Feature Fusion with Deep Supervision for Remote-Sensing Image Scene Classification

Usman Muhammad, Weiqiang Wang
School of Computer and Control Engineering
University of Chinese Academy of Sciences
Beijing, China
usman@mail.bnu.edu.cn

Abdenour Hadid
Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu
Finland

***Abstract*—The convolutional neural networks (CNNs) have shown an intrinsic ability to automatically extract high level representations for image classification, but there is a major hurdle to their deployment in the remote-sensing domain because of a relative lack of training data. Moreover, traditional fusion methods use either low-level features or score-based fusion to fuse the features. In order to address the aforementioned issues, we employed a deep supervision (DS) strategy to enhance the generalization performance in the intermediate layers of the AlexNet model for remote-sensing image scene classification. The proposed DS strategy not only prevents from overfitting, but also extracts the features more transparently. Secondly, the canonical correlation analysis (CCA) is adopted as a feature fusion strategy to further refine the features with more discriminative power. The fused AlexNet features achieved by the proposed framework have much higher discrimination than the pure features. Extensive experiments on two challenging datasets: 1) UC MERCED data set and 2) WHU-RS dataset demonstrate that the two proposed approaches both enhance the performance of the original AlexNet architecture, and also outperform several state-of-the-art methods currently in use.**

***Index Terms*—Pre-trained AlexNet, Canonical Correlation Analysis(CCA), Deep Supervision (DS), Scene Classification**

I. Introduction

With the rapid increase of remote sensing satellites over the past decade, the mutil-angle and high-resolution remote sensing (HRRS) images are available to study the structural and spatial patterns with large detail. However, inter-class similarity among scene categories or identical land-covers make the classification tasks very challenging. For example, images from river and beach, which are two typical scene categories, may both consists of water, trees, and boats at the same time but differs in the density and spatial distribution of these three thematic classes. In this regard, it becomes a crucial task for researcher to formulate efficient and effective descriptors for the scene classification and existing methods can be roughly divided into three main classes [1] : hand-crafted or manually created methods (low-level), mid-level methods, and deep feature learning based methods (high-level). Recently, a significant progress has been made for learning high-level semantic features due to the development of convolutional neural network (CNN), since it has an end-to-end advanced structure to efficiently encode spectral and spatial information based on stack of convolutional filters. One approach is to use pre-trained CNN as a activation vector (transfer learning). Another approach is to construct the CNN from the start. However, the back-propagation process, the stochastic gradient descent (SGD) strategy or training a deep CNN from the start is a highly time-consuming work. Moreover, overfitting is the main reason due to small number of training samples. To achieve transparent and better representation of hidden layers, a deep supervision (DS) strategy is introduced [8]. It is an effective companion objective operation which increase transparency of the intermediate layers to boost the classification performance. DS strategy can also minimize the gradient vanishing and help CNN model to prevent overfitting.

In order to use multi-layer features, features fusion is an efficient step to get the benefits of transfer learning features in scene understanding. Among fusion methods, the multi kernel learning (MKL) [3] and metric learning (ML) [4] are the famous approaches to fuse features from different layers of the CNN model. The work in [5], fused the fully connected layer features based on the discriminant correlation analysis (DCA). In [6], authors attempt to fuse the convolutional and fully-connected layer features based on fisher kernel coding approach (MIFK). Nonetheless these methods present an improved classification accuracy, no literature has attempted to use canonical correlation analysis (CCA) [7] as a feature fusion strategy to fuse the different layers of AlexNet model, so far. This make us so inquisitive to examine the impact of fusing the AlexNet for the application of remote-sensing image scene classication. Therefore, we attempt to address the problem of remote sensing classification by fusing AlexNet features and give detail of the fusion, how it can be explored to make a complementary feature space. The major contributions of this article can be summarized as follows.

(1) A simple but an efficient data expansion technique based on affine transformations is proposed to expand the training data.

(2) A deep supervision strategy is proposed to prevent overfitting and to maintain the features robustness through both final-layer supervision and intermediate-layer supervision of the pre-trained AlexNet architecture.

(3) We employ CCA transformations and combine the features of two fully-connected layers of the model into a single one with more discriminative power, which allows an

improved classification performance.
The remainder of this paper is organized as follows. In section II, two proposed scenarios are explained using the pre-trained AlexNet architecture. Experiments on two data sets and the results are presented in section III. Finally, in section IV , we draw conclusions about the proposed methods.

II. The Proposed Methodology

In this section, two scenarios are proposed for investigating the effectiveness of pre-trained AlexNet architecture [21]. In the first scenario, we incorporate a deep supervision strategy, and in the second scenario, a feature fusion approach is proposed to combine relevant information from two fully connected layers into a single one with more discriminative power than any of the input feature vectors.

A. Scenario(I): The deep supervision (DS) strategy

The goal of the pre-trained AlexNet architecture is to provide supervision at the output layer and propogating this supervision back to earlier layers. However, this single supervision is not enough because the features learned at hidden layers (early hidden layers in particular) are not always transparent to deal with the classification error minimization. To alleviate the single supervision in the pre-trained AlexNet architecture, and to provide integrated direct supervision to the hidden layers, a deeply-supervised (DS) [8] strategy is incorporated into the model. DS is a strong convex strategy, and advantage of such integrated deep supervision is evident on three aspects of:(1) to minimize the non-transparency in the intermediate layers of the Alexnet architecture; (2) discriminativeness and robustness of learned features, through both final-layer supervision and intermediate-layer supervision; (3) training effectiveness in the face of "exploding" and " vanishing" gradients. This integrated direct hidden layer supervision is added by the companion objective function for each hideen layer, and can be regarded as an additional constraint within the learning process. An illustration is given below.

Let's assume that the input sample $x_i \in R^n$ comprises the raw input data and $y_i \in \{1, \cdots, K\}$ represents the corresponding ground truth label for sample X_i. Suppose that M denotes the total number of layers in the pre-trained Alexnet architecture, $W = (W^{(1)}, ..., W^{(M)})$ are the weight combinations of the model, Meanwhile, the corresponding weights $w = (w^{(1)}, ..., w^{(M-1)})$ are associated for each classifier in each hidden layer. Equations (1) and (2) refers the weight parameters and the filters of the network:

$$Z^{(m)} = f(Q^{(m)}) \; and \; Z^{(0)} = x \tag{1}$$

$$Q^{(m)} = W^{(m)} * Z^{(m-1)} \tag{2}$$

where the specific layer of the pre-trained model is denoted by m. $W^m, m = 1...M$ are the weights to be learned; $Q^{(m)}$ concerns to the convolved responses based on the previous map, and $f(\cdot)$ is a pooling function on Q. The total object function is defined in Equation (3).

$$F(W) = P(W) + Q(W) \tag{3}$$

P(W) denotes the output objective, and $Q(W)$ is the summed companion objectives, defined in Equations (4) and (5), respectively.

$$P(W) = ||w^{(out)}||^2 + L(W, w^{(out)}) \tag{4}$$

$$Q(W) = \sum_{m=1}^{M-1} \alpha_m [||w^{(m)}||^2 + l(W, w^{(m)} - r]_+ \tag{5}$$

where the classifier weight for the output layer is denoted by $w^{(out)}$. The final combined objective function of the architecture is shown in Equation (6).

$$||w^{(out)}||^2 + L(W, w^{(out)} + \sum_{m=1}^{M-1} \alpha_m [||w^{(m)}||^2 + l(W, w^{(m)} - r]_+ \tag{6}$$

where $||w^{(out)}||^2$ and $L(W, w^{(out)})$ are respectively the margin and the (squared) hinge loss of the support vector machine (SVM) classifier. For the AlexNet architecture, $||w^{(m)}||^2$ and $l(W, w^{(m)})$ are respectively the margin and (squared) hinge loss of the classifier at each hidden layer. The overall loss of the output layer and the companion loss of the intermediate layers is defined by:

$$L(W, w^{(out)}) = \sum_{y_k \neq y} [1 - < w^{(out)}, \phi(Z^{(M)}, y) - \phi(Z^M, y_k) >]_+^2 \tag{7}$$

and

$$l(W, w^{(m)}) = \sum_{y_k \neq y} [1 - < w^{(out)}, \phi(Z^{(M)}, y) - \phi(Z^M, y_k) >]_+^2 \tag{8}$$

From the above formulations , it can be observed, the deep supervision via companion objectives make the pre-trained AlexNet architecture not only learn the convolutional kernels $W^\star$, but implie as an additional (soft) constraint at each hidden layer of the architecture (or as a new regularization) within the learning process. It is illustrated [8] that for each $l(W, w^{(m)})$, the $w^{(m)}$ directly depends on $Z^{(m)}$, which is dependent on $W^1, ..., W^m$ up to the mth layer. During the course of training, the second term often goes to zero. In this regard, the aim of enhancing the performance at output layer is not altered and the companion objective acts in the form of proxy or regularization for discriminative features. In order to perform this operation, we set γ (a hyper parameter) as a threshold in the second term of Eq.(6).In this way, the companion objective function vanishes once falls to γ (or below) and will stop contributes to the gradient update during learning process. The mth balance parameter α_m emphasizes the importance of the error between the output objective and the companion objective. To make it more clear, diagrammatical representation is provided in Fig.1.

B. Scenario(II): Features Fusion Based on Canonical Correlation Analysis

Feature level fusion in scene classification is believed to be more effective than the other levels of fusion (score-based), and an intuitive way of integrating these features is to concatenate them into a single vector [3]. Our objective is to combine relevant information from two fully connected layers of pre-trained AlexNet model into a single one with more discriminative power to construct final representation

of scenes images. In this regard, we adopt the canonical correlation analysis (CCA) [7], a feature level fusion technique that establish the correlation criterion function between the two groups of feature vectors, to extract their canonical correlation features. Suppose that $X \in R^{p\times n}$ and $Y \in R^{q\times n}$ represent BoW features of n training images. p and q are the dimensions of each vector.

Let's assume that $S_{xx} \in R^{p\times p}$ and $S_{yy} \in R^{q\times q}$ define the within sets covariance matrices of X and Y respectively. Furthermore, the $S_{xy} \in R^{p\times q}$ is referred the between set covariance matrix such as $S_{yx} = S_{xy}^T$. The overall covariance matrix S is then written as:

$$S = \begin{pmatrix} cov(x) & cov(x,y) \\ cov(y,x) & cov(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \quad (9)$$

It is complicated to follow the relationship between these two sets of vector from matrix S because these feature sets may not comply a consistent pattern. However, the objective of CCA is to find the linear combinations, $\overset{*}{X} = W_x^T X$ and $\overset{*}{Y} = W_y^T X$, such that to maximizes the pair-wise correlations across the two feature sets. The pairwise correlation is defined as:

$$corr(\overset{*}{X}, \overset{*}{Y}) = \frac{cov(\overset{*}{X}, \overset{*}{Y})}{var(\overset{*}{X}).var(\overset{*}{Y})} \quad (10)$$

Where $cov(\overset{*}{X}, \overset{*}{Y}) = W_x^T S_{xy} W_y, var(\overset{*}{X}) = W_x^T S_{xx} W_x$ and $var(\overset{*}{Y}) = W_y^T S_{yy} W_y$. Maximization is conducted using Lagrange multipliers subject to satisfy the following variables $var(\overset{*}{X}) = var(\overset{*}{Y}) = 1$. Both transformation matrices, W_x and W_y, can be modeled using the eigenvalue equations:

$$\begin{cases} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \hat{W}_x = R^2 \hat{W}_x \\ S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \hat{W}_y = R^2 \hat{W}_y \end{cases} \quad (11)$$

where $\hat{W}_x$ and $\hat{W}_y$ are the eigenvectors and R^2 is the eignvalue diagonal matrix. Hence, it is possible to perform feature-level fusion either by concatenation or summation of the transformed feature vectors and can be represented as:

$$Z_1 = \begin{pmatrix} \overset{*}{X} \\ \overset{*}{Y} \end{pmatrix} = \begin{pmatrix} W_x^T X \\ W_y^T Y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \quad (12)$$

or

$$Z_2 = \overset{*}{X} + \overset{*}{Y} = W_x^T X + W_y^T Y = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \quad (13)$$

where Z_1 and Z_2 are called the Canonical Correlation Discriminant Features (CCDFs). The architecture of AlexNet with CCA is schematically illustrated in Fig.1. Hence, the feature fusion can be performed either concatenation or summation using the above Eq.(12)(13). Both layers features (FC6 and FC7) are fed into a canonical correlation analysis (CCA) subspace to compute the new transformations. After performing CCA, we can simply concatenate the new transformed features into single and final feature vector to get the final classification performance. This simple concatenation further improves the classification accuracy as shown in Table I.

TABLE I
OVERALL CLASSIFICATION ACCURACY (%)OF REFERENCE AND PROPOSED METHODS ON THE UC-MERCED AND WHU-RS DATA SETS.

Methods	UC-Merced	WHU-RS
GoogLeNet [20]	92.80±0.61	93.00
FK-S [23]	91.63±1.49	-
OverFeat [19]	90.91±1.19	-
MARTA GANs [22]	94.86±0.80	-
D-CNN with AlexNet [4]	96.67±0.10	-
D-CNN with GoogLeNet [4]	97.07±0.12	-
KCRC [12]	93.80±0.58	93.70±0.57
MTJSLRC [13]	91.07±0.67	91.74±1.14
MS-CLBP1 [14]	90.60±1.40	93.30±0.80
CaffeNet [16]	95.02±0.81	96.24±0.56
VGG-VD-16 [16]	95.21±1.20	96.05±0.91
Fusion by addition [5]	97.42±1.79	98.70±0.22
MLF [6]	89.62±1.67	88.16±2.76
AlexNet-SPP-SS [15]	96.67±0.94	95.00±1.12
SPP-net-+MKL [3]	96.38±0.92	95.07±0.79
salM^3LBP-CLM [2]	95.75±0.80	96.75±0.86
Pre-trained AlexNet	94.30±0.80	95.60±1.54
Pre-trained-AlexNet-Aug	**96.63±0.40**	**97.13±0.60**
Pre-trained-AlexNet-Aug-DS	**97.10±0.50**	**97.90±0.50**
Pre-trained-AlexNet-Aug-DS-CCA	**97.80±0.30**	**98.80±0.30**
Fusion by concatenation (Proposed)	**98.10±0.20**	**99.17±0.20**

III. DATASETS AND EXPERIMENTAL SETUP

In this section, we explore the effectiveness of the proposed methods, and compare with several state-of-the-art approaches on two datasets, namely the widely utilized UC Merced dataset [9] , and the WHU-RS dataset [10].

A. Data sets

The UC Merced dataset is acquired from the USGS National Map Urban Area Imagery collection, contains 21 distinctive scene categories. The image size is 256×256 pixels with a pixel resolution of one foot. Each class is composed of 100 samples. The second data set chosen for evaluating the performance of the proposed AlexNet-DS-CCA model is the public WHU-RS data set. It consists of 950 images with a size of 600×600 pixels with various resolutions, illumination, scale, collected from Google Earth, which make it more complicated than the UCM dataset. This dataset is relatively small than the UC Merced dataset because of containing 19 scene categories.

B. Experimental Setup

In each experiment, the images are resized to $227 \times 227 \times 3$ for the pre-defined size requirement of the AlexNet model. To increase the diversity and preventing from overfitting issue, we algorithmically expand both data sets by applying affine transformations to the raw images. This simple data augmentation was adopted by incorporating three types of transformations including rotation, translation and scaling to all the images. Rotation was done by rotating the images 40 degree clockwise using bilinear interpolation. Translation was peformed by using the translation vector [80 80]. Images were scaled using factor 5 by bicubic interpolation. For training and

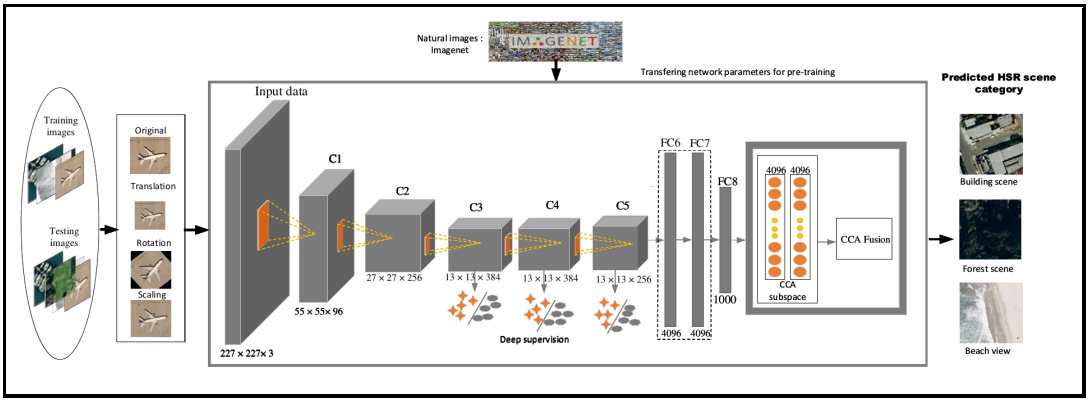

Fig. 1. The pre-trained AlexNet with affine transformations, deep supervision and canonical correlation analysis. First, the affine transformations are used as a data augmentation to expand the datasets. The Deep Supervision (DS) approach is used to prevent from overfitting. Canonical Correlation Analysis (CCA) based transformations are performed to fuse the features in order to obtain discriminative feature representation. Finally, features are simply concatenated to obtain the final classification performance.

testing, The linear SVM is trained on a set of 80% samples per category and validated on the rest of data. We use LIBSVM library package for the linear SVM. The pre-trained model used in this paper is available on Caffe Model Zoo [11]. For the scenario I, the optimization is performed using the SGD algorithm in MATLAB and the gradient functions in a same way following the AlexNet architecture. For the scenario II, we utilize only the 4096-dimensional activation vectors from the first and second fully-connected (FC) layers in a feed forward way to CCA subspace, and then fuse these features sets by concatenation. In an overall view, the pre-trained AlexNet-DS-CCA architecture, as shown in Fig.1, illustrates the proposed methods with affine transformations to obtain an improved classification performance. All the experiments are conducted using MATLAB 2017b on an Intel Core i7-3770 with a 8GB of RAM memory. The experiments were conducted 5 times to obtain convincing results for both datasets.

C. Results

In Table I, results from other state-of-the-art methods are shown, and summarized here for a comparison analysis. To generate a class-specific codebook, an improved class-specific codebook using kernel collaborative representation based classification (KCRC) is proposed in [12]. Multiple features, e.g., shape, color and textual features, are used in [13]. Then, a multi-task joint sparse and low-rank representation is adopted to combine the features. The fisher kernel (FK) coding framework is introduced to extend the BOVW model in [23], by characterizing the low-level features with a gradient vector. Authors report in [14], introduce the completed local binary patterns (CLBP) operator for the first time on remote sensing land-use scene classification. Unsupervised learning is proposed [22] to learn a representation using only unlabeled data. In [5], discriminant correlation analysis (DCA) is used to prove that a feature fusion can be performed efficiently by fusing two fully connected layers of VGG-Net architecture. The AlexNet is explored with spatial pyramid pooling (SPP-net), and then transfer learning is performed to ensure the effectiveness of each layer. In order to fuse the multi features, the multi-kernel learning is used [15]. The work presented in [16], attempts to tune the weights of CaffeNet using the fine-tuning approach based on VGG-VD-16 architecture. Metric learning (ML) [4] has been utilized frequently into the convolutional neural models to further increase the discrimination of deep representations. A new approach is introduced to extract features using sparse autoencoder in [17]. A large patch convolutional neural network (LPCNN) is proposed in [18], where authors replace the fully-connected layer with global average pooling layer to decrease the kernels parameters. Other approaches including, OverFeat [19], a fused feature representation between salM 3 LBP and CLM [2] and six fine-tuned ConvNets [20]. As shown in table I, it can be observe that even proposed data augmentation (affine transformation) strategy achieves high accuracy than most of comparison methods. This phenomenon may be caused by lacking enough training samples, as there are 100 training samples of each class of UC merced data set and 50 to 60 of each class of WHU-RS data set. Making a comparison with these methods, our proposed fusion achieves the best accuracy (99.17%) for WHU-RS dataset and obtain an impressive accuracy (98.10%) for UC Merced dataset. Hence, the proposed fusion framework achieves very competitive accuracy in the literature of scene classification when compared with low-level based approaches, high-level methods, and deep learning frameworks.

For further analysis, a confusion matrix of WHU-RS dataset and UC Merced dataset is shown in Fig.2 and Fig.3, respectively. From Fig.2, the Pond category in WHU-RS dataset, which is hard to be classified in all proposed methods because of inter-class similarity with port and farmland categories, achieving low accuracy. From the confusion matrix of UC Merced dataset as shown in Fig.3, it can be observe that other classes such as storage tanks, tennis court, sparse residential, medium residential, forest, river, and dense residential are easily confused due to similar structures and background color.

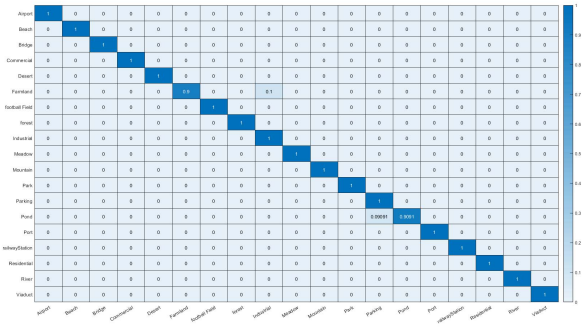

Fig. 2. Confusion matrix for the proposed model with WHU-RS data set.

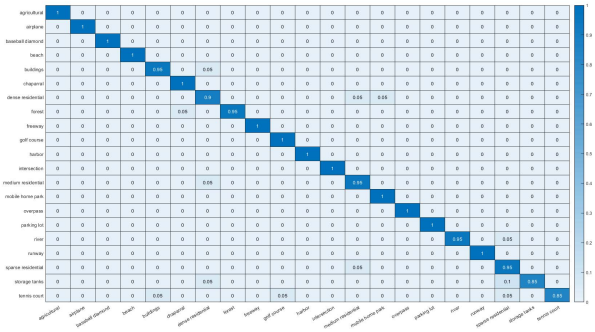

Fig. 3. Confusion matrix for the proposed model with UC-Merced data set

In summary, these data sets are challenging, even though we have achieved a comparable performance.

IV. Conclusion

This paper proposes an improved pre-trained CNN architecture named the pre-trained AlexNet-DS-CCA to classify the remote sensing scene images. It could be observed that the proposed data augmentation strategy leads to very encouraging classification results and competes most of the existing results on the same data sets. However, our approach does not neglect the importance of manually collecting more training data. To provide rich semantic information, and to prevent from overfitting, we integrate direct supervision by incorporating a deep supervision strategy (DS) into the intermediate layers, rather than the standard approach that provides supervision only at the output layer. The incorporation of DS strategy can, to some extent, alleviate over-fitting problem. Furthermore, a CCA method was adopted to fuse the two fully-connected layers features, which allows an improved classification performance than previous fusion methods. The experiments on two classical satellite data sets have demonstrated that the proposed framework boost the classification results compared with state-of-the-arts methods used now.

References

[1] Cheng, Gong, Junwei Han, and Xiaoqiang Lu. "Remote sensing image scene classification: benchmark and state of the art." Proceedings of the IEEE 105.10 (2017): 1865-1883.

[2] Bian, Xiaoyong, et al. "Fusing local and global features for high-resolution scene classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10.6 (2017): 2889-2901.

[3] Liu, Qingshan, et al. "Learning multiscale deep features for high-resolution satellite image scene classification." IEEE Transactions on Geoscience and Remote Sensing 56.1 (2018): 117-126.

[4] Cheng, Gong, et al. "When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs." IEEE transactions on geoscience and remote sensing 56.5 (2018): 2811-2821.

[5] Chaib, Souleyman, et al. "Deep Feature Fusion for VHR Remote Sensing Scene Classification." IEEE Trans. Geosci. Remote Sens 55.8 (2017): 4775-4784.

[6] Li, Erzhu, et al."Integrating Multilayer Features of Convolutional Neural Networks for Remote Sensing Scene Classification." IEEE Transactions on Geoscience and Remote Sensing 55.10 (2017): 5653-5665.

[7] Sun, Quan-Sen, et al. "A new method of feature fusion and its application in image recognition." Pattern Recognition 38.12 (2005): 2437-2448.

[8] Lee, Chen-Yu, et al. "Deeply-supervised nets." Artificial Intelligence and Statistics. 2015: (pp. 562-570).

[9] UC MERCED DATASET, http://vision.ucmerced.edu/datasets/landuse.html. Last accessed 20 February 2018

[10] SIRI-WHU Dataset, http://www.lmars.whu.edu.cn/xia/AID-project.html. Last accessed 20 February 2018

[11] Caffe Model Zoo , https://github.com/BVLC/caffe/wiki/Model-Zoo. Last accessed 20 February 2018

[12] Yan, Li, et al. "Improved Class-Specific Codebook with Two-Step Classification for Scene-Level Classification of High Resolution Remote Sensing Images." Remote Sensing 9.3 (2017): 223.

[13] Qi, Kunlun, et al. "Multi-Task Joint Sparse and Low-Rank Representation for the Scene Classification of High-Resolution Remote Sensing Image." Remote Sensing 9.1 (2016): 10.

[14] Chen, Chen, et al. "Land-use scene classification using multi-scale completed local binary patterns." Signal, image and video processing 10.4 (2016): 745-752.

[15] Han, Xiaobing, et al. "Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification." Remote Sensing 9.8 (2017): 848.

[16] Xia, Gui-Song, et al. "AID: A benchmark data set for performance evaluation of aerial scene classification." IEEE Transactions on Geoscience and Remote Sensing 55.7 (2017): 3965-3981.

[17] Li, Erzhu, et al. "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10.3 (2017): 1068-1081.

[18] Zhong, Yanfei, Feng Fei, and Liangpei Zhang. "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery." Journal of Applied Remote Sensing 10.2 (2016): 025006.

[19] Penatti, Otvio AB, Keiller Nogueira, and Jefersson A. dos Santos. "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?." Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on. IEEE, 2015.

[20] Nogueira, Keiller, Otvio AB Penatti, and Jefersson A. dos Santos. "Towards better exploiting convolutional neural networks for remote sensing scene classification." Pattern Recognition 61 (2017): 539-556.

[21] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[22] Lin, Daoyu, et al. "MARTA GANs: Unsupervised Representation Learning for Remote Sensing Image Classification." IEEE Geoscience and Remote Sensing Letters 14.11 (2017): 2092-2096.

[23] Zhao, Bei, et al. "The Fisher kernel coding framework for high spatial resolution scene classification." Remote Sensing 8.2 (2016): 157.