Remote Sensing Image Scene Classification With Self-Supervised Paradigm Under Limited Labeled Samples

Preprint *in* IEEE Geoscience and Remote Sensing Letters · October 2020

DOI: 10.1109/LGRS.2020.3038420

CITATIONS
71

READS
693

5 authors, including:

Chao Tao
Central South University
103 PUBLICATIONS **4,778** CITATIONS
SEE PROFILE

Ji Qi
Central South University
23 PUBLICATIONS **559** CITATIONS
SEE PROFILE

Hao Wang
Central South University
9 PUBLICATIONS **278** CITATIONS
SEE PROFILE

Haifeng Li
Central South University
146 PUBLICATIONS **7,094** CITATIONS
SEE PROFILE

Remote Sensing Image Scene Classification with Self-Supervised Paradigm under Limited Labeled Samples

Chao Tao *Member, IEEE*, Ji Qi, Weipeng Lu, Hao Wang and Haifeng Li*, *Member, IEEE*

***Abstract*—With the development of deep learning, supervised learning methods perform well in remote sensing images (RSIs) scene classification. However, supervised learning requires a huge number of annotated data for training. When labeled samples are not sufficient, the most common solution is to fine-tune the pre-training models using a large natural image dataset (e.g. ImageNet). However, this learning paradigm is not a panacea, especially when the target remote sensing images (e.g. multispectral and hyperspectral data) have different imaging mechanisms from RGB natural images. To solve this problem, we introduce a new self-supervised learning (SSL) mechanism to obtain the high-performance pre-training model for RSIs scene classification from large unlabeled data. Experiments on three commonly used RSIs scene classification dataset demonstrated that this new learning paradigm outperforms the traditional dominant ImageNet pre-trained model. Moreover, we analyze the impacts of several factors in SSL on RSIs scene classification task, including the choice of self-supervised signals, the domain difference between source and target dataset, and the amount of pre-training data. The insights distilled from our studies can help to foster the development of SSL in remote sensing community. Since SSL could learn from unlabeled massive RSIs which are extremely easy to obtain, it will be a potentially promising way to alleviate dependence on labeled samples and thus efficiently solve many problems, such as global mapping.**

***Index Terms*—Remote sensing image, scene classification, self-supervised learning (SSL), unlabeled pre-training, limited labeled samples.**

I. Introduction

REMOTE sensing technologies are playing an increasingly important role in global observing missions due to the wide range of observations and high temporal resolution. Particularly, RSIs scene classification, which aims to classify scene images into different semantic categories, has been a hot topic driven by applications such as land resource management and urban planning [1], [2].

To achieve accurate scene classification, how to extract discriminative features from RSIs to precisely represent the semantic content of scene has attracted wide attention. In recent years, with the powerful hierarchical feature extraction capabilities of deep convolutional neural networks (DCNNs), deep learning based methods have made significant progresses in RSIs scene understanding [2]–[4]. However, learning DCNNS generally requires large datasets, and building a big remote sensing dataset like ImageNet (with more than 10 million annotated natural image samples) is almost impossible, as the accurate annotation of RSIs is tedious work requiring rich experience and sound geographic knowledge. Pre-training methods can solve such problem effectively. These methods pre-train the DCNNs on a large labeled RGB natural image dataset (source data) and then fine-tune the network with small remote sensing data (target data). Although many researches [5]–[7] have demonstrated that the features extracted from ImageNet pre-trained DCNNs can generalize well to aerial image scene classification tasks, pre-training CNNs on ImageNet has two limitations: 1) It may provide no benefit if there exists significant domain difference between source and target datasets. 2) Fine-tuning cannot be applied directly if the data types and imaging mechanisms of source and target dataset (natural RGB data vs. multispectral data) are quite different.

Most recently, a new trend is observed in machine learning, which is learning representations by self-supervised learning (SSL) methods without any additional annotation cost [8]. SSL methods can first learn potential useful knowledge from a large amount of unlabeled source data by solving pre-designed tasks (called pretext tasks), then transfer them to target tasks. Inspired by recent advances of SSL in applications like natural language processing [9], nature image classification [10] and object detection [11], we believe that this kind of feature learning mechanism is a more effective and robust way for RSIs scene understanding when labeled data is insufficient. The main reasons could be: first, SSL provides a flexible pre-training architecture, because we can use any type of large-scale remote sensing data without human annotation to pre-train DCNNs; second, since we can choose a source dataset similar to the target dataset at low cost for pre-training, the new learning paradigm can potentially alleviate the domain difference and thus ensure the performance of the learned representations on target RSIs scene classification task. Therefore, we introduce the self-supervised feature learning mechanism for RSIs scene classification, and evaluate the feature learning impact of three commonly used pretext tasks on target RSIs scene classification. As far as we know, this is the first time to use the self-supervised learning mechanism in RSIs scene classification. The contributions of this work are mainly in two aspects:

1) We demonstrate that SSL is an entirely new paradigm which learns feature from unlabeled massive images for remote sensing image understanding. This paradigm is extremely suited to RSI understanding tasks because we have very easy access to a large number of RSIs, over different areas and at different times.

2) Experiments on three RSIs scene classification datasets

This work was supported by National key research and development projects (grant number 2018YFB0504500), National Natural Science Foundation of China (grant number 41771458, 41301453), Young Elite Scientists Sponsorship Program by Hunan Province of China under Grant 2018RS3012, and Hunan Science and Technology Department Innovation Platform Open Fund Project under Grant 18K005.

C. Tao, J. Qi, W.P. Lu, H. Wang and H.F. Li are with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China (Corresponding author: H.F. Li, E-mail: lihaifeng@csu.edu.cn).

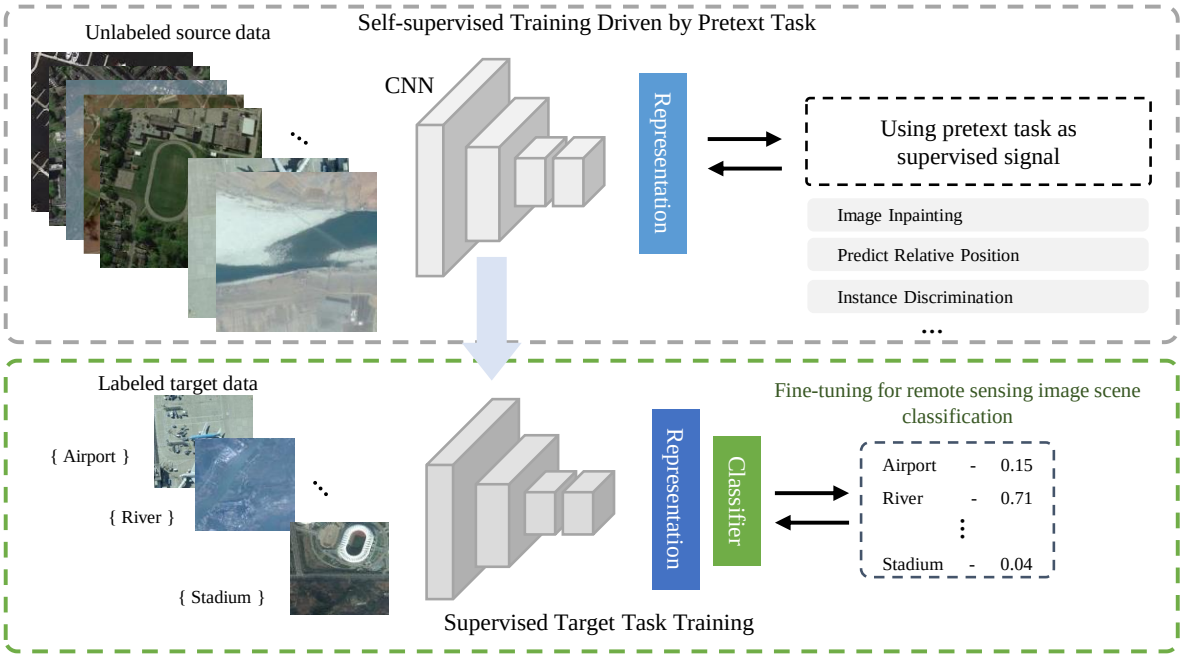

Fig. 1. Flowchart of the self-supervised learning paradigm for remote sensing image scene classification.

show that the proposed method overpass the traditional dominant ImageNet pre-training approach when labeled data is insufficient.

3) We analyze the effects of several factors on the performance of SSL, which contributes to a deeper understanding of what enables useful self-supervised feature representation for RSIs scene understanding.

II. Methodology

A. Overview of self-supervised learning paradigm

In this letter, we suggest a self-supervised learning framework for RSIs scene classification. As shown in Fig. 1, the general pipeline of self-supervised learning paradigm consists of two phases. In the self-supervised training phase, a DCNN is trained to solve predefined pretext tasks for learning potential useful representations on large unlabeled source data. And the learned representations are stored as parameters of the encoder of the DCNN. In the second phase, the learned representations are transferred to target tasks as a pre-trained model. Compared with training from scratch, fine-tuning the pre-trained model with good representations can overcome overfitting and achieve higher performance on target task, especially when labeled samples are insufficient.

B. Learning representations by solving pretext tasks

In the self-supervised learning framework, image inpainting [12], predicting relative position [13] and instance-wise contrastive learning [14] are three common pretext tasks for training DCNN encoders. In the following, we brief the learning algorithms of these three pretext tasks, and evaluate their feature learning impacts on target RSIs scene classification by experiments.

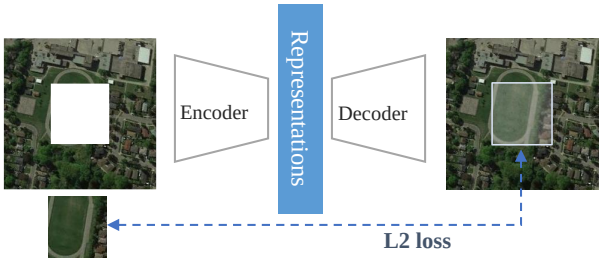

Fig. 2. Illustration of image inpainting pretext task. Given the corrupted image (left), the model is used to restore the missing part (right) based on the rest of the image.

1) Image Inpainting: In image inpainting, model $f(\cdot)$ takes the corrupted image $\tilde{\boldsymbol{x}}_i$ as input and is trained to predict the original image. $\tilde{\boldsymbol{x}}_i$ is obtained by masking arbitrary regions of $\boldsymbol{x}_i$. Generally, the objective function for such a task uses L2 loss as shown in (1). By optimizing Eq. (1), the encoder of DCNN is driven to model pixel-level relationships and local contextual relations within the image for guessing the missing regions based on the rest of the image (Fig. 2).

$$\mathcal{L}_{inpainting} = \|f(\tilde{\boldsymbol{x}}_i) - \boldsymbol{x}_i\|_2^2 \tag{1}$$

2) Predict the Relative Position: Image parts have rich complex spatial or sequential relations, especially natural images. For instance, in portrait photos the head is above the body. Therefore, various models regard recognizing relative positions between parts of images as the pretext task for self-supervised learning. The relative positions could be between two patches from a sample [15], or between shuffled

Fig. 3. A typical example for predicting the relative position: 3×3 Jigsaw puzzles. A DCNN takes nine patches as input and predict their positions.

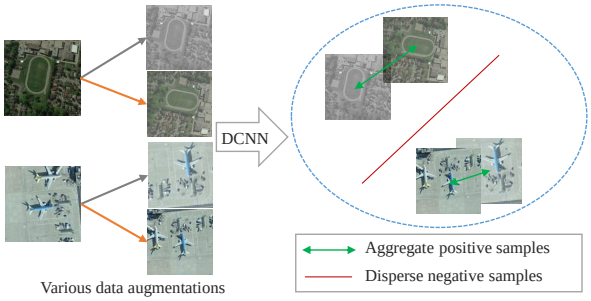

Fig. 4. Illustration of the instance discrimination pretext task. A DCNN is asked to draw near multiple augmentation views of an image sample, and pull away one sample from the other samples by embedding them to a proper feature space.

segments of an image (solve jigsaw) [13], as shown in Fig. 3. Given an image meshed into $m \times n$ patches $\boldsymbol{p}_{i,j}(i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n)$, model $f(\cdot)$ learns contextual relationships between patches by optimizing the loss function in Eq. (2):

$$\mathcal{L}_{jigsaw} = -\sum_{i=1}^{m}\sum_{j=1}^{n} P_{i,j} \log\left(f\left(\boldsymbol{p}_{i,j}\right)\right), \tag{2}$$

where $P_{i,j}$ is the ground truth position of $\boldsymbol{p}_{i,j}$.

3) Instance Discrimination: Different from the above pretext tasks, instance discrimination (ID, or instance-wise contrastive learning) tasks classify examples as their own labels [10], [11]. Specifically, the ID-based self-supervised learning (IDSSL) method takes different augmented views of a sample as positive samples, and takes other different samples as negative ones. Then, a DCNN is trained to distinguish between positive and negative samples by embedding them to a proper feature space with learned representation $f(\cdot)$ (Fig. 4). A minibatch of N samples is augmented to be $2N$ samples $\tilde{\boldsymbol{x}}_i (i = 1, 2, \ldots, 2N)$. For a pair of positive samples $(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)$, other $2(N-1)$ samples are negative ones. Then, a pairwise contrastive loss, the NT-Xent loss [10], is defined as Eq. (3):

$$\ell_{i,j} = -\log \frac{\exp\left(\operatorname{sim}\left(f\left(\hat{\boldsymbol{x}}_i\right), f\left(\hat{\boldsymbol{x}}_j\right)\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{I}_{[k\neq i]} \exp\left(\operatorname{sim}\left(f\left(\hat{\boldsymbol{x}}_i\right), f\left(\hat{\boldsymbol{x}}_k\right)\right)/\tau\right)}, \tag{3}$$

where τ denotes a temperature parameter, and $\mathbb{I}_{[k\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff$[k \neq i]$. The similarity measure function $\operatorname{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ denotes the cosine similarity between vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. $\ell_{i,j}$ is asymmetrical, so the total loss of a minibatch can be computed by Eq. (4):

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}\left[\ell_{2i-1,2i} + \ell_{2i,2i-1}\right]. \tag{4}$$

III. Experiment

A. Datasets Description and Experiment Designing

The scene classification experiments used three public datasets with few labeled samples, which are EuroSAT, AID and NR. These datasets can be divided into low-resolution multi-spectral RSIs datasets and high-resolution RSIs datasets:

1) Multi-spectral RSIs datasets

EuroSAT [16] contains 27,000 samples of 10 categories, collected from Sentinel-2 satellite. These images in 13 bands have a spatial resolution of 30-10m. We used all the 27,000 samples without annotation as the source data for SSL pre-training, and 5,400 samples were used for test.

TABLE I
THE TRAINING-TESTING SET CONFIGURATION OF THREE DATASET IN THE EXPERIMENTS

Dataset	EuroSAT	AID	NR
Classes	10	45	30
Unlabeled samples used for SSL	21,600	25,200	8,000
Labeled samples used for fine-tuning	5/class	5/class	5/class
Samples used for testing	5,400	6,300	2,000

2) High-resolution RSIs datasets

Aerial Image dataset (AID) [17] contains 10,000 samples of 30 classes, collected form Google Earth. These overhead scene images in RGB color space have a resolution of approximately 8-0.5m. We used all the 10,000 samples without annotation for SSL pre-training, and 8,000 for testing.

NWPU-RESISC45 dataset (NR) [18] contains 31,500 samples of 45 scene categories, collected from Google Earth. The spatial resolution varies from about 30m to 0.2m per pixel for most images. We used all the 31,500 samples without annotation for SSL pre-training, and 25,200 for testing.

To evaluate the performance of the SSL-based method on RSIs scene classification task, we carried out the following two experiments in PyTorch environment under the CentOS 7.5 platform with four NVIDIA Tesla V100 (memory 16 GB). The overall accuracy (OA) is used to compare the performance quantitatively.

An overview of the two experiments is as follows:

- Experiment I aims at analyzing several factors that may affect the pre-training performance of SSL on the target RSIs scene classification task, including the choice of self-supervised signals, the domain difference between source and target dataset, and the amount of pre-training data.
- Experiment II aims at evaluating the performance of SSL on the task of RSIs scene classification, and demonstrating the advantage of SSL over other methods when labeled training samples are insufficient.

TABLE II
RESULTS ON THE CHOICE OF SELF-SUPERVISED SIGNALS FOR EXPERIMENT I

Pretext task	OA on target scene classification tasks		
	EuroSAT	AID	NR
Image Inpainting	53.81+1.62	41.15+1.13	17.58+1.22
Predict Relative Position	53.15+1.64	50.32+0.79	34.23+0.85
Instance Discrimination	76.10+0.27	76.80+0.30	80.63+0.03

TABLE III
RESULTS ON THE DOMAIN DIFFERENCE FOR EXPERIMENT I

Source data	OA on target scene classification tasks for experiment I		
	EuroSAT	AID	NR
EuroSAT (30-10m)	76.10+0.27	42.80+0.47	26.85+0.08
AID (8-0.5m)	61.84+0.52	76.80+0.30	54.77+0.15
NR (30-0.2m)	71.45+0.20	74.10+0.07	80.63+0.03

TABLE IV
RESULTS ON THE AMOUNT OF UNLABELED DATA FOR SSL FOR EXPERIMENT I

Source data	Number	OA on target scene classification tasks		
		EuroSAT	AID	NR
EuroSAT	2,160	57.06+0.31	\	\
EuroSAT	10,800	69.12+0.14	\	\
EuroSAT	21,600	76.10+0.27	\	\
AID	800	\	46.88+0.21	\
AID	4,000	\	63.80+0.27	\
AID	8,000	\	76.80+0.30	\
NR	2,520	\	\	39.79
NR	12,600	\	\	68.52
NR	25,200	\	\	80.63+0.03
AID and NR	33,200	72.30+0.15	84.20+0.50	81.13+0.10
AID, NR and EuroSAT	54,800	67.69+0.42	78.60+0.23	75.25+0.24

B. Experiment I

In this section, we performed controlled studies to investigate several factors that may affect the pertaining performance of SSL on the target RSIs scene classification task. In the stage of SSL training, we pre-trained ResNet50 [19] model using the Adam optimizer with a batch size of 256 samples. The learning rate was initially set to 1e-4 and was reduced in a cosine manner within 400 epochs. In the stage of fine-tuning, we used only five labeled samples per category to fine-tune the target task.

1) Study of the choice of self-supervised signals: We evaluated the feature learning performance of different self-supervised signals or pretext tasks on target scene classification task. As shown in TABLE II, the pretext of instance discrimination consistently outperforms other two pretext tasks by a large margin on all three datasets. These results indicate that choosing an appropriate pretext task is crucial, and the correlation between pretext and target tasks plays an important role in learning representative and transferable features. For solving instance discrimination tasks, models are required to learn high-level abstract features with semantic information, which is crucial for classifying RSIs scenes. While in image inpainting and predicting relative position tasks, models are mainly concerned with the pixel-level relationships and local context.

2) Study of domain difference: In this study, we compared the performance of models pre-trained using various source datasets by instance discrimination-based SSL on target datasets[1]. TABLE III shows that pre-training the model using a source dataset similar to the target dataset can make the learned representations suit the target dataset better. For example, source-target pairs with strong domain similarity (e.g., EuroSAT→EuroSAT, AID→AID, and NR→NR) achieves the best results for the three datasets. On the other hand, the performance drops dramatically if the domain difference is large (e.g. from low-resolution multi-spectral RSIs dataset (EuroSAT) to high-resolution RSIs dataset (NR, AID)). This phenomenon exists in most machine learning methods. However, considering SSL training does not need human annotated data, it is efficient and low-cost to obtain source datasets similar to the target dataset.

3) Study of the amount of pre-training data: Since SSL training does not need human annotated data, it is interesting to see whether more pre-training data can bring performance improvements. To this end, given a source training dataset, we created two subsets by randomly sampling 10% and 50% of samples and then evaluated the effect of the amount of pre-training data using different subsets of each dataset. The experiment result is shown in TABLE IV. As can be seen, enlarging the training data size generally leads to better performance because deep learning-based method is data hungry. For all three datasets, the performance improved significantly when the amount of pre-training data increases from 10% to 50%, but the growth rate gradually slows down when the amount of pre-training data increases from 50% to 100%.

C. Experiment II

In this section, we compared the instance discrimination-based SSL (IDSSL) method with two techniques, which are (a) ImageNet pre-trained ResNet50 and (b) multiple-layer feature-matching generative adversarial networks (MARTA GANs) [20]. Here, MARTA GANs is also an unsupervised representation learning method for RSIs scene classification. It learns features with multiscale spatial information from unlabeled images by proposing a multiple-feature-matching layer combined with GANs. For target tasks, we chose the optimal fine-tuning strategy for each method. Considering the non-negligible domain differences between ImageNet and the above three remote sensing datasets, we fine-tuned the entire ImageNet pre-trained ResNet50. For IDSSL, we fixed the parameters of models' encoder and optimized the full connectivity layer. We used Adam optimizer with a batch size of 64. And the learning rate was initially set to 1e-4 and was reduced in a cosine manner within 200 epochs. Consist with [20], we regarded the model obtained by MARTR GANs as an feature extractor. The extracted features were fed into the SVMs for classification. Since the models pre-trained

[1]Here only RGB bands were used for EuroSAT to ensure that the pre-trained models on the three datasets are transferable to each other datasets.

TABLE V
OA OF THE STATE-OF-THE-ART METHODS ON THREE DATASETS FOR EXPERIMENT II

Method	EuroSAT		AID		NR	
	Number of samples per category					
	5	20	5	20	5	20
ResNet50 from scratch [19]	53.60+0.58	74.14+0.69	40.75+1.53	59.28+0.88	32.70+0.40	58.29+1.10
ImageNet pre-trained ResNet50	69.24+0.62	80.91+0.28	72.77+0.09	81.60+0.17	59.63+0.06	73.74+0.17
MATAR GANs [20]	50.92+1.16	68.60+0.28	53.08+0.44	61.90+0.60	43.52+0.18	59.01+0.24
IDSSL	76.10+0.27	84.68+0.03	76.80+0.30	80.62+0.22	80.63+0.03	85.80+0.15

on ImageNet (natural RGB images) cannot be directly fine-tuned on EuroSAT datasets (multi-spectral RSIs), only RGB channels were used on this dataset for ImageNet pre-trained ResNet50.

From the experimental results in TABLE V, we got the following findings. First, IDSSL consistently outperforms ImageNet pre-trained model on all three datasets, especially when using only 5 labeled training samples per category. One possible reason is that the domain difference between ImageNet and remote sensing datasets reduces the performance of the ImageNet pre-trained model. In contrast, SSL provides a flexible pre-training architecture, because we can use any type of large-scale remote sensing data with human annotation to pre-train DCNNs. As a result, the new learning paradigm can potentially alleviate the domain difference and ensure the performance of the learned representations on target RSIs scene classification task. Second, IDSSL is less sensitive to the amount of labeled data than other methods. For example, when the fine-tuning data on NR reduces from 20 per class to 5 per class, the performance of IDSSL decreases by only 6.02%, whereas the relative reduction of ImageNet pre-trained ResNet50 and MATAR GANs are 19.13% and 26.24%, respectively. This might be caused by that with robust representations, IDSSL has little risk of overfitting to labels of small-sized target data.

IV. CONCLUSION AND FUTURE WORKS

In this study, we introduce a new learning paradigm, SSL, for RSIs scene classification for the cases of lacking labeled data. Moreover, we performed comprehensive comparative study by analyzing several factors in SSL on RSIs scene classification task and uncovered that the choice of self-supervised signals, the domain difference between source and target dataset, and the amount of pre-training data strongly affect the pre-training performance of SSL. By combining our findings, the SSL based method outperforms traditional dominant ImageNet pre-training approach as well as other state-of-the-art methods by a large margin when labeled data is insufficient. Future work aims at constructing large scale, publicly available benchmarks for different sensors to foster the development of new SSL methods in remote sensing communities and also using the proposed method to applications such as global mapping which struggle with the limited labeled samples and transferability problems.

REFERENCES

[1] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[2] Y. Gu, Y. Wang, and Y. Li, "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection," *Applied Sciences*, vol. 9, no. 10, p. 2110, 2019.

[3] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[4] C. Tao, W. Lu, J. Qi, and H. Wang, "Spatial information considered network for scene classification," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[5] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, Conference Proceedings, pp. 44–51.

[6] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2015.

[7] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[8] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Conference Proceedings, pp. 4171–4186.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 10 709–10 719.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, Conference Proceedings, pp. 9729–9738.

[12] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, Conference Proceedings, pp. 2536–2544.

[13] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, Conference Proceedings, pp. 69–84.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[15] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, Conference Proceedings, pp. 1422–1430.

[16] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classi-

fication," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[17] G. S. Xia, J. W. Hu, F. Hu, B. G. Shi, X. Bai, Y. F. Zhong, L. P. Zhang, and X. Q. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[18] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, Conference Proceedings, pp. 770–778.

[20] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "Marta gans: Unsupervised representation learning for remote sensing image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2092–2096, 2017.