A NOVEL SELF-TAUGHT LEARNING FRAMEWORK USING SPATIAL PYRAMID MATCHING FOR SCENE CLASSIFICATION

Yi Yang[1], Daoye Zhu[1], Fuhu Ren[1], Chengqi Cheng[1,2,*]

[1] Center for Data Science, Peking University, Beijing 100871, P.R.China - (pkuyangyi, zhudaoye, renfh)@pku.edu.cn
[2] College of Engineering, Peking University, Beijing 100871, P.R.China - ccq@pku.edu.cn

KEY WORDS: Remote Sensing, Scene Classification, Self-taught Learning, Spatial Pyramid Matching, High Resolution Imagery

ABSTRACT:

Remote sensing earth observation images have a wide range of applications in areas like urban planning, agriculture, environment monitoring, etc. While the industrial world benefits from availability of high resolution earth observation images since recent years, interpreting such images has become more challenging than ever. Among many machine learning based methods that have worked out successfully in remote sensing scene classification, spatial pyramid matching using sparse coding (ScSPM) is a classical model that has achieved promising classification accuracy on many benchmark data sets. ScSPM is a three-stage algorithm, composed of dictionary learning, sparse representation and classification. It is generally believed that in the dictionary learning stage, although unsupervised, one should use the same data set as classification stage to get good results. However, recent studies in transfer learning suggest that it might be a better strategy to train the dictionary on a larger data set different from the one to classify.
In our work, we propose an algorithm that combines ScSPM with self-taught learning, a transfer learning framework that trains a dictionary on an unlabeled data set and uses it for multiple classification tasks. In the experiments, we learn the dictionary on Caltech-101 data set, and classify two remote sensing scene image data sets: UC Merced LandUse data set and Changping data set. Experimental results show that the classification accuracy of proposed method is compatible to that of ScSPM. Our work thus provides a new way to reduce resource cost in learning a remote sensing scene image classifier.

1. INTRODUCTION

Remote sensing plays an important role in earth observation, and in this area remote sensing image scene classification is one fundamental problem (Cheng et al., 2017). With the increasing development of remote sensing imaging techniques, huge amounts of high spatial resolution images have been acquired. Detailed contents in these images, however, make automatic classification a challenging task.

In the past decade machine learning based methods have made notable success in computer vision, several of which, for example support vector machine (SVM) (Mountrakis et al., 2011) and stacked auto-encoder (Yao et al., 2016), have been applied to high resolution remote sensing image processing. Recent studies show that the bag-of-visual-words (BOVW) model is an effective and robust feature encoding approach (Yang, Newsam, 2010) (Zhu et al., 2016), generated by which the higher level image representations can improve performance of machine learning classifiers like SVM. Spatial pyramid matching using sparse coding (ScSPM) (Yang et al., 2009) is a BOVW based model that has achieved state-of-the-art performance on several open remote sensing image data sets (Yang et al., 2015) (Wu et al., 2016). Basically, ScSPM uses dictionary learning with sparse coding to train a dictionary that captures salient features, then low-level features are encoded by the dictionary and represented in a spatial pyramid way to form higher level features, which are used as input to the classifier. One bottleneck of ScSPM is the dictionary learning process. The learning objective function is generally difficult to optimize, and what's more, if ScSPM is applied to classify some data set B, then the dictionary should be trained on B. This way of dictionary learning is sometimes called "task- specific". Hence if we want to use ScSPM on multiple tasks, say to classify data set B, C, and D, then it can be very time and computation consuming to train three dictionaries on their corresponding data sets.

In the computer vision community, dictionary learning (DL) is also a topic that attracts lots of attention, on which many of the works focus on non-task-specific dictionary learning (Maurer et al., 2013) (Zhu, Shao, 2014). In self-taught learning (Raina et al., 2007), it is for the first time proposed that the dictionary can be efficiently trained in an unsupervised way. One important conclusion of self-taught learning is that a dictionary learned on a large, unlabeled data set A can be used for feature encoding of another labeled data set B, and the classifier using these encodings on B can get promising results.

Inspired by this, we propose a self-taught learning framework using spatial pyramid matching (S-ScSPM) on remote sensing scene classification from high resolution imagery. We show in our experiments that using S-ScSPM, to classify labeled data set B and C, a dictionary trained on data only from unlabeled data set A is sufficient. While the overall classification accuracy using S-ScSPM is compatible to and sometimes outperform that of original ScSPM on labeled data sets, in S-ScSPM the dictionary is learned only on one unlabeled data set, and thus the resource cost of learning is significantly reduced.

2. METHOD

2.1 Backgrounds

Formally, we use $\mathcal{D}_l$ to denote a labeled data set, and $\mathcal{D}_u$ to denote an unlabeled data set. Self-taught learning mainly consists of three stages: dictionary learning, sparse representation and classification. ScSPM consists of the same three stages, however these two methods have the following differences. First,

* Corresponding author

in self-taught learning, one trains a dictionary using $\mathcal{D}_u$, while computes sparse representations for images in $\mathcal{D}_l$ and classifies $\mathcal{D}_l$. In ScSPM, one uses $\mathcal{D}_l$ for dictionary learning stage and the same $\mathcal{D}_l$ for the latter two stages. Second, in self-taught learning, one generally uses raw pixels as image descriptors to learn a dictionary and calculate sparse codes for images, while in ScSPM, the dictionary is learned on dense SIFT features of images, and sparse codes computed using the dictionary are further processed through spatial pyramid matching (SPM) and max pooling, before fed to the classifier.

Our proposed S-ScSPM is therefore a combination of the self-taught learning framework and ScSPM. S-ScSPM consists of the same three stages as discussed above, with the details shown in Figure 1.

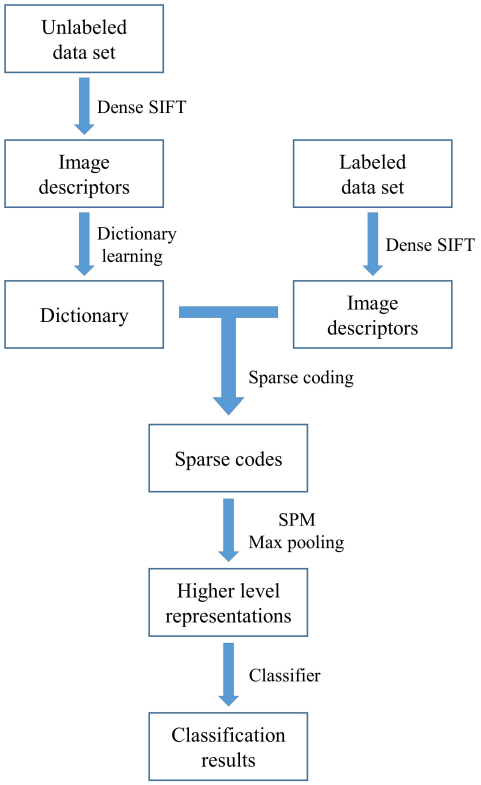

Figure 1. Flowchart of S-ScSPM

2.2 Dictionary learning

We use dense SIFT features extracted on unlabeled images to train a dictionary via dictionary learning using sparse coding. For simplicity we let $\mathcal{D}_u$ denote the set of extracted dense SIFT features. The problem of dictionary learning is defined as the following optimization problem:

$$\begin{aligned} \underset{B,a_u}{\text{minimize}} \quad & \sum_{i=1}^{N_u} ||x_u^{(i)} - B a_u^{(i)}||_2^2 + \lambda |a_u^{(i)}|_1 \\ \text{s.t.} \quad & ||B_k||_2^2 \leq 1, \ \forall k = 1, 2, ..., K \end{aligned} \tag{1}$$

where $x_u \in \mathbf{R}^d$, $a_u \in \mathbf{R}^K$, $B \in \mathbf{R}^{d \times K}$
$x_u^{(i)}$ = ith dense SIFT feature from $\mathcal{D}_u$
B = dictionary, with k denoting its kth column
$a_u^{(i)} = x_u^{(i)}$'s sparse codes under B
λ = regularization parameter that controls sparsity of $a_u^{(i)}$
K = total number of elements (columns) in B
N_u = total number of training features from $\mathcal{D}_u$

We follow the convention that B is an overcomplete basis set, i.e., $K > d$, and that an $l2$-norm constraint is applied on codeword B_k to avoid trivial solutions (Yang et al., 2009). The objective function of (1) balances two terms: (i) the first quadratic term encourages $x_u^{(i)}$ to be well reconstructed by a linear combination of all the codewords in B, with $a_u^{(i)}$ being the combination weights; and (ii) the $l1$-norm penalty on a_u forces a_u to be sparse, so that the codewords capture salient patterns in x_u. This problem is optimized by alternately updating a_u and B. B is obtained once the optimization of (1) converges. We follow the details described as (Yang et al., 2009) in our implementation.

2.3 Sparse representation

We extract dense SIFT features from labeled scene images to form $\mathcal{D}_l$, and the features are fed into the dictionary and further encoded using standard ScSPM algorithm. Specifically, with the dictionary B trained and fixed, we compute the sparse codes $a_l^{(i)}$ for feature $x_l^{(i)}$ from labeled data by optimizing

$$\underset{a_l^{(i)}, i=1,2,...,N_l}{\text{minimize}} \quad ||x_l^{(i)} - B a_l^{(i)}||_2^2 + \lambda |a_l^{(i)}|_1 \tag{2}$$

Then by ScSPM, a spatial pyramid is built over the image and participates the image into $\{2^l\}_{l=0}^L$ regions, where L is the pyramid level. A max-pooling function is applied to all M sparse codes in one region, and a higher level representation z of this region is formed by $z_j = \max\{|a_{l(1j)}|, |a_{l(2j)}|, \ldots, |a_{l(Mj)}|\}$, where z_j denotes jth component of representation vector z. Finally we concatenate z obtained from all the regions in image n as $\mathbf{z}^{(n)}$.

2.4 Classification

Once the representations $\mathbf{z}^{(n)}$ are obtained, we combine them with image labels to form a training data set $\{(\mathbf{z}^{(n)}, y^{(n)})\}_{n=1}^N$ which we use to train a multiclass linear SVM classifier. Here the label $y \in \mathcal{Y} = \{1, 2, ..., C\}$ and C denotes number of classes.

Following ScSPM, we take the one-vs-all strategy to train C linear SVM classifiers for multiclass classification, each optimizing an objective function of the form

$$\underset{w_c}{\text{minimize}} \quad ||w_c||_2^2 + C \sum_{n=1}^{N} l(w_c; y_c^{(n)}, \mathbf{z}^{(n)}) \tag{3}$$

where w_c = SVM parameters (for class c)
$y_c^{(n)} = 1$ if $y^{(n)} = c$, otherwise $y_c^{(n)} = -1$
$l(w_c; y_c^{(n)}, \mathbf{z}^{(n)})$ = the hinge loss term

Instead of the standard hinge loss, we adopt the squared hinge loss

$$l(w_c; y_c^{(n)}, \mathbf{z}^{(n)}) = [\max(0, w_c^\top \mathbf{z}^{(n)} \cdot y_c^{(n)} - 1)]^2 \tag{4}$$

such that the objective function (3) is differentiable and thus can be trained using gradient based methods.

Finally for some test data $\mathbf{z}^{(n)}$, the class label $\hat{y}$ is given by

$$\hat{y} = \max_{c \in \mathcal{Y}} w_c^\top \mathbf{z}^{(n)} \tag{5}$$

3. EXPERIMENTS AND DISCUSSION

3.1 Data sets

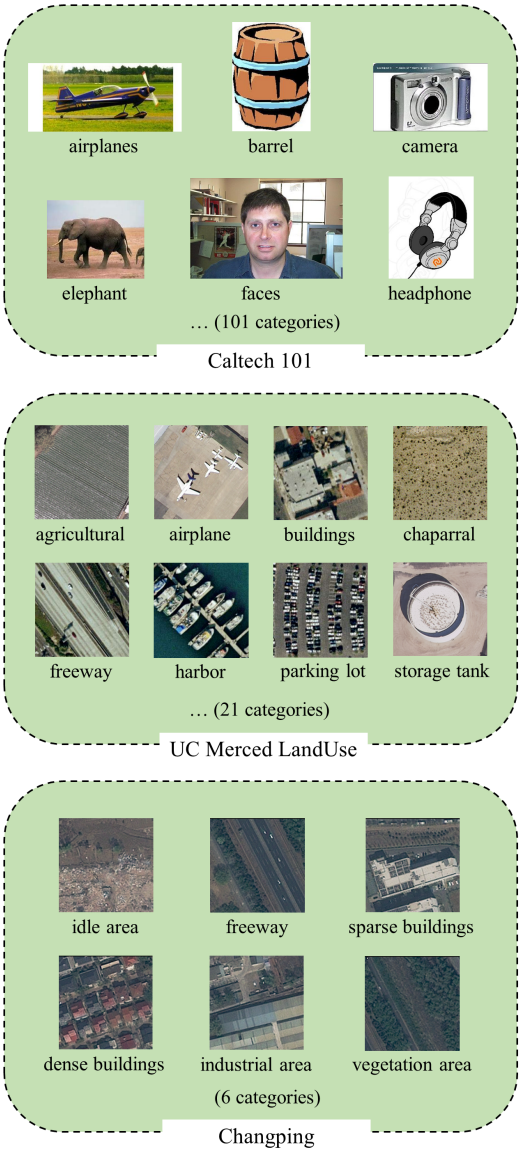

Figure 2. Sample images from three data sets

We test our proposed method on the following three data sets: we choose the Caltech-101 data set (Fei-Fei et al., 2004) for the unlabeled data set D_u, while we evaluate the performance of S-ScSPM on two labeled data sets: UC Merced LandUse (UCM) data set (Yang, Newsam, 2010) and Changping data set. Caltech-101 data set is a data set of natural scene/object images, while the latter two data sets contain only remote sensing scene images. Some typical images in these data sets are shown in Figure 2.

Caltech-101 data set contains 101 classes, including images of animals, faces, vehicles, etc. The number of images per class varies from 31 to 800, with most categories having about 50 images. The size of each image is roughly 300×200 pixels. Although for each image in this data set there is a label, the dictionary learning stage of our algorithm does not require any information from labels.

UC Merced LandUse data set is one of the most popular remote sensing scene image data sets, which contains 21 scene categories including agricultural, airplane, buildings, etc., with each category having 100 images of 256×256 pixels. The spatial resolution of these images is 1 foot.

Changping data set is acquired by the Gaofen-2 sensor and covers a certain area in Changping District, Beijing, China. The original image size is 4736×3200 pixels, with spatial resolution of 0.8m. In total 6 categories of scenes are obtained from the original image by a non-overlapping grid with a cell size of 128×128 pixels. The categories are idle area, freeway, sparse buildings, dense buildings, industrial area and vegetation area, each containing 117, 101, 154, 56, 103 and 85 images, respectively. For scenes that cannot be categorized as any of the above 6 classes, scenes being a mixture of several classes for example, we label them as "undefined" and do not use these scenes in either dictionary learning or classification. The image and annotations are shown in Figure 3.

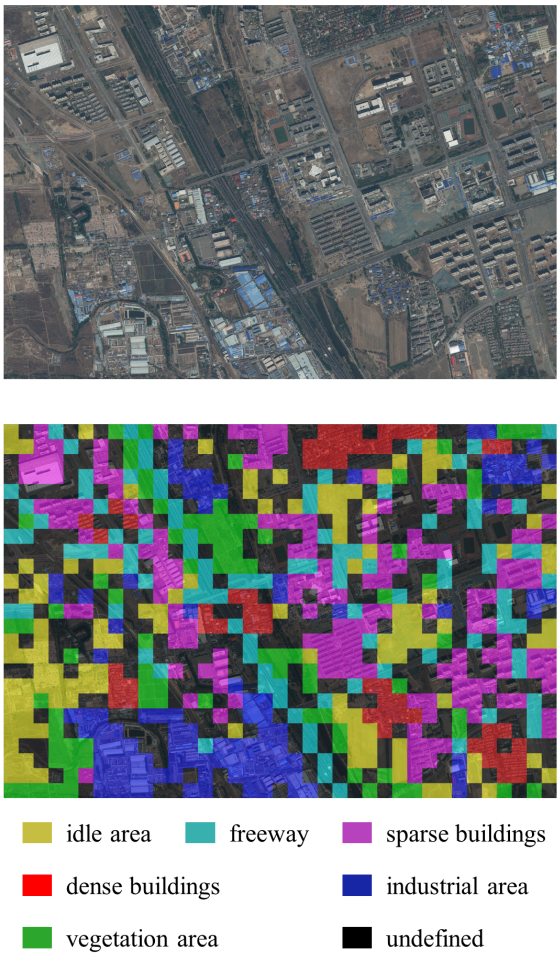

Figure 3. Changping data set

3.2 S-ScSPM v.s. ScSPM

In this section we compare our method with the original ScSPM algorithm. We evaluate and compare performance of both algorithms on UCM and Changping data set. In the experimental group, we evaluate S-ScSPM by using dictionary trained only on the Caltech-101 data set, and the SVM classifier is trained on either UCM or Changping data set. In the control group, ScSPM is performed on UCM data set and Changping data set. The dictionary is trained task-specifically.

In S-ScSPM dictionary learning stage, dense SIFT patch size is set to 16×16 pixels and step size is 8 pixels. The SIFT descriptors are extracted on gray scale images, and 200,000 descriptors are random selected and used to train the dictionary. Dictionary size K and regularization term λ are set to 1024 and 0.15, respectively.

In ScSPM dictionary learning stage, we set dense SIFT patch size to 16×16 pixels and step size is 6 pixels on UCM data set, and 8×8 pixels and 3 pixels on Changping data set. On both data sets, $K = 1024$, $\lambda = 0.15$ and number of training patches is 200,000, same as the settings in S-ScSPM.

In the sparse representation stage, for both algorithms we set the encoding regularization term λ to 0.15 and spatial pyramid level L to 2.

Following the common benchmarking procedures, we repeat all the classification experiments 10 times with different random initialization and report average classification rate with its standard deviation. Furthermore, for both experimental and control group, the SVM is trained on $20\%, 50\%$, and 80% of the data and tested on the rest.

Method	Accuracy/%		
	20% training	50% training	80% training
S-ScSPM	73.58 ± 1.19	81.86 ± 1.00	86.12 ± 1.93
ScSPM	72.08 ± 1.94	81.27 ± 1.11	85.43 ± 2.10

Table 1. Overall classification accuracy on UC Merced data set

Method	Accuracy/%		
	20% training	50% training	80% training
S-ScSPM	58.52 ± 2.03	67.56 ± 2.30	72.25 ± 4.61
ScSPM	59.82 ± 1.79	68.80 ± 1.94	72.05 ± 3.38

Table 2. Overall classification accuracy on Changping data set

Overall classification accuracy on UCM data set and Changping data set is shown in Table 1 and Table 2, respectively. On UCM data set, S-ScSPM averagely outperforms ScSPM by 1%, and the standard deviations are also smaller. On Changping data set, S-ScSPM achieves 1% lower classification rate than ScSPM under the setting of 20% and 50% training data, while slightly outperforms ScSPM when using 80% data for training. S-ScSPM has generally larger standard deviation.

The above results indicate that S-ScSPM can perform at least as well as ScSPM, and thus it is possible to learn a dictionary on a single unlabeled data set $\mathcal{D}_u$ and adopt it for classifying multiple labeled data set $\mathcal{D}_l$s, even if the distribution of images are very different between $\mathcal{D}_u$ and $\mathcal{D}_l$. Such a difference also suggests that it is possible to train a dictionary using large open source computer vision data sets like ImageNet, which is much cheaper than obtaining and labeling a large remote sensing image data set, and use the dictionary for feature encoding for further scene classification.

3.3 S-ScSPM v.s. DL based methods

In this section we compare the performance of S-ScSPM with that of several other DL based methods on UCM data set. These methods include the original BOVW, unsupervised feature learning (UFL) (Cheriyadat, 2013), multipath unsupervised feature learning (MP-UFL) (Fan et al., 2017) and bi-layer dictionary learning (BL-DL) (Yang et al., 2016). Average classification rates of the above algorithms on UCM data set, all using 80% of the data for training and repeated 10 times, are shown in Table 3.

BOVW	UFL	S-ScSPM	MP-UFL	BL-DL
76.81%	81.67%	86.12%	91.95%	93.67%

Table 3. Performance of different DL methods on UC Merced data set

In Table 3, reading from left to right, the model complexity goes higher while the classification rate gets improved. UFL basically uses SPM without sparse coding. In MP-UFL, the dictionary objective function is slightly different from S-ScSPM, and image descriptors go through a procedure called multipath consisting of multiple dictionary learning - sparse coding operations to form a higher level representation. BL-DL builds several dictionaries to seperate commonality and class-particularity dictionary atoms for better classification performance.

All algorithms listed above, except S-ScSPM, are trained task-specifically. We suspect that the self-taught learning framework may not work out fine with BOVW and UFL, for they do not require the encodings be sparse, while sparse coding is a key feature that makes self-taught learning successful. On the other hand, it can be difficult to fully combine methods like MP-UFL and BL-DL with self-taught learning, for they learn multiple dictionaries and are more complicated than ScSPM. It is possible, however, to learn some low-level dictionaries in these complex methods on unlabeled data and apply them to multiple tasks.

4. CONCLUSION

In this paper we integrate ScSPM with self-taught learning framework and propose the S-ScSPM framework for remote sensing scene classification. The experiments show that the pre-trained dictionary applies surprisingly well to both UC Merced LandUse data set and Changping data set, with our classification rate slightly outperforming traditional ScSPM. Using S-ScSPM, time and resources required for training can be significantly reduced.

ACKNOWLEDGEMENTS

We are grateful to Sensing IntelliGence and MAchine learning group, Wuhan University, for their open source remote sensing scene extract software. We adopted the software for preprocessing Changping data set.

REFERENCES

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10), 1865–1883.

Cheriyadat, A. M., 2013. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1), 439–451.

Fan, J., Chen, T., Lu, S., 2017. Unsupervised feature learning for land-use scene recognition. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4), 2250–2261.

Fei-Fei, L., Fergus, R., Perona, P., 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 conference on computer vision and pattern recognition workshop*, IEEE, 178–178.

Maurer, A., Pontil, M., Romera-Paredes, B., 2013. Sparse coding for multitask and transfer learning. *International conference on machine learning*, 343–351.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259.

Raina, R., Battle, A., Lee, H., Packer, B., Ng, A. Y., 2007. Self-taught learning: transfer learning from unlabeled data. *Proceedings of the 24th international conference on Machine learning*, 759–766.

Wu, C., Zhang, L., Zhang, L., 2016. A scene change detection framework for multi-temporal very high resolution remote sensing images. *Signal Processing*, 124, 184–197.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification. *2009 IEEE Conference on computer vision and pattern recognition*, IEEE, 1794–1801.

Yang, M. Y., Al-Shaikhli, S., Jiang, T., Cao, Y., Rosenhahn, B., 2016. Bi-layer dictionary learning for remote sensing image classification. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 3059–3062.

Yang, M. Y., Jiang, T., Al-Shaikhli, S., Rosenhahn, B., 2015. A novel dictionary learning method for remote sensing image classification. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 4364–4367.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.

Yao, X., Han, J., Cheng, G., Qian, X., Guo, L., 2016. Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6), 3660–3671.

Zhu, F., Shao, L., 2014. Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1-2), 42–59.

Zhu, Q., Zhong, Y., Zhao, B., Xia, G.-S., Zhang, L., 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. *IEEE Geoscience and Remote Sensing Letters*, 13(6), 747–751.