Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery

Grant J. Scott, *Member, IEEE*, Matthew R. England, William A. Starms, *Student Member, IEEE*, Richard A. Marcum, and Curt H. Davis, *Fellow, IEEE*

***Abstract*— Deep convolutional neural networks (DCNNs) have recently emerged as a dominant paradigm for machine learning in a variety of domains. However, acquiring a suitably large data set for training DCNN is often a significant challenge. This is a major issue in the remote sensing domain, where we have extremely large collections of satellite and aerial imagery, but lack the rich label information that is often readily available for other image modalities. In this letter, we investigate the use of DCNN for land–cover classification in high-resolution remote sensing imagery. To overcome the lack of massive labeled remote-sensing image data sets, we employ two techniques in conjunction with DCNN: *transfer learning* (TL) with fine-tuning and *data augmentation* tailored specifically for remote sensing imagery. TL allows one to bootstrap a DCNN while preserving the deep visual feature extraction learned over an image corpus from a different image domain. Data augmentation exploits various aspects of remote sensing imagery to dramatically expand small training image data sets and improve DCNN robustness for remote sensing image data. Here, we apply these techniques to the well-known UC Merced data set to achieve the land–cover classification accuracies of $97.8 \pm 2.3\%$, $97.6 \pm 2.6\%$, and $98.5 \pm 1.4\%$ with CaffeNet, GoogLeNet, and ResNet, respectively.**

***Index Terms*— Deep convolutional neural network (DCNN), deep learning, high-resolution remote sensing imagery, land–cover classification, transfer learning (TL).**

I. Introduction

FULLY automated land–cover classification is a complex problem that is constantly pushing the envelope of machine learning and computer vision. Classifying land use from remote sensing imagery is vital to monitoring and managing human development in a way that is not possible in scope and scale from a ground perspective with human-centric methods. Moreover, automated classification via machine learning is critically important for understanding the constantly changing surface of the earth, especially anthropogenic changes. Automated land–cover classification of high-resolution remote sensing imagery is a highly desirable capability, yet presents many challenges due to the sheer volume and variety of satellite and aerial images collected on a daily basis. This is further complicated by the wide variety of human-related features and scales present across geographically diverse areas on the surface of the earth.

Conventional image feature extraction (FE) and image analysis techniques have been driven by approaches using handcrafted FE. These methods typically use common computer vision FE methods such as scale invariant feature transform [1], histograms of oriented gradients [2], and bag of visual words (BoVW) [3]. A variety of other features have also been used (see [4]–[6]), each with their respective advantages and drawbacks. All of these approaches have found success to varying degrees in a wide variety of image classification modalities, including regular photography, medical imaging, and remote sensing, to list just a few. However, *convolutional neural networks* (CNNs) and deep learning have experienced a resurgence within the last few years, finally realizing the promise from their conceptual development in the 1980s starting with [7]. Deep learning techniques, such as the Caffe framework [8], GoogLeNet [9], and ResNet [10], have recently established themselves as the highest performing methods for a wide variety of machine learning tasks.

This resurgence in deep learning has created many new and promising avenues of investigation. *Deep CNNs* (DCNNs) are a common deep neural architecture that attempt to mimic the brain's ability to learn and develop hierarchical feature representations within a multilevel FE phase. These learned visual features provide the classification phase with the features necessary to discriminate images into their respective classes. Deep learning often outperforms conventional handcrafted computer vision feature extractors, such as those mentioned previously.

However, DCNN require very large labeled training data sets to allow the complex networks to sufficiently learn both FE and classification simultaneously. Unfortunately, large remote sensing image data sets with labels are not readily available. In order to mitigate the limited availability of labeled training data, we accelerated learning by applying transfer learning (TL) from pretrained DCNN models, CaffeNet [8], GoogLeNet [9], and ResNet [10], and expanding the remote sensing image training data set through augmentation strategies. CaffeNet, GoogLeNet, and ResNet are prime candidates for TL, also known as inductive transfer [11], as they were originally trained for FE and classification in a similar domain:

Manuscript received January 27, 2016; revised June 30, 2016 and December 3, 2016; accepted December 20, 2016. Date of publication February 17, 2017; date of current version March 3, 2017. *(Corresponding author: Grant J. Scott.)*

The authors are with the Center for Geospatial Intelligence, University of Missouri-Columbia, Columbia, MO 65211 USA (e-mail: GrantScott@missouri.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2017.2657778

Fig. 1. Sample image chips from the UCM data set showing examples from all 21 land-cover classes.

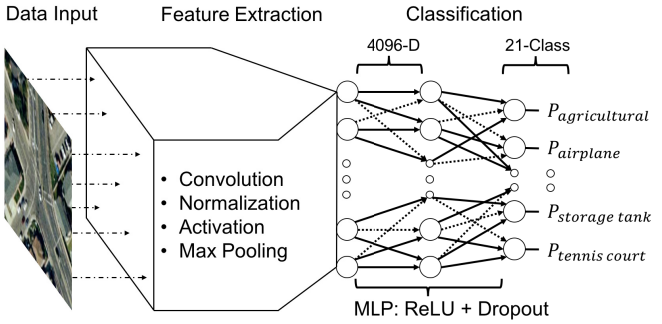

Fig. 2. Deep neural network architecture comprising two phases: visual FE and land–cover classification (21-class UCM shown).

ground-based digital photos. By initializing our deep learning with a pretrained model, we greatly decrease the amount of training time required by repurposing and refining the already learned weights in the FE phase.

II. Data Sets and Prior Research

Many researchers have used the well-known *UC Merced (UCM) Land Use* data set [3] to investigate land–cover classification methods for high-resolution remote sensing imagery. The UCM data set is a high-resolution remote-sensing image data set with 21 different labeled classes. A major drawback of using the UCM data set in deep neural network techniques is its relatively small size, only 100 images per class, which must then be divided between training and validation data. Fig. 1 shows a sample image chip from each class of the UCM data set.

The UCM data are widely used. A few noteworthy recent results include [5] with 90.6% overall classification accuracy using completed LBP, [12] at 94.3% using compressive sensing techniques, and [6] at 89.1% using a pyramid of spatial relations. Li *et al.* [13] applied unsupervised multilayer feature learning using two layers of FE and K-means clustering to train a deep neural network and then performed K-nearest neighbor classification. Luus *et al.* [14] created a deep network from scratch and achieved a 93.5% accuracy on the UCM data. Other approaches have classified UCM, such as 81.2% [3] and 77.8% [15], with methods like spatial partitioning and BoVW, respectively. Marmanis *et al.* [16] used the OverFeat model [17], which like CaffeNet is a slight variant on AlexNet [18], as an FE phase. They then passed the last two hidden MLP layers as input into a smaller CNN and achieved 92.4%. Penatti *et al.* [19] achieved 93.4% for CaffeNet and then 99.43% by combining CaffeNet with OverFeat FE outputs that were fed into an SVM.

In addition to the UCM data set, we evaluated another remote sensing image data set from [20] that includes 19 land–cover classes with at least 50 images/class. This second data set, referred herein as RSD, is characteristically different from UCM where the RSD source image chips are 600×600 pixels and are JPEG compressed. We start by downscaling the RSD image chips to 256×256 pixels so that the augmentation pipelines described in Section V for UCM are repeatable. For the RSD data set, Chen *et al.* [5] recently achieved 93.4% classification accuracy and the current state of the art is [21] with 93.6% overall accuracy.

III. Deep Network Architecture

In this letter, we first utilize CaffeNet [8] as the seed of our deep learning structure, which is a derivative of AlexNet [18]. Fig. 2 provides a high-level view of our deep neural network structure, where remote sensing image chips are fed into the network and a classification confidence vector is generated at the output. Our DCNN is broken down into two primary phases, each with various sublayers. The first phase of the network is FE, where the deep visual features are extracted. The FE uses multiple stages of convolution, activation, max pooling, and normalization in various sequences. The readers interested in further details of the CaffeNet architecture are encouraged to review [8], [18]. The second phase of our DCNN is the classification phase comprising two sets of inner product layers, each with 4096 nodes, and a final node layer (21 nodes for UCM and 19 for RSD) for land–cover classification.

The most critical aspect of our DCNN is the FE phase that extracts the relevant visual features. The convolutional layers apply a set of convolution kernels across the input image space. Unlike handcrafted FE techniques, the weights in a convolutional layer are learned during training, adapting to input data simultaneously while learning classification. These learnable filters provide deep FE over the input space, extracting visual features from the imagery that help distinguish complex land–cover classes from each other. The pooling layers apply simple aggregate statistical kernels over the input space, such as mean and max. For in-depth analysis of the various layers and their effect on network performance and the training hyperparameters, the reader is referred to [18].

The DCNN classification phase uses the extracted high-level visual features to classify the image chips. This is accomplished by feeding the output features into a $4096 \times 4096 \times N$ multilayer perceptron (MLP) neural network. This phase utilizes rectified linear units [22] to provide a nonsaturating activation function defined as $f(x) = \max(0, x)$. The two 4096-D MLP layers are also configured to use 50% dropout to reduce overfitting [18]. The final layer of the DCNN is 21 or 19 nodes (N) corresponding to the land–cover classes of the UCM or RSD data set, respectively.

IV. Transfer Learning

TL applies knowledge previously learned in one domain to another domain. In the context of this effort, TL uses a

Fig. 3. UCM image chip with a few illustrative rotations. (Top) Seed image and seven rotations. (Botttom) Transposed image and seven rotations, in total 60 for each seed and transpose image. Rotations are 0°, 7°, 90°, 97°, 180°, 187°, 270°, and 277°, respectively, in each row.

previously trained model for a new task, where the original model was trained using a very large data set of images. We bootstrapped our DCNN FE phase with the learned FE weights originally generated by the training on the ImageNet data [18] (1.2 million regular photographic images). Pan and Yang [23] survey how TL can save a significant amount of training effort. Others, such as Marmanis *et al.* [16] and Penatti *et al.* [19], have investigated TL for land-cover classification as well. For our limited data set, this technique provides a significant head start toward learning appropriate visual features salient for land–cover classification.

As previously stated, the UCM data set is very small relative to the connectivity in our DCNN, with only 2100 images in total. In one TL approach, the FE phase of the pretrained network can be completely locked, meaning the weights of the convolution filters are not updated and the pretrained network is used solely to generate features. Locking the FE phase prevents extractor weight modification during retraining, forcing the classifier to interpret the output of the extractor as it was provided. This is the approach adopted in much of the previous work, where the extracted features are taken from the very last 4096-node layer. However, in this effort, we also investigated further refinement of the convolution layers during the training of the DCNN classification phase to assess potential improvement over the locked TL approach. In this alternate approach, we allowed the FE weights to also be modified during training so that the prior extractor weights from CaffeNet, GoogLeNet, and ResNet could be *fine-tuned* for the remote sensing imagery. We utilized a small learning rate, 0.001, for FE fine-tuning and empirically evaluated a few momentum settings. We found that momentum values of 0.3, 0.6, and 0.9 generated classification accuracies with insignificant differences: $97.67 \pm 2.5\%$, $97.90 \pm 2.3\%$, and $97.71 \pm 2.5\%$, respectively, for the augmented UCM data.

V. Data Augmentation

Deep learning requires a significant volume and a variety of training data to develop the complex network structures that are required to achieve the best classification accuracies. This is a significant problem for the UCM and RSD data sets, which have, in total, only 2100 and 950 images, respectively. As a result, we adopted a novel strategy to augment the limited amount of training image data in this study to enable more robust deep network learning. This was done by increasing the volume and variety of the training data presented to our DCNN. While the Caffe library provides basic data augmentation tools, remote sensing image data have specific properties that can be exploited to further augment the training data beyond the basic capabilities provided by the Caffe library.

We applied various transforms to the seed images, including image transpose and rotations, to augment the UCM and RSD training data. The motivation for this approach is as follows: remote sensing imagery is generally captured from a viewpoint significantly high above the regions and/or objects of interest. In this case, there is not an upright orientation in observed classes, such as buildings, roads, and land–cover types (forest, water), as there is in ground-level images (e.g., the sky is up). Whereas recognition of objects and scenes in ground photographs may rely on the spatial distribution and configuration relative to expected sensor orientation, this alignment is not expected in remote sensing imagery acquired from aerial and/or space borne platforms. Therefore, augmentation allows the network to learn feature classification *without regard to feature orientation*. This is especially important for small remote sensing data sets where the limited training data may unintentionally have feature orientation biases that are completely irrelevant for object or scene recognition.

Image transpositions create horizontally and vertically mirrored versions of a given seed image. We then apply a series of rotations to the seed and transposed images to expand the available training data. A sample of these rotations is provided in Fig. 3 (top and bottom) for a seed image and a transposed version of the same seed image, respectively. Labels and the associated seed image are linked to the new augmented data sets, allowing a massive expansion of the available labeled training data. One important consideration is the size of the training seed images relative to the pretrained network's input image size. Being unaltered, the CaffeNet FE phase inputs a 227×227 window of image pixels. Since the UCMerced seed images are 256×256, this enables the rotation of the seed and transposed images by more than basic 90° increments. In fact, each image can be rotated $-7°$ through $+7°$ from each 90° rotated image (15 rotations for each 90° basic rotation) and still provide valid pixels in the entire input window required by CaffeNet. Rotations exceeding this $\pm 7°$ range from the cardinal 0°, 90°, 180°, and 270° would introduce values not

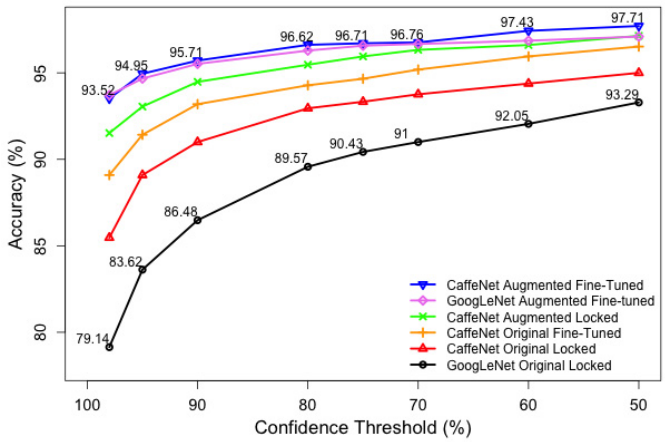

Fig. 4. Overall land–cover classification accuracy with varying confidence thresholds for four CaffeNet and two GoogLeNet-derived DCNNs.

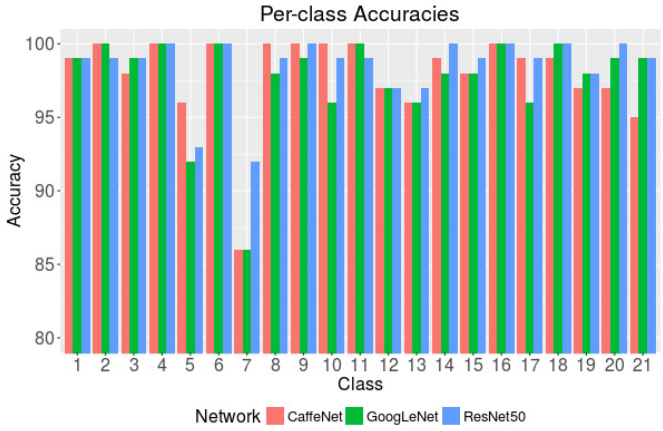

Fig. 5. Per-class land–cover classification accuracies for the UCM data set for the highest performing CaffeNet, GoogLeNet, and ResNet DCNNs evaluated using augmented training data and FE tuning.

part of the original image chip, e.g., no data pixels. As the final step in our augmentation, we center crop all images to 227 × 227 pixels and then train our DCNN. Using this approach, we expanded the UCM data set by 120× from 2100 to 252 000 images.

VI. Results

We first evaluated land–cover classification performance using a CaffeNet-derived DCNN for the UCM data set in four sets of experiments. The experiments sets were as follows:

1) original data set without augmentation with no FE fine-tuning;
2) original data set without augmentation with FE fine-tuning;
3) augmented data set with no FE fine-tuning;
4) augmented data set with FE fine-tuning.

In addition, we also evaluated classification performance for the UCM data set using a GoogLeNet-derived DCNN for: 1) original data set without augmentation with no FE fine-tuning and 2) augmented data set with FE fine-tuning. All experiments used fivefold cross validation, i.e., within each of the five folds, 80% of the data were used for training the DCNN and the remaining 20% were withheld for validating the DCNN. As mentioned in Section V, seed images are *always* linked to their augmented forms. This ensures that augmented forms of a seed image do not simultaneously occur in both the training and validation data sets. Fig. 4 shows the overall land–cover classification accuracy of the DCNN when using an increasing confidence value cut-off threshold, i.e., we generated a set of receiver operating characteristic (ROC) plots. Notably, the classification accuracies exceeded 95% for both the CaffeNet and GoogLeNet DCNNs at a 90% confidence level when the augmented data sets with fine-tuning FE were used. Our experiments demonstrate that the CaffeNet DCNN with the original data (i.e., no augmentation) and locked FE achieved a 95.0 ± 3.2% classification accuracy for the UCM data set, while fine-tuning the FE increased performance to 96.5 ± 2.8% for the original data set. The CaffeNet DCNN trained with the augmented data and locked FE achieved 97.1 ± 2.7% classification accuracy, while fine-tuning increased performance to 97.8 ± 2.3%. GoogLeNet achieved 93.9 ± 3.9% and 97.6 ± 2.6% on the UCM data set for original data with locked FE and augmented data with fine-tuned FE, respectively. When tabulating the overall accuracy, an image was considered correctly labeled if and only if the correct class was the maximum value in the network's softmax classifier output vector [18].

Fig. 5 provides the per-class breakdown for the CaffeNet and GoogLeNet DCNN with augmented data and fine-tuned FE. The class labels correspond to the left-to-right and top-to-bottom order of Fig. 1 and are as follows: 1) *Agricultural*; 2. *Airplane*; 3) *Baseball diamond*; 4) *Beach*; 5) *Buildings*; 6) *Chaparral*; 7) *Dense residential*; 8) *Forest*; 9) *Freeway*; 10) *Golf course*; 11) *Harbor*; 12) *Intersection*; 13) *Medium residential*; 14) *Mobile home park*; 15) *Overpass*; 16) *Parking lot*; 17) *River*; 18) *Runway*; 19. *Sparse residential*; 20) *Storage tanks*; 21) *Tennis court*. The CaffeNet, GoogLeNet, and ResNet DCNNs all achieved 100% classification accuracy for the *Airplane*, *Beach*, *Chaparral*, and *Parking lot* image classes. As with prior work on the UCM data set, the various residential classes (*Dense*, *Medium*, and *Sparse residential* and *Mobile home park*) were the most difficult to correctly classify. These classes can be seen in Fig. 1 to be composed of similar object content, differing primarily in the density of the objects within the image chip. Specifically, images for these classes typically contain trees, houses, and roads but with variable amount/frequency. Table I summarizes the classification results on the two data sets and the networks evaluated in this letter. We conducted experiments for the RSD data set using only the data augmentation and FE fine-tuning approach given that this produced the best result for the UCM data experiments. The CaffeNet DCNN achieved an overall classification of 97.3 ± 3.2%, while the GoogLeNet DCNN achieved 98.1 ± 3.1% overall classification.

Furthermore, we have conducted experimentation using our data augmentation and fine-tuning on the ResNet50 [10] architecture. ResNet architectures have residual connections that bypass two or more convolutional layers at a time and are an adaptation of the VGG net models [24]. ResNet50, ResNet101,

TABLE I

CLASSIFICATION ACCURACIES: CaffeNet, GoogLeNet, AND ResNet50 USING DATA AUGMENTATION AND FINE-TUNING

UCM		RSD	
CaffeNet	97.8	CaffeNet	97.3
GoogLeNet	97.6	GoogLeNet	**98.1**
ResNet50	**98.5**	ResNet50	97.8

and ResNet152 models bypass a stack of three convolutional layers at a time with residual connections, which allows better error back propagation. We have conducted testing with the ResNet50 network and achieved an accuracy of $98.5 \pm 1.4\%$ on the UCM data and the ResNet50 DCNN achieved $97.8 \pm 2.7\%$ on the RSD.

VII. CONCLUSION

In this letter, we demonstrated DCNNs that: 1) incorporated a novel data augmentation strategy designed specifically to improve network training for remote sensing imagery and 2) utilized fine-tuning of the FE phase for TL. Three DCNNs, derived from CaffeNet, GoogLeNet, and ResNet50, respectively, were shown to significantly improve overall land–cover classification accuracies on two different remote sensing data sets (UCM and RSD). Fine-tuning of the FE phase improved the classification accuracies for both the original and augmented data sets. However, data augmentation was shown to deliver the best performance improvement, increasing the overall classification accuracies for both the locked and fine-tuned FE cases. As shown in Table I, using FE fine-tuning and data augmentation on the UCM data set, we achieved 97.6%, 97.1%, and 98.5% overall classification accuracies from the CaffeNet, GoogLeNet, and ResNet50-derived DCNNs, respectively. This includes 100% classification accuracy for many of the classes within the data set. Additionally, we achieved state-of-the-art results on the RSD data set of 97.3%, 98.1%, and 97.8% overall classification accuracies for CaffeNet, GoogLeNet, and ResNet50 DCNN, respectively (shown in Table I). As deep learning network models become more and more complex, the use of TL with FE fine-tuning and data augmentation will become increasingly important. Finally, we are making all our data augmentation, cross-fold, training preparation, and other utility scripts available to the community. These are available at github.com/scottgs/ML_DataAugmentation.

REFERENCES

[1] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 3278–3285.

[2] R. Negrel, D. Picard, and P.-H. Gosselin, "Evaluation of second-order visual features for land-use classification," in *Proc. 12th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2014, pp. 1–5.

[3] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (ACM GIS)*, 2010, pp. 270–279.

[4] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[5] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal, Image Video Process.*, vol. 10, no. 4, pp. 745–752, 2016.

[6] S. Chen and Y. L. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

[7] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.

[8] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[9] C. Szegedy *et al.* (2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/1409.4842

[10] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[11] T. G. Dietterich, L. Pratt, and S. Thrun, "Special issue on inductive transfer," *Mach. Learn.*, vol. 28, no. 1, p. 5, Jul. 1997.

[12] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, Oct. 2015.

[13] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, Feb. 2016.

[14] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015.

[15] Y. Jiang, J. Yuan, and G. Yu, "Spatial partition for scene recognition," in *Computer Vision—-ECCV* (Lecture Notes in Computer Science), vol. 7573, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 730–743.

[16] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Red Hook, NY, USA: Curran & Associates Inc., 2012, pp. 1097–1105.

[19] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.

[20] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.

[21] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[22] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, Apr. 2011, pp. 315–323.

[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[24] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556