2023 International Conference on New Trends in Computational Intelligence (NTCI 2023)

Boosting Few-shot Remote Sensing Image Scene Classification with Language-guided Multimodal Prompt Tuning

Haixia Bi*, Zhangwei Gao*, Kang Liu*, Qian Song†, Xiaotian Wang*
*National Engineering Research Center of Offshore Oil and Gas Exploration,
School of Information and Communications Engineering,
Xi'an Jiaotong University, China
†Technical University of Munich, Germany

2023 International Conference on New Trends in Computational Intelligence (NTCI) | 979-8-3503-8086-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/NTCI60157.2023.10403750

***Abstract*—Remote sensing image Scene classification is an important research topic in remote sensing community and has evoked a growing concern with the recent development of deep learning techniques. However, the requirement of a large amount of annotations brings great challenges to deep learning-based scene classification approaches. Visual-linguistic pretraining models, which improve the transferability of visual models using the supervision information of text, create a new way for the task under label scarcity scenario. In this paper, we explore the novel approach of prompt engineering, aiming to achieve satisfactory performance of multi-modal pretraining models on downstream remote sensing image scene classification task with minimal amounts of training data. Experiments were conducted on multiple publicly available datasets. The results indicate that training the learnable prompts with a small number of samples can yield impressive results, surpassing the few-shot transfer learning results of the best-performing pre-trained models.**

***Index Terms*—Remote sensing image scene classification, Prompt tuning, Few-shot learning, Multi-modal pretraining.**

I. Introduction

Remote sensing image scene classification, which aims to classify scene images into different semantic categories, is a crucial component in remote sensing image processing realm. It has been widely applied in various fields, such as land resource management and environmental monitoring [1]. The early remote sensing image scene classification methods utilize manually crafted features, such as spectral features and texture features, and then performs classification with classifiers [2]. These handcrafted features-based methods require the involvement of numerous efforts of human experts, and their generalization ability is weak [3], [4].

Deep learning methods have become one of the research focuses in the field of remote sensing image scene classification due to their powerful hierarchical feature extraction capabilities [5]. Literature [6] utilizes multi-layer convolutional layers and dense residual blocks to obtain more detailed features for remote sensing images. [7] combines graph neural network (CNN) and graph neural network (GNN) to make full use of the adjacency and disjointness relationship among geographical objects. However, the training of these models requires a large amount of annotations, while annotating is a cumbersome and costly task [8], [9], [10]. Therefore, improving the scene classification performance with a meager amount of labels is still a challenging task in remote sensing field.

Most recently, pretrained models based on multimodal data have shown tremendous potential in representation learning. The deep features learned by such models exhibit strong generalization abilities and demonstrate great versatility across various downstream tasks. The most outstanding multimodal pretrained models include CLIP [11], ALBEF [12] and VLMo [13]. It should be noted that these models successfully promoted the performances of vision tasks by exploring the heterogeneity and correlation between lanaguage and vision. For instance, the CLIP pretrained model [11] achieves comparable accuracy to ResNet-50 [14] on the ImageNet dataset, even without using its 1.28 million training samples. Some advanced CLIP models even achieved zero-shot inference accuracy of over 80.0% on the ImageNet-1K dataset.

Motivated by the strong generalization performance of the vision-language multimodal pretrained models, we propose the language-guided remote sensing image scene classification approach in this paper. Specifically, to overcome the poor performance of handcrafted hard prompts when encountering domain-specific tasks or data, we explore the learnable prompt tuning-based transfer learning paradigm for remote sensing image scene classification task. Experimental results on two benchmark datasets show that learnable prompts can significantly enhance model performance, even outperforming specialized remote sensing pretrained models in few-shot learning scenario.

The rest of this paper is organized as follows. Section II introduces the the proposed method. Section III presents the experimental results. Section IV concludes the paper.

II. Methodology

In this section, we will first introduce the utilized vision-language multimodal model CLIP, and then explain the prompt tuning-based transfer learning method.

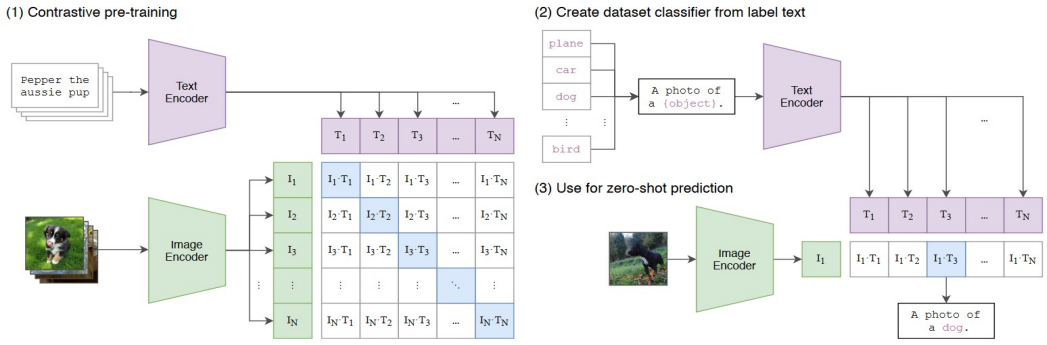

Fig. 1. The architecture of CLIP.

A. CLIP Multimodal pretraining model

CLIP is a groundbreaking multimodal pretrained model that provides a new approach by aligning text and images in a shared feature space, with separate encoders for images and text. The text encoder is based on BERT [15], while the visual encoder can utilize ResNet [14] or ViT [16]. Through large-scale pretraining, the model can learn diverse visual concepts and easily transfer to any downstream task using prompts. CLIP represents a successful practice of leveraging large-scale multimodal pretraining for zero-shot learning.

CLIP achieves remarkable zero-shot task transfer capability by only employing contrastive pretraining on a large-scale dataset of image-text pairs. As CLIP utilizes multimodal pretraining to transform classification into a retrieval task, specifically predicting whether an image matches a given text description, it naturally lends itself to zero-shot recognition. This is accomplished by comparing the synthesized classification weights of image features with the text encoder, which takes class-specific text descriptions as input.

The CLIP architecture is illustrated in Figure 1. For a given image *I*, it undergoes segment embedding to obtain $E_0 \in \mathbb{R}^{(M\times d_v)}$. The image encoder V consists of K Transformer layers $V_{i(i=1)}^{K}$. We introduce a special learnable token as the classification token c_i. For the i-th layer, we have:

$$[c_i, E_i] = V([c_{i-1}, E_{i-1}]), \quad i = 1, ..., K \tag{1}$$

To obtain the final image representation, the $ImageProj(\cdot)$ function projects the class token c_i into the latent embedding space of visual-textual domain:

$$x = ImageProj(c_K), \quad x \in \mathbb{R}^d, \tag{2}$$

where d denotes the dimension of the embedding space.

Similarly, for the text encoder L, with an input of a word embedding sequence $W_0 = [w_0^1, w_0^2, ..., w_0^N] \in \mathbb{R}^{N\times d_l}$, for the i-th layer:

$$W_i = L_i(W_{i-1}), \quad i = 1, 2, \ldots, K \tag{3}$$

The final text representation is obtained by projecting the embedding of the last token from the last layer using $TextProj(\cdot)$ into the latent embedding space of visual-textual representations:

$$z = TextProj(w_K^N), \quad z \in \mathbb{R}^d \tag{4}$$

So the image I extracts image features as x, and the text encoder generates a set of feature vectors $\{z_i\}_{i=1}^{C}$, where C represents the number of classes, and each z_i is derived from prompts of the form "a photo of [class]", with class labels replaced by specific class names like "cat", "dog", or "car". Then, the predicted probabilities are computed as:

$$p(y = i|I) = \frac{\exp(\frac{\cos(z_i, x)}{\tau})}{\sum_{j=1}^{C} \exp(\frac{\cos(z_j, x)}{\tau})}, \tag{5}$$

where τ is a parameter learned during the training process of CLIP, and $cos(\cdot)$ represents cosine similarity.

Compared to traditional classifier learning methods that learn closed-set visual concepts from random vectors, visual-language pretraining allows exploration of open-set visual concepts through high-capacity text encoders, enabling a broader semantic space and facilitating transferability of learned representations to downstream tasks. Moreover, CLIP possesses powerful zero-shot transfer capabilities, endowing it with open-ended detection abilities.

B. Learnable Prompt Tuning

In this section, we will introduce the employed prompt tuning method called MaPle, as illustrated in Figure 2. Compared to other prompt tuning approaches, MaPle incorporates learnable prompts in both modal branches, allowing both modalities to adapt to downstream tasks. The learnable prompts are based on the pretrained visual-textual model CLIP, with the visual Transformer serving as the visual encoder.

1) Deep Language Prompts:: To learn language prompts, we introduce b learnable tokens $\{P^i \in \mathbb{R}^{d_l}\}_{i=1}^{b}$ to the CLIP text branch. The input embeddings are then $[P^1, P^2, ..., P^b, W_0]$, where $W_0 = [w^1, w^2, ..., w^N]$ represents fixed token embeddings of the input text. We incorporate such

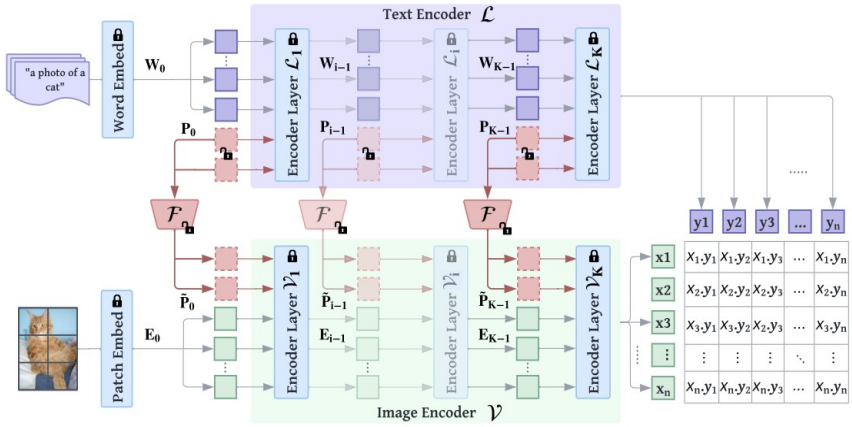

Fig. 2. The architecture of MaPle.

learnable tokens at deep levels as well. Specifically, for the i-th layer:

$$[\_, W_i] = L_i([P_{i-1}, W_{i-1}]), \quad i = 1, 2, \ldots, J, \tag{6}$$

where $[\cdot, \cdot]$ denotes the concatenation operation, and J represents the layer at which we need to insert the prompts. After the J-th layer, we handle the prompts in the following manner:

$$[P_i, W_i] = L_i([P_{i-1}, W_{i-1}]), \quad i = J+1, \ldots, K, \tag{7}$$

where K represents the total number of layers in that branch. The final text representation is obtained by projecting the embedding of the last token from the final layer using $TextProj(\cdot)$ into the visual-textual latent embedding space:

$$z = TextProj(w_K^N), \quad z \in \mathbb{R}^d, \tag{8}$$

where d represents the dimension of the embedding space.

2) Deep Visual Prompts:: Similar to the language branch, we introduce b learnable prompt tokens $\{\tilde{P}^i \in \mathbb{R}^{d_v}\}_{i=1}^{b}$ to the visual branch. For the i-th layer of the visual encoder:

$$[c_i, E_i, \_] = V([c_{i-1}, E_{i-1}, \tilde{P}_{i-1}]), \quad i = 1, 2, \ldots, J \tag{9}$$

$$[c_i, E_i, \tilde{P}_i] = V([c_{i-1}, E_{i-1}, \tilde{P}_{i-1}]), \quad i = J+1, \ldots, K \tag{10}$$

To obtain the final image representation, we utilize the function $Imageproj(\cdot)$ as well to project the category token c_i into the latent embedding space:

$$x = ImageProj(c_K), \quad x \in \mathbb{R}^d \tag{11}$$

Compared to independent prompts, cross-stage shared prompts are more effective. Therefore, unlike the early stages, late stages do not provide free prompts for independent learning.

3) Coupling Visual and Language Prompts:: If there is no interaction during the learning and adjustment process of these prompts, the two modal branches naturally lack synergy. In order to couple the prompts from both modalities, we introduce the language prompt token from the J-th layer into the corresponding visual branch using a linear layer as the coupling function:

$$[c_i, E_i, \_] = V([c_{i-1}, E_{i-1}, F_{i-1}(\tilde{P}_{i-1})]), \quad i = 1, 2, \ldots, J, \tag{12}$$

$$[c_i, E_i, \tilde{P}_i] = V([c_{i-1}, E_{i-1}, \tilde{P}_{i-1}]), \quad i = J+1, \ldots, K, \tag{13}$$

where $F(\cdot)$ represents the coupling function.

III. Experiments

We conducted a series of experiments using two publicly available remote sensing scene classification datasets: the AID dataset [17] and the UCM dataset [18], to evaluate the performance of prompt tuning in few-shot transfer learning scenario. We first introduce the experimental datasets and deployment details used in this study, then validate the effectiveness of our proposed method on these two datasets. Additionally, we compare our approach with the state-of-the-art algorithm, namely the ViT model [16] pre-trained on large-scale remote sensing datasets, on both the AID and UCM datasets to highlight the strengths and weaknesses of both methods.

A. Datasets and Experimental Settings

AID Dataset: It is a large-scale dataset used for aerial image classification in remote sensing scenarios. It consists of 10,000 sample images of 30 different types of aerial scenes. Each category contains 200 to 400 samples. The images have a resolution of 600×600 pixels.

UCM dataset: It consists of 2,100 images in total. It includes 21 different land-use categories, with 100 samples per category. The images have a resolution of 256×256 pixels. These

TABLE I
ACCURACY OF ZERO-SHOT AND FEW-SHOT TRANSFER ON AID DATASET (%)

Class	Zero-shot	8-shot	16-shot	Class	Zero-shot	8-shot	16-shot
Airport	99.4	100	100	Bare land	0.0	80.0	87.1
Baseball field	99.1	90.9	93.6	Beach	98.0	98.5	98.5
Bridge	93.3	96.7	96.7	Center	11.8	80.0	80.0
Church	65.8	85.0	84.2	Commercial	12.6	83.4	80.0
Dense residential	74.2	87.3	90.2	Desert	93.3	96.7	97.3
Farmland	89.2	94.1	97.8	Forest	95.2	99.2	98.4
Industrial	75.4	74.4	82.6	Meadow	9.3	93.6	92.9
Medium residential	40.0	90.3	92.4	Mountain	82.9	95.3	99.4
Park	67.4	85.7	94.9	Parking	100.0	96.4	98.5
Playground	3.2	71.4	92.4	Pond	95.2	98.1	98.1
Port	89.5	93.7	93.7	Railway station	87.9	92.3	86.2
Resort	66.9	77.2	84.1	River	79.5	96.6	99.5
School	50.0	74.7	87.3	Sparse residential	0.0	95.3	86.0
Square	20.6	87.3	79.4	Stadium	97.4	95.9	93.1
Storage tanks	0.0	83.9	93.9	Viaduct	80.0	99.0	95.7
Overall	63.7	89.7	92.2				

images were manually extracted from large-scale images of urban areas in various cities across the United States. The pixel resolution of these publicly available images is 1 foot.

Evaluation Metrics: We use accuracy as the metric to measure the performance of our algorithm, which is defined as:

$$Acc = \frac{\sum_{i=1}^{k} x_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} x_{ij}}, \tag{14}$$

where k represents the number of categories. In this case, the dataset used has 30 classes and 21 classes respectively. x_{ij} denotes the number of samples from class i that are classified as class j. x_{ii} represents the number of samples correctly classified in class i.

1) Results for Few-shot Transfer Learning: This section discusses the performance of multimodal pretraining models like CLIP on remote sensing scene classification datasets using fine-tuning. In the subsequent use of the CLIP model, unless otherwise specified, we will be using the OpenAI pretrained CLIP model. We utilize the data Transformer base model (ViT-base) as our visual encoder, where H=8, d_m=768, K=12 layers, and the image patch resolution is 16×16. The input image size is set to 224×224. Therefore, the visual encoder branch used in this paper has approximately 86M parameters.

The CLIP model is pretrained on a dataset of 400 million images To begin with, we employ manually crafted prompt templates to fine-tune the CLIP model on the AID dataset and UCM dataset. The prompt template used is "an aerial remote sensing photo of [class]", where "[class]" represents the specific remote sensing scene category. To evaluate the performance of the algorithm, we select 50% of the data as the test set, while the remaining 50% as training set.

We conducted 8-shot and 16-shot transfer experiments using the prompt tuning approach described in Section 2. Specifically, we created new training sets by randomly selecting 8 (or 16) samples per class from the original training set. We then trained the model with prompts following the parameter settings of MaPLe. For the 16-shot transfer training, we performed 15 epochs, while for the 8-shot transfer, we conducted 10 epochs. All experiments were deployed on a compute platform consisting of 4×NVIDIA RTX A5000 24G GPUs. The experimental results on two datasets are presented in Table I and Table II respectively.

From the results, it is evident that the CLIP model achieves impressive performance using manually crafted templates. However, there are two prominent issues with using manually designed templates for zero-shot transfer learning. *1)* during the process of manually designing templates, the performance of the CLIP model on these datasets is sensitive to the selection of manual templates. For instance, if we choose "a photo of [class]" as the prompt template, where "[class]" represents the remote sensing scene category, the test accuracy on the AID dataset is only 55.3%. *2)* looking at the specific test accuracies for different categories, it is observed that manually designed templates perform poorly on certain classes. For example, on the AID dataset, categories such as "Bare land", "Meadow", "Storage tanks", "playground" and "Sparse residential" exhibit low accuracies. Similarly, on the UCM dataset, categories like "Dense residential", "Medium residential", "parse residential" and "Storage tanks" also have almost 0 accuracy.

The main reasons for these issues can be attributed to two factors. Firstly, during the CLIP pretraining phase, there is limited exposure to remote sensing-specific image-text pairs, resulting in a lack of specific visual information related to remote sensing domains. Secondly, some categories exhibit minimal differences in text or images, making it challenging to distinguish between them. For example, on the AID dataset, categories like "Playground" and "Square" or "Commercial" and "Railway station," as well as on the UCM dataset, categories like "Mobilehome park" and "Parking lot" are easily confused due to their subtle distinctions.

2) Comparative Results: To further investigate effective transfer methods with different types of parameters, we designed an experimental comparison with existing transfer learning approaches based on the CLIP model. These include Clip-Adapter, CoOp and the MaPle method used in this paper, where the numerical results are shown in Table III.

Clip-Adapter is a transfer learning approach that adds an additional adapter module after the backbone network of the

TABLE II
ACCURACY OF ZERO-SHOT AND FEW-SHOT TRANSFER ON UCM DATASET (%)

Class	Zero-shot	8-shot	16-shot	Class	Zero-shot	8-shot	16-shot
Agricultural	62.0	94.0	92.0	Airplane	74.0	98.0	100.0
Baseball diamond	100.0	100.0	100.0	Beach	98.0	100.0	100.0
Buildings	44.0	82.0	88.0	Chaparral	14.0	100.0	100.0
Dense residential	4.0	88.0	90.0	Forest	96.0	100.0	98.0
Freeway	82.0	82.0	80.0	Golf course	96.0	100.0	100.0
Harbor	54.0	100.0	100.0	Intersection	98.0	98.0	98.0
Medium residential	4.0	68.0	80.0	Mobilehome park	100.0	84.0	96.0
Overpass	62.0	92.0	98.0	Parking lot	96.0	98.0	98.0
river	84.0	96.0	98.0	runway	46.0	96.0	98.0
Sparse residential	0.0	96.0	100.0	Storage Tanks	6.0	90.0	100.0
Tennis court	94.0	92.0	96.0	Overall	62.6	93.0	95.7

TABLE III
COMPARISON WITH DIFFERENT TRANSFER LEARNING METHODS (%)

Method	UCM dataset	AID dataset
Clip-Adapter	83.7	86.0
CoOp	94.0	91.6
MaPle	**95.7**	**92.2**

CLIP model, rather than inserting adapters into the backbone network itself, thus preserving the integrity of the structure. In addition, this method utilizes residual connections to mix the original zero-shot visual-linguistic embeddings with the corresponding fine-tuned features, enabling the model to leverage both the knowledge stored in the original CLIP and the newly learned knowledge from few-shot training.

From the results, it can be observed that using Clip-Adapter has limited performance on both datasets. This may be due to the significant disparity between remote sensing images and natural images. Inserting the adapter module after the model makes it challenging to enhance the model's ability to extract remote sensing-specific features at the frontend of the model. In addition, compared with CoOp which inserts tokens only in the first layer of the text modality, the coupling MaPle prompt tuning achieves the best performance, which validates the effectiveness of coupling the language and visual prompts.

IV. CONCLUSION

This paper explores the novel transfer learning approach of prompt tuning, which enables the effective utilization of limited downstream data and achieves promising results for downstream tasks using multimodal pretrained models. Based on this efficient prompt tuning approach, the multimodal pretrained model is applied to remote sensing scene classification. The method is experimented on multiple publicly available remote sensing scene classification datasets, and the results demonstrate that training the learnable prompts with a small amount of samples can yield impressive performance, surpassing the few-shot transfer results of the best-performing pretrained models on these datasets.

REFERENCES

[1] Tao, C., Lu, W., Qi J., Wang H.: Spatial information considered network for scene classification. IEEE Geoscience and Remote Sensing Letters 18(6):984-988 (2020).

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

[3] Bi, H., Yao, J., Wei, Z., Hong, D., Chanussot, J.:PolSAR image classification based on robust low-rank feature extraction and Markov random field. IEEE Geoscience and Remote Sensing Letters, 19: 4005205 (2022).

[4] Bi., H., Perello-Nieto, M., Santos-Rodriguez, R., Flach, P., Craddock, I., An active semi-supervised deep learning model for human activity recognition, Journal of Ambient Intelligence and Humanized Computing, 14:13049–13065 (2023).

[5] Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.: Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13:3735-3756 (2020).

[6] Zhao, X., Zhang, J., Tian, J.: Residual Dense Network Based on Channel-Spatial Attention for the Scene Classification of a High-Resolution Remote Sensing Image. Remote Sensing 12(11):1887 (2020).

[7] Lu, W., Tan, W., Qi, K., Zhang, X., Zhu, Q.:Multi-Output Network Combining GNN and CNN for Remote Sensing Scene Classification. Remote Sensing 14(6):1478 (2022).

[8] Bi, H., Xu, F., Wei, Z., Xue, Y., Xu, Z.: An active deep learning approach for minimally supervised PolSAR image classification. IEEE Transactions on Geoscience and Remote Sensing 57(11):9378-9395 (2019).

[9] Bi, H., Sun, J., Xu, Z. Unsupervised PolSAR image classification using discriminative clustering. IEEE Transactions on Geoscience and Remote Sensing 55(6):3531-3544 (2017).

[10] Bi, H., Xu, L., Cao, X., Xue, Y., Xu, Z. Polarimetric SAR image semantic segmentation with 3D discrete wavelet transform and Markov random field. IEEE Transactions on Image Processing 29:6601-6614 (2020).

[11] Radford, A., Kim, J., Hallacy, C.: Learning transferable visual models from natural language supervision. International conference on machine learning: 8748-8763 (2021).

[12] Li, J., Selvaraju, R., Gotmare, A.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems 34: 9694-9705 (2021).

[13] Bao, H., Wang, W., Dong, L.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems 35: 32897-32912 (2022).

[14] He, K., Zhang, X., Ren, S.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition: 770-778 (2016).

[15] Devlin, J., Chang, M., Lee, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[17] Xia, G., Hu, J., Hu, F.: AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing 55(7): 3965-3981 (2017).

[18] Yang, Y., Newsam, S.: 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 (2010).