Recurrent Transformer Networks for Remote Sensing Scene Categorisation

Zan Chen[1]
zanchen2@gmail.com
Shidong Wang[2]
shidong.wang@uea.ac.uk
Xingsong Hou[1]
houxs@xjtu.edu.cn
Ling Shao[3]
ling.shao@ieee.org

[1] School of Electronic and Information Engineering
Xi'an Jiaotong University
Xi'an, China

[2] School of Computing Sciences
University of East Anglia
Norwich, UK

[3] Inception Institute of Artificial Intelligence
Abu Dhabi, United Arab Emirates

Abstract

Remote sensing scene categorisation is a task to distinguish the basic level scene images in accordance with the contents of the subordinate level feature representations. This gives rise to a significant semantic gap between subordinate level features and the basic level scene contents. In this paper, we propose recurrent transformer networks (RTN) to mitigate the above problem. RTN incorporates learning transformation-invariant regions with transformer based attention mechanism, thus reducing the semantic gap efficiently. It also can learn the canonical appearance for the most relevant regions based on the subordinate level contents of the remote sensing scene images. The predictions of both transformation parameters and classification score are derived from the bilinear CNN pooling regression. The whole network is differentiable and can be learned end-by-end by only acquiring the basic level labels. Through extensive experiments, we demonstrate that our RTN is able to achieve state-of-the-art performance on several public remote sensing scene datasets.

1 Introduction

During the last decades, the rapid development of remote sensing observation technologies has made it easier to accumulate gigabytes of high spatial resolution image data on a daily basis. Such kind of remote sensing (RS) images are associated with a wide range of applications such as land use and land cover (LULC) determinations, urban planning, environmental monitoring, vegetation mapping and natural hazards detection [1, 2, 3]. Hence, the remote sensing scene categorisation (RSSC) is critically important for human to smart interpret and understand the contents of RS images.

RSSC referred as categorising the complex arrangement scene images into a set of semantic classes, which is a fairly challenging task due to the following problems. First, RSSC

(a) (b)

Figure 1: Examples from NWPU-RESISC45 [6].(a) within-class diversity: palace (1st row), church (2nd row), and railway station (3rd row). (b) between-class similarity: railway station vs. stadium vs. church; airport vs. railway vs. free way; Dense residential vs. commercial area vs. industrial area; meadow vs. forest vs. wetland (from top to bottom and from left to right).

aims to abstract the basic-level semantic information of the RS images from subordinate-level feature representations, which will produce a semantic gap between the scene labels and the image contents [4]. Second, unlike natural scene images, RS image datasets have the characteristics of rich image variations, i.e. large within-class diversity and high between-class similarity as shown in Fig. 1, which makes the general methods insufficient to challenge the RSSC tasks [5]. Third, it is even more challenging to precisely predict labels of a large number of testing data by utilising a small number of training samples [6].

Many methods have been explored to handle the above issues. In the early stage, most methods rely on specific hand-crafted features for RSSC, but these methods lack the ability to bridge the semantic gap between the represent features and the ground truth [7]. A recent trend in RSSC is to take advantages of the convolutional neural networks (CNN) to learn discriminative feature representation, which has demonstrated superior performances compared to the hand-crafted methods [8]. However, CNN-based methods usually neglect the impacts on the classification which are yielded by the aforementioned semantic gaps and variances problems [9].

To cope the above issues, we design recurrent transformer networks (RTN) for RSSC tasks inspired by the spatial transformer networks (STN) [10]. In contrast with the original STN, our RTN can gradually learn to attend on multiple discriminative regions by leveraging the inter-scale loss between each two STN streams. We briefly summarise our contributions as follows: (1) Our RTN can progressively localise discriminative regions and learn robust transformation-invariant features, which reduces the semantic gaps between the basic level categories and the subordinate level categories. (2) Our RTN guarantees to retain the information by using bilinear pooling. Meanwhile, it can progressively learn the subtle differences for small regions of the images by introducing the inter-scale loss. (3) We conduct extensive experiments and achieve state-of-the-art accuracy on three challenging RS image datasets. The RTN can be efficiently trained by end-to-end with only requiring the labels of the basic level categories.

2 Related Work

In recent years, RSSC task has been explored extensively. The core of existing methods is learning robust feature representations. In this section, we will introduce the milestone works from two aspects which are extracting hand-crafted features and CNN-based features.

Hand-crafted Feature Approaches. In the early stage, remote sensing images were in low resolution, many approaches referred to utilise hand-crafted features for RSSC tasks. The handcrafted features incorporate colour, texture and shape information to learn task-specific descriptors. For example, [11] learns the texture information by applying the completed local binary pattern (LBP) features. [12] further fuse LBP features with the different local features such as SIFT and fisher vectors. Considering handcrafted features are insufficient to compose the rich semantic information of the remote sensing images, many works [13, 14, 15] have been proposed to employ clustering or to encode multi-local features to obtain more discriminative features. Applying the bag of visual words (BOVW) model [2, 16] to generate the semantic features is also popular in RSSC domains.

CNN-based Feature Approaches. Recently, deep learning based architectures have been intensively exploited to improve the performance of RSSC tasks and achieved impressive results. For example, [17] introduces a two-stage framework to extract deep features from the pre-trained model, while using a supervised CNN classifier to predict final classification score. [18] proposes a stacked sparse autoencoder model to learn the feature representation for the land-use classification task. In addition, [8] provides an analysis of how to better applying CNN for RSSC tasks and demonstrate their best results produced by using the combination of the fine-tuning and the linear support vector machine (SVM). As CNN features are limited by lacking the ability to learn invariance of the input data, [9] proposes plugging a new rotation-invariant layer into the CNN. Besides, [19] investigates the profits of the data augmentation methods, such as flip, translation, and rotation. One notable work has been proposed by [5], which imposes metric learning regularisation on the CNN features and has achieved state-of-the-art results on popular RSSC datasets.

In contrast with aforementioned methods, our framework can efficiently handle the variances of the input data by exploiting STN [10] without the data augmentation process, and meanwhile, learn the discriminative features by utilising bilinear pooling methods [20].

3 Approach

In this section, we will introduce the proposed recurrent transform networks (RTN) for remote sensing scene categorisation. The core of our RTN is to recursively discover the transformation-invariant regions and learn the latent relationship based region feature representations. As shown in Figure 2, our RTN framework composes of three major parts which are recurrent warp operation, bilinear pooling operation, and intra-scale loss L_{intra} and inter-scale loss L_{inter}. The proposed RTN architecture can automatically discover multi-scale transformed regions and learn their canonical appearances. With employing bilinear pooling function and inter-scale loss function, RTN is efficient to handle the variances of the input data in a mutually reinforced manner and achieve competitive results on the public RSSC datasets.

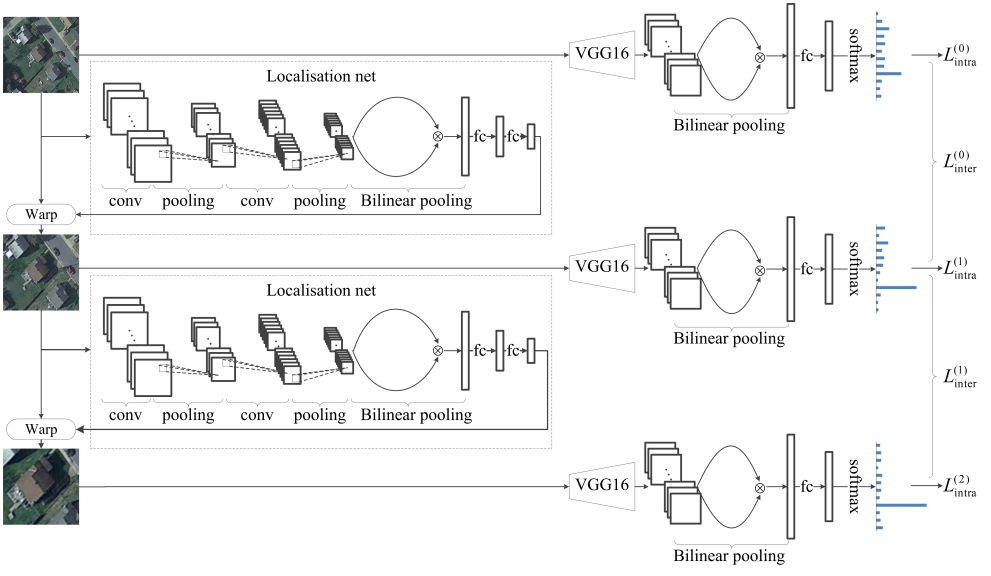


Figure 2: The architecture of recurrent transformer networks (RTN). Given an input image, the localisation network will learn to predict the transformer parameters. With recurrently applying the warp operation, the network can progressively attend to the discriminative regions and produce multi-scale relevant sub-images (*i.e.*, three streams including raw image). The classification loss L_{intra} is designed to evaluate the results for each stream, while the inter-scale loss L_{inter} is employed to discover the relationship of the neighbouring streams. All of the regression layers are based on bilinear pooling (*i.e.*, $\otimes$). '*conv*', '*pooling*' and 'fc' denote the convolutional layer, max pooling layer, and fully-connected layer respectively.

3.1 Recurrent Warp Operation

The recurrent warp operation is to handle CNN variance problems by learning multi-scale discriminative regions inspired by spatial transformer networks [10]. The original STN includes multiple independent streams, and each stream learns its own spatial transformation independently, which neglects the latent relationship of each stream.

To address these disadvantages, we propose recurrent warp operation which extracts the relevant multi-scale region-based feature representations progressively. Specifically, our warp operation runs in a recurrent manner, which can be denoted as

$$I^{(s)} = f_{warp}(\theta^{(s)}\tau^{(s)}, I^{(s-1)}), \tag{1}$$

where $I^{(s)}$ is the s-th scale image (e.g., $I^{(0)}$ is the raw image), θ^s is the transformation parameters computed by the localisation function $\theta^s = f^{(s)}_{loc}(I^{(s-1)})$, and $\tau^{(s)}$ is the target coordinates of the regular grid in the output image. Each warp operation f_{warp} has the similar progress to the original STN [10]. Suppose the i-th target point of the output image as

$\tau_i^{(s)} = [x_i^{(s)}, y_i^{(s)}, 1]^T$, the corresponding source coordinates are generated by

$$\begin{bmatrix} x_i^{(s-1)} \\ y_i^{(s-1)} \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_{1,1}^s & \theta_{1,2}^s & \theta_{1,3}^s \\ \theta_{2,1}^s & \theta_{2,2}^s & \theta_{2,3}^s \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i^{(s)} \\ y_i^{(s)} \\ 1 \end{bmatrix}. \tag{2}$$

Then, it allows doing the warp operations, such as crop, scale, and translation. The warp operation requires applying a sampling kernel on the input image $I^{(s-1)}$, to produce the value at a particular pixel in the finer scale image $I^{(s)}$. In this work, we obtain finer scale image by employing the standard bilinear interpolation, and the i-th target point value of $I^{(s)}$ can be written as

$$I_i^{(s)} = \sum_{\tilde{h}=1}^{H} \sum_{\tilde{w}=1}^{W} I_{\tilde{h}\tilde{w}}^{(s-1)} \max(0, 1 - |x_i^{(s-1)} - \tilde{w}|) \max(0, 1 - |y_i^{(s-1)} - \tilde{h}|)), \tag{3}$$

where H and W denote the height and width of the input image $I^{(s-1)}$. With repeatedly calling the warp operation, the network can progressively yield multi-scale discriminative regions.

3.2 Intra-scale Loss and Inter-scale Loss

As the recurrent warp operation produces the relevant regions from coarser scales to finer scales, we extract feature representation for each stream by the pre-trained deep architecture (e.g., VGG16 [21] in this work). To maintain the selectivity of the spatial information, we employ the bilinear pooling method [20] to generate the final classification scores. The standard bilinear pooling can be written as

$$B(\mathcal{X}) = \sum_{i=1}^{hw} x_i x_i^T, \tag{4}$$

where $B(\mathcal{X}) \in \mathbb{R}^{c \times c}$ denotes the bilinear pooled feature, and $x_i \in \mathbb{R}^c$ denotes the feature vector at the i-th channel of the given VGG feature $\mathcal{X} \in \mathbb{R}^{h \times w \times c}$. Equation 4 captures the second order statistics of the feature map. Then, we can apply a fully-connected layer followed with a softmax layer to map the bilinear pooled feature to the probability distribution of the responding category entries.

Based on the generated probability, an alternative way is to optimise it directly. However, it lacks considering the relationship of the neighbouring scales. To cope this issue, we merge intra-scale loss for each stream and inter-scale loss for neighbouring streams to optimise the network. The final loss is defined as

$$L = \sum_{s=0}^{S} L_{intra}^{(s)} + \alpha \sum_{s=0}^{S-1} L_{inter}^{(s)}, \tag{5}$$

where α is a hyper-parameter to adjust the total loss and learn the latent relationship between the neighbouring scales. Suppose $P^{(s)}$ and P^* as the predicted label vector from a specific scale and the ground truth label respectively, then the intra-scale loss $L_{intra}^{(s)}$ can be written as

$$L_{intra}^{(s)} = -\sum_{k=1}^{n} P_k^* \log P_k^{(s)}, \tag{6}$$

where n is the number of the classification. To ensure the streams learning in a mutual reinforcement way, we impose inter-scale loss for the adjoining scales and define it as

$$\begin{aligned} L_{inter}^{(s)} &= \max(0, \sum_{k=1}^{n} P_k^* (\log P_k^{(s)} - \log P_k^{(s+1)}) - \text{margin}) \\ &= \max(0, L_{intra}^{(s+1)} - L_{intra}^{(s)} - \text{margin}), \end{aligned} \tag{7}$$

which enforces $L_{intra}^{(s+1)} < L_{intra}^{(s)} + \text{margin}$ during the training phase. In such way, each scale can refer to the adjoining scales to progressively learn sub-region feature representations. With gradually attending at the finer scale, the extracted features are able to decrease the semantic gap by degrees and boost the performance of the proposed architecture on the RSSC datasets.

3.3 Backpropagation of the Model

The framework of our RTN has been demonstrated as above, we will illustrate the backpropagation process of the RTN. The warp operation and the bilinear pooling have been proven by [10] and [20], which is differentiable within the CNN. We will provide the update of a specific network parameter in the scenarios of merging inter-scale and intra-scale loss functions. Without loss of generality, we consider a convolutional weight $\bar{w}$ in the VGG-based feature extraction part at scale s. Its update can be calculated by the stochastic gradient descent (SGD)

$$\begin{aligned} \bar{w} =& \bar{w} - \frac{\eta}{m} \sum_{i=1}^{m} \frac{\partial L_i}{\partial \bar{w}} \\ =& \bar{w} - \frac{\eta}{m} \sum_{i=1}^{m} \frac{\partial (L_{intra,i}^{(s)} + \alpha L_{inter,i}^{(s-1)} + \alpha L_{inter,i}^{(s)})}{\partial \bar{w}} \\ =& \bar{w} - \frac{\eta}{m} \sum_{i=1}^{m} ((1 + \alpha \delta_i^{(s-1)} - \alpha \delta_i^{(s)}) \frac{L_{intra,i}^{(s)}}{\partial \bar{w}}), \end{aligned} \tag{8}$$

where η denotes the learning rate, L_i refers to the value of the loss function at the i-th training example, m is the batch size, and α is introduced in Equation 5. δ is associated with Equation 7 to denote the options of the returned value, which is defined as

$$\delta_i^{(s-1)} = \begin{cases} 1, & \text{if } L_{intra,i}^{(s-1)} < L_{intra,i}^{(s)} - \text{margin} \\ 0, & \text{otherwise} \end{cases}. \tag{9}$$

More specifically, the value of the δ refers to the degree of relevance of the adjacent scales. For instance, if the intra-scale loss of the $I^{(s)}$ is significantly higher than $I^{(s-1)}$, it acquires increasing the learning rate of weights by α to reduce the distances between $I^{(s)}$ and $I^{(s-1)}$, and vice versa.

4 Experiments

4.1 Datasets

We conduct experiments on three publicly available remote sensing image datasets, including NWPU-RESISC45[6], UC Merced[22], and AID[23]. We provide a summary statistics and

show them in Table 1. For the NWPU-RESISC45 dataset, we set the ratios of the training set to 10% and 20%, and the left 90% and 80% for testing. For AID and UC Merced datasets, the training ratios are set to 20% and 80% respectively, and the rest of the data for testing. More details of the split information can be found in Table 2.

Table 1: The statistics of the datasets used for experiments.

Datasets	Images per class	Scene Class	Total Images	Image size
NWPU-RESISC45 [6]	700	45	31500	256×256
AID [23]	220∼420	30	10000	600×600
UC Merced [22]	100	21	2100	256×256

4.2 Implementation Details

We evaluate our framework with employing VGG16 [21] architecture, which has been pre-trained on the ImageNet. '*conv5_3*' features are extracted to predict the classification score. The localisation net is composed of two convolutional layers, each of which is followed with a max-pooling layer. Then, we compute the bilinear pooling features, with two subsequent fully-connected layers to predict the transformation parameters. All the input images are resized to 224×224 resolutions.

We set the initial learning rate as 0.0001 and 0.01 for the localisation net and classification net with weight decay rate 0.0005. The input batch size is 36. To keep the model training stable, we empirically set the α and margin in Equation 5 and Equation 7 as 0.1 and 0.05. The model is roughly trained for $80k$ iterations with standard SGD.

4.3 Experimental Results and Comparisons

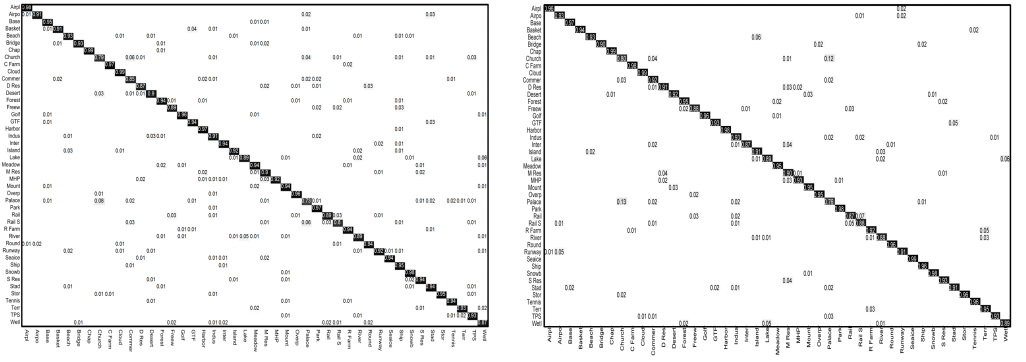
Existing methods for RSSC tasks can be partitioned into two categories, which are hand-crafted feature approaches and deep CNN-based feature approaches. We show the classification results produced by some representative methods. Note that the relevant results are borrowed from the original papers. As shown in Table 2, we can obvious that our method outperforms the hand-crafted feature approaches by a significant margin. CNN-based feature approaches have a much better performance on predicting the categories for RSSC tasks compared with the hand-craft feature approaches. From Table 2, we can obvious that both GoogLeNet [24] and VGG16 [21] are able to obtain the acceptable results on three experiment datasets. On UC Merced dataset [22], VGG16 with the support vector machine (SVM) gets 97.14% accuracy, which is only 1.82% lower than our RTN framework. Such disparities become larger and larger, along with the growing size of the datasets and the number of training samples.

The best accuracies on RSSC datasets are made by the recently proposed D-CNN method, which takes the metrics learning as the regularisation term [5]. Compared to the state-of-the-art results of D-CNN, our RTN achieves improvements to the categorisation accuracies on all experiment datasets. Specifically, we obtain 92.44% on AID database, with 1.62% relative gain. Moreover, we provide the confusion matrices of our RTN and D-CNN in the same settings, to demonstrate the detailed differences between two methods. As shown in Fig. 3, our RTN achieves 76% classification accuracy on the most challenging category-*Palace*, with 3% gain compared to D-CNN. Additionally, 32 out of 45 classes obtain better or the similar

result for per-class accuracy. This leads to our RTN attain a new state-of-the-art performance for RSSC tasks.

Table 2: Comparison results of our RTN to baselines and previous work. We experiment the accuracy in different training ratios on three public RSSC datasets.

	Method	NWPU-RESISC45		AID	UC-Merced
		10%	20%	20%	80%
Handcraft Feature	SPM+SIFT [25]	27.83	32.96	38.43	60.02
	LLC+SIFT [26]	38.81	40.03	58.06	72.55
	BoVW+SIFT [27]	41.72	44.97	62.49	75.52
Deep Feature	GoogLeNet+SVM [24]	82.57	86.02	87.51	96.82
	VGG16+SVM [21]	87.15	90.36	89.33	97.14
	D-CNN with GoogLeNet [5]	86.89	90.49	88.79	97.07
	D-CNN with VGG16 [5]	89.22	91.89	90.82	98.93
	RTN with VGG16 (ours)	**89.90**	**92.71**	**92.44**	**98.96**



(D-CNN method) (Our RTN)

Figure 3: Confusion matrices of D-CNN [5] and our RTN on the NWPU-RESISC45 dataset (20% ratio for training).

Table 3: The classification results of our RTN framework at the different scales on NWPU-RESISC45 dataset (20% ratio for training).

Scales.	scale 0	scale 1	scale 2	scale (0+1)	scale (0+1+2) w/o L_{inter}	scale (0+1+2) w/ L_{inter}
Acc.	91.20%	91.84%	90.20%	92.35%	92.49%	**92.71%**

4.4 Qualitative Analysis and Visualisations

We show the accuracy of the attended regions from multiple scales of our RTN framework for qualitative analysis. All experiments are evaluated at the same settings with the different scales. As shown in Table 3, scale 1 presents the higher accuracy than both scale 2 and scale 1 (e.g., the raw image). This reflects the finer scale is discriminative to corresponding categories, but it should not zoom in without any limitation. The finer scale learns the subtle differences of the images but meanwhile appears incorporating less information. An efficient

Figure 4: Visualisations of test images of our method on NWPU-RESISC45 dataset. The first row are the raw images and the rest rows are respect to the finer scale images.

way is to stack multiple scales to generate the classification results (e.g., scale (0+1) achieves 92.35% accuracy). With imposing the inter-scale loss L_{inter}, our RTN obtain the best result (92.71%), even surpassing scale (0+1+2) without using L_{inter}. It confirms the truth of L_{inter} can strengthen the relevancy of abutting scales during learning.

To better understand our RTN model, we visualised the model responses for specific input images. We have randomly picked six groups of test images and have shown the raw images and the attended finer scale regions in Figure 4. The visualisations suggest that our RTN is capable of removing cluttered backgrounds and gradually focusing on specific discriminative parts. Apart from learning the relevant regions, our RTN also enables to discover the canonical appearances of the finer regions and improves the final classification results.

5 Conclusions

In this paper, we have presented novel recurrent transformer networks (RTN) for remote sensing scene categorisation. The model learns invariance of the input data, which is achieved by progressively focusing multiple discriminative regions from coarser scales to finer scales and leveraging the extra inter-scale loss for the neighbouring regions to enforce the sub-regions learning in mutually reinforcement way. The proposed framework can be trained end-by-end and achieves state-of-the-art accuracy on the public RSSC datasets.

Acknowledgement

Xingsong Hou is the corresponding author. This work was supported in part by the NSFC under Grant 61373113, u1531141, 61732008 and 61772407, the National Key R&D Program of China under Grant 2017YFF0107700, Guangdong Provincial Science and Technology Plan Project under Grant 2017A010101006 and 2016A010101005.

References

[1] Gong Cheng, Junwei Han, Lei Guo, Zhenbao Liu, Shuhui Bu, and Jinchang Ren. 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, 2015.

[2] Gong Cheng, Lei Guo, Tianyun Zhao, Junwei Han, Huihui Li, and Jun Fang. 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, 2013.

[3] Qiong Hu, Wenbin Wu, Tian Xia, Qiangyi Yu, Peng Yang, Zhengguo Li, and Qian Song. Exploring the use of google earth imagery and object-based methods in land use/cover mapping. *Remote Sensing*, 5(11):6026–6042, 2013.

[4] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience & Remote Sensing Magazine*, 4(2):22–40, 2016.

[5] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 2018.

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[7] Gong Cheng, Peicheng Zhou, Junwei Han, Lei Guo, and Jungong Han. 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, 2015.

[8] Keiller Nogueira, Otávio AB Penatti, and Jefersson A dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017.

[9] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[11] Chen Chen, Baochang Zhang, Hongjun Su, Wei Li, and Lu Wang. Land-use scene classification using multi-scale completed local binary patterns. *Signal, image and video processing*, 10(4):745–752, 2016.

[12] Xiaoyong Bian, Chen Chen, Long Tian, and Qian Du. Fusing local and global features for high-resolution scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(6):2889–2901, 2017.

[13] Anil M Cheriyadat. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):439–451, 2014.

[14] Mohamed L Mekhalfi, Farid Melgani, Yakoub Bazi, and Naif Alajlan. Land-use classification with compressive sensing multifeature fusion. *IEEE Geoscience and Remote Sensing Letters*, 12(10):2155–2159, 2015.

[15] Xiaoqiang Lu, Yuan Yuan, and Xiangtao Zheng. Joint dictionary learning for multispectral change detection. *IEEE transactions on cybernetics*, 47(4):884–897, 2017.

[16] Qiqi Zhu, Yanfei Zhong, Bei Zhao, Gui-Song Xia, and Liangpei Zhang. 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, 2016.

[17] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2016.

[18] Xiwen Yao, Junwei Han, Gong Cheng, Xueming Qian, and Lei Guo. Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3660–3671, 2016.

[19] Xingrui Yu, Xiaomin Wu, Chunbo Luo, and Peng Ren. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, 54(5):741–758, 2017.

[20] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Sigspatial International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.

[23] Gui Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–17, 2016.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions.

[25] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2010.

[26] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. 119(5):3360–3367, 2010.

[27] Qiqi Zhu, Yanfei Zhong, Bei Zhao, Gui Song Xia, and Liangpei Zhang. 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, 2017.