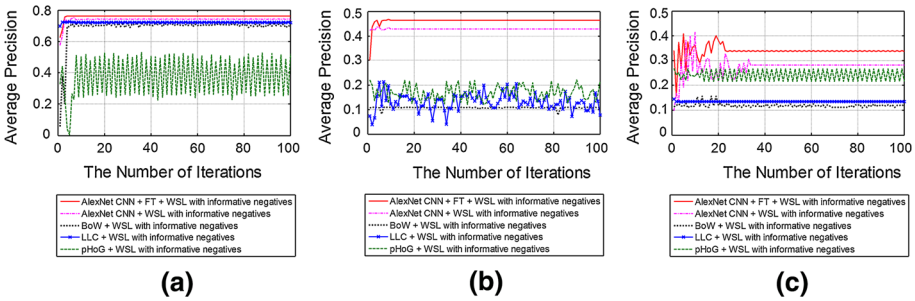

Fig. 5 The performance comparison of target detectors trained with different features and iteration numbers: **a** Google Earth dataset for airplane detection, **b** ISPRS dataset for vehicle detection and **c** Landsat dataset for airport detection

5.2.2 Transferred deep features versus traditional features

To validate the effectiveness of our transferred deep model for feature extraction, we applied the proposed framework on three RSI datasets with five types of features. These five types of features are transferred deep feature denoted by "AlexNet CNN+FT", AlexNet CNN (Krizhevsky et al. 2012), pHOG (Bosch et al. 2007), BoW (Csurka et al. 2004), and LLC (Wang et al. 2010). The parameter settings for each type of feature are all the same as the work of Jia et al. (2014), Bosch et al. (2007), Csurka et al. (2004), and Wang et al. (2010). Figure 5 shows the performance comparison of target detectors trained with different features and iteration numbers. As can be seen from Fig. 5: (1) The feature extracted by transferred deep model is much stronger than all other hand-designed features extracted by pHOG (Bosch et al. 2007), BoW (Csurka et al. 2004), and LLC (Wang et al. 2010). (2) Using the transferred deep feature we obtained better performance than that using the feature extracted with AlexNet CNN (Krizhevsky et al. 2012) directly. The comparison results demonstrate the effectiveness of the proposed transferred deep model.

5.2.3 Evaluation of our WSL method

To quantitatively evaluate the proposed framework, we compared it with our previously published method (Zhou et al. 2015) and three state-of-the-art WSL methods (Zhang et al. 2015). In our previous work (Zhou et al. 2015), we just used the AlexNet CNN model in (Krizhevsky

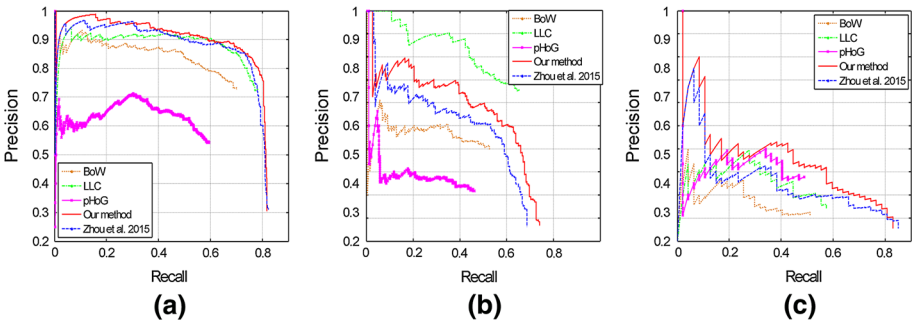

Fig. 6 Precision-recall curves of the proposed framework and four state-of-the-art approaches: **a** Google Earth dataset for airplane detection, **b** ISPRS dataset for vehicle detection and **c** Landsat dataset for airport detection

Table 3 Performance comparisons of five different methods in terms of AP values

Data sets/target classes	Our method	Zhang et al. (2015)			Zhou et al. (2015)
		BoW	LLC	pHoG	
Google Earth/airplane	**0.7626**	0.6183	0.6928	0.4038	0.7558
ISPRS/vehicle	0.4647	0.2829	**0.5119**	0.1770	0.3933
Landsat/airport	**0.3365**	0.1184	0.1845	0.2099	0.2293

Bold entries denote the best APs for each target class

et al. 2012) to extract features directly and did not consider the similarities between negative samples and positive samples when updating negative samples. The other three comparison methods were proposed by Zhang et al. (2015), which employed conventional WSL to detect targets with three different hand-designed features including BoW (Csurka et al. 2004), LLC (Wang et al. 2010), and pHoG (Bosch et al. 2007). The parameter settings for all three methods are the same as the work of Zhang et al. (2015). Briefly, BoW represents each image patch as a histogram of visual words from a codebook. LLC uses locality constraint to select five similar bases from the codebook and learns a linear combination weight of these bases to reconstruct each descriptor. The pHoG feature captures the shape property of each image patch by calculating a histogram of orientation gradients which are discretized into 16 bins with orientations in the range [0, 180]. As the implementation of Zhang et al. (2015), the codebook used in BOW and LLC was generated by extracting 128-dimensional SIFT descriptors (Lowe 2004) in training images and then clustering them into 1024 visual words via k-means algorithm. For better addressing the object variations in rotation, the global level of the pyramid representation in pHOG and LLC was also used. Figure 6 and Table 3 show the quantitative comparison results of five different methods, measured by PRC and AP values for each target class, respectively. As shown in Fig. 6 and Table 3, the performance of the proposed WSL method surpasses our previous work and other two state-of-the-art methods (BoW and pHoG) significantly, which demonstrates the effectiveness and superiority of the proposed framework. Although the average precision of our method is lower than the LLC feature for vehicle detection, considering the size of vehicle target is far smaller than the requirement of the pixel size for the inputs of the transferred deep model, there is information loss when up-sampling the vehicle targets, so the result is reasonable.

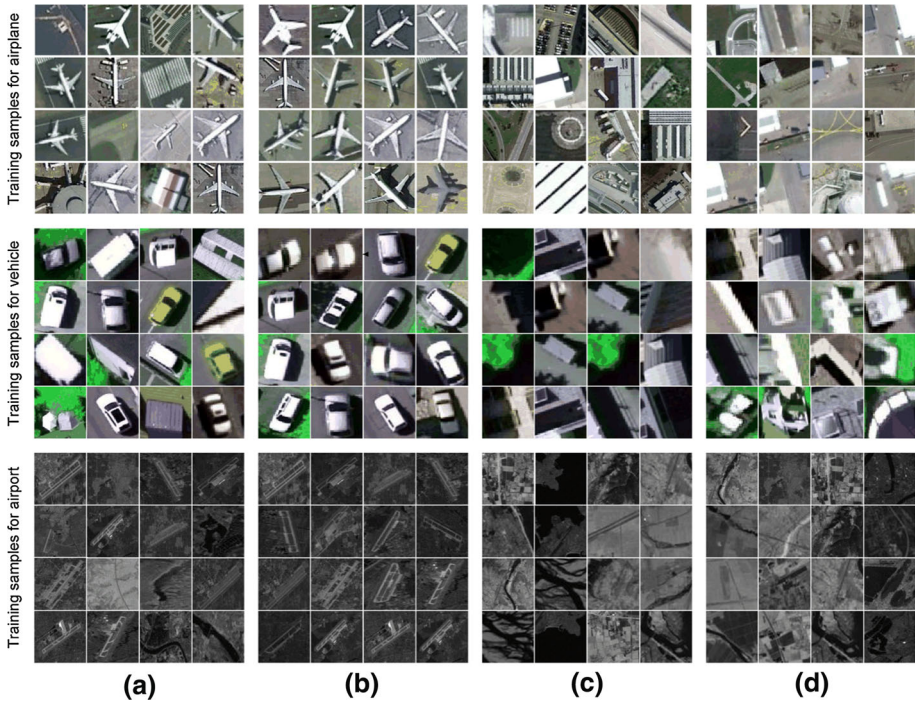

Fig. 7 The initial positive samples and negative samples and their corresponding updated samples: **a** Initial positive samples obtained by saliency-based self-adaptive segmentation, **b** updated positives obtained by the proposed method after 100 iterations, **c** negative samples collected by randomly sampling and **d** updated negatives obtained by the proposed method after 100 iterations

5.2.4 Qualitative analysis of the proposed method

To qualitatively evaluate the influence of the training samples on target detector training, we visualize some initial training samples and their corresponding updated training samples after 100 iterations used for different target detectors training in Fig. 7. As can be seen from Fig. 7a, b, after 100 iterations, the noisy samples in initial positive samples set are removed. The updated negatives in Fig. 7d obtained by the proposed method have similar visual representation to positives. They are deemed to be more informative and hence can be used to train more refined target detector. On the contrary, the negatives in Fig. 7c collected by randomly sampling do not have this characteristic. In addition, Fig. 8 gives some detection results by using the proposed approach where the true positives, false negatives, and false positives are highlighted by red, white and yellow rectangles, respectively. As can be seen, our method can effectively locate most of the targets with diverse orientations and sizes from different RSIs.

6 Conclusion

In this paper, we developed a novel framework for weakly supervised target detection in RSIs based on transferred deep features and negative bootstrapping. On one hand, we employed

Fig. 8 Some target detection results by using the proposed method on three RSIs datasets: **a** Google Earth dataset for airplane detection, **b** ISPRS dataset for vehicle detection and **c** Landsat dataset for airport detection

a transferred deep model for better extracting domain-specific features of remote sensing images. On the other hand, we integrated negative bootstrapping scheme into iterative detector training process to make the detector converge more stably and faster by selecting the most discriminative training samples. Comprehensive evaluations on three datasets and comparisons with several state-of-the-art methods demonstrate the effectiveness and superiority of the proposed method. In the future work, we will (1) integrate some discriminative information between positives and negatives to train more effective target detector and (2) apply this method to other applications, such as binary decision diagram (Li et al. 2014).

Acknowledgments This work was partially supported by the National Science Foundation of China under Grants 61401357 and U1261111HZ, the China Postdoctoral Science Foundation under Grants 2014M552491 and 2015T81050, and the Aerospace Science Foundation of China under Grant 20140153003.

References

Bosch, A., Zisserman, A., & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval* (pp. 401–408).

Capobianco, L., Garzelli, A., & Camps-Valls, G. (2009). Target detection with semisupervised kernel orthogonal subspace projection. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(11), 3822–3833.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27.

Cheng, G., Guo, L., Zhao, T., Han, J., Li, H., & Fang, J. (2013a). Automatic landslide detection from remote-sensing imagery using a scene classification method based on boVW and pLSA. *International Journal of Remote Sensing*, *34*(1), 45–59.

Cheng, G., Han, J., Guo, L., & Liu, T. (2015a). Learning coarse-to-fine sparselets for efficient object detection and scene classification. In *Proceedings of the 28th IEEE conference on computer vision and pattern recognition* (pp. 1173–1181).

Cheng, G., Han, J., Guo, L., Liu, Z., Bu, S., & Ren, J. (2015b). Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(8), 4238–4249.

Cheng, G., Han, J., Guo, L., Qian, X., Zhou, P., Yao, X., et al. (2013b). Object detection in remote sensing imagery using a discriminatively trained mixture model. *ISPRS Journal of Photogrammetry and Remote Sensing*, *85*, 32–43.

Cheng, G., Han, J., Zhou, P., & Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, *98*, 119–132.

Cheng, G., Zhou, P., Han, J., Guo, L., & Han, J. (2015c). Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images. *IET Computer Vision*, *9*(5), 639–647.

Cramer, M. (2010). The DGPF-test on digital airborne camera evaluation—Overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*, *2*, 73–82.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (pp. 1–2).

Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., & Fei-Fei, L. (2012). ImageNet large scale visual recognition competition 2012 (ILSVRC2012).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., & Tzeng, E., et al. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531.

Feng, Y., Ren, J., & Jiang, J. (2011). Object-based 2D-to-3D video conversion for effective stereoscopic content generation in 3D-TV applications. *IEEE Transactions on Broadcasting*, *57*(2 PART 2), 500–509.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524.

Han, J., He, S., Qian, X., Wang, D., Guo, L., & Liu, T. (2013a). An object-oriented visual saliency detection framework based on sparse coding representations. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(12), 2009–2021.

Han, J., Ji, X., Hu, X., Zhu, D., Li, K., Jiang, X., et al. (2013b). Representing and retrieving video shots in human-centric brain imaging space. *IEEE Transactions on Image Processing*, *22*(7), 2723–2736.

Han, J., Ngan, K. N., Li, M., & Zhang, H.-J. (2006). Unsupervised extraction of visual attention objects in color images. *IEEE Transactions on Circuits and Systems for Video Technology*, *16*(1), 141–145.

Han, J., Zhang, D., Cheng, G., Guo, L., & Ren, J. (2015a). Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(6), 3325–3337.

Han, J., Zhang, D., Hu, X., Guo, L., Ren, J., & Wu, F. (2015b). Background prior based salient object detection via deep reconstruction residual. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(8), 1309–1321.

Han, J., Zhang, D., Wen, S., Guo, L., Liu, T., & Li, X. (2015c). Two-stage learning to predict human eye fixations via SDAEs. *IEEE Transactions on Cybernetics*, online published.

Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., et al. (2014). Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *ISPRS Journal of Photogrammetry and Remote Sensing*, *89*, 37–48.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., & Girshick, R., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM international conference on multimedia* (pp. 675–678).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 1097–1105). South Lake Tahoe, NV: NIPS foundation.

Li, S., Si, S., Dui, H., Cai, Z., & Sun, S. (2014). A novel decision diagrams extension method. *Reliability Engineering & System Safety*, *126*, 107–115.

Li, X., Snoek, C. G., Worring, M., Koelma, D., & Smeulders, A. W. (2013). Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia*, *15*(4), 933–945.

Li, X., Snoek, C. G., Worring, M., & Smeulders, A. W. (2011). Social negative bootstrapping for visual categorization. In *Proceedings of the 1st ACM international conference on multimedia retrieval*.

Liu, L., Shao, L., Zheng, F., & Li, X. (2014). Realistic action recognition via sparsely-constructed Gaussian processes. *Pattern Recognition*, *47*(12), 3819–3827.

Liu, Q., Liao, X., & Carin, L. (2008). Detection of unexploded ordnance via efficient semisupervised and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(9), 2558–2567.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Natsev, A. P., Naphade, M. R., & TešiĆ, J. (2005). Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proceedings of the 13th annual ACM international conference on multimedia* (pp. 598–607).

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *27th IEEE Conference on computer vision and pattern recognition* (pp. 1717–1724).

Pandey, M., & Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models. In *Proceedings of the 2011 IEEE international conference on computer vision* (pp. 1307–1314).

Ren, J., & Jiang, J. (2009). Hierarchical modeling and adaptive clustering for real-time summarization of rush videos. *IEEE Transactions on Multimedia*, *11*(5), 906–917.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.
Shao, L., Liu, L., & Li, X. (2014a). Feature learning for image classification via multiobjective genetic programming. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(7), 1359–1371.
Shao, L., Wu, D., & Li, X. (2014b). Learning deep and wide: A spectral method for learning deep networks. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(12), 2303–2308.
Shi, Z., Hospedales, T. M., & Xiang, T. (2013). Bayesian joint topic modelling for weakly supervised object localisation. In *Proceedings of the 2013 IEEE international conference on computer vision* (pp. 2984–2991).
Sirmacek, B., & Unsalan, C. (2009). Urban-area and building detection using SIFT keypoints and graph theory. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(4), 1156–1167.
Siva, P., Russell, C., & Xiang, T. (2012). In defence of negative mining for annotating weakly labelled data. In *Proceedings of the 12th European conference on computer vision* (pp. 594–608).
Siva, P., & Xiang, T. (2011). Weakly supervised object detector learning with model drift detection. In *Proceedings of the 2011 IEEE international conference on computer vision* (pp. 343–350).
Sun, H., Sun, X., Wang, H., Li, Y., & Li, X. (2012). Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geoscience and Remote Sensing Letters*, *9*(1), 109–113.
Tao, D., Tang, X., Li, X., & Wu, X. (2006). Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(7), 1088–1099.
Tello, M., López-Martínez, C., & Mallorqui, J. J. (2005). A novel algorithm for ship detection in SAR imagery based on the wavelet transform. *IEEE Geoscience and Remote Sensing Letters*, *2*(2), 201–205.
Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). Locality-constrained linear coding for image classification. In *IEEE Conference on computer vision and pattern recognition* (pp. 3360–3367).
Yang, W., Dai, D., Triggs, B., & Xia, G.-S. (2012). SAR-based terrain classification using weakly supervised hierarchical Markov aspect models. *IEEE Transactions on Image Processing*, *21*(9), 4232–4243.
Zhang, D., Han, J., Cheng, G., Liu, Z., Bu, S., & Guo, L. (2015). Weakly Supervised Learning for Target Detection in Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, *12*(4), 701–705.
Zhang, L., Zhen, X., & Shao, L. (2014). Learning object-to-class kernels for scene classification. *IEEE Transactions on Image Processing*, *23*(8), 3241–3253.
Zhao, C., Li, X., Ren, J., & Marshall, S. (2013). Improved sparse representation using adaptive spatial support for effective target detection in hyperspectral imagery. *International Journal of Remote Sensing*, *34*(24), 8669–8684.
Zhou, P., Zhang, D., Cheng, G., & Han, J. (2015). Negative bootstrapping for weakly supervised target detection in remote sensing images. In*Proceedings of the 2015 IEEE international conference on multimedia big data* (pp. 318–323).
Zhu, F., & Shao, L. (2014). Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, *109*(1–2), 42–59.

Peicheng Zhou received the B.S. degree from Xi'an University of Technology, Xi'an, China, in 2011, and the M.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2014. He is currently a Ph.D. student in Northwestern Polytechnical University, Xi'an, China. His research interests are computer vision and pattern recognition.

Gong Cheng received the B.S. degree from Xidian University, Xi'an, China, in 2007, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2013, respectively. He is currently a postdoctoral fellow at Northwestern Polytechnical University, Xi'an, China. His main research interests are computer vision and remote sensing image analysis.

Zhenbao Liu received the Ph.D. degree in computer science from College of Systems and Information Engineering, University of Tsukuba, Japan in 2009. He was a visiting scholar in the GrUVi Lab of Simon Fraser University in 2012. He is currently an associate professor with Northwestern Polytechnical University, Xi'an, China. His research interests include 3D shape and scene analysis, computer vision, and remote sensing.

Shuhui Bu received the M.S. and Ph.D. degrees in College of Systems and Information Engineering from University of Tsukuba, Japan in 2006 and 2009. He was an assistant professor (2009-2011) at Kyoto University, Japan. He is currently an associate professor at Northwestern Polytechnical University, Xi'an, China. His research interests are concentrated on computer vision and robotics, including 3D shape analysis, image processing, pattern recognition, 3D reconstruction, and related fields.

Xintao Hu received his M.S. and Ph.D degrees from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively. He is currently an associate professor with Northwestern Polytechnical University. His research interests include computational brain imaging and computer vision.