MRSSC: A BENCHMARK DATASET FOR MULTIMODAL REMOTE SENSING SCENE CLASSIFICATION

Kang Liu [1], Aodi Wu [1,2], Xue Wan[1], Shengyang Li[1,2]*

[1] Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China – (liukang, wuaodi20, wanxue, shyli)@csu.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

KEY WORDS: Benchmark Dataset, Domain Adaptation, Remote Sensing Classification, Optical Remote Sensing Images, Near-nadir SAR

ABSTRACT:

Scene classification based on multi-source remote sensing image is important for image interpretation, and has many applications, such as change detection, visual navigation and image retrieval. Deep learning has become a research hotspot in the field of remote sensing scene classification, and dataset is an important driving force to promote its development. Most of the remote sensing scene classification datasets are optical images, and multimodal datasets are relatively rare. Existing datasets that contain both optical and SAR data, such as SARptical and WHU-SEN-City, which mainly focused on urban area without wide variety of scene categories. This largely limits the development of domain adaptive algorithms in remote sensing scene classification. In this paper, we proposed a multi-modal remote sensing scene classification dataset (MRSSC) based on Tiangong-2, a Chinese manned spacecraft which can acquire optical and SAR images at the same time. The dataset contains 12167 images (optical 6155 and 6012 for optical and SAR, resp.) of seven typical scenes, namely city, farmland, mountain, desert, coast, lake and river. Our dataset is evaluated by state-of-the-art domain adaptation methods to establish a baseline with average classification accuracy of 79.2%. The MRSSC dataset will be released freely for the educational purpose and can be found at China Manned Space Engineering data service website (http://www.msadc.cn). This dataset will fill the gap between remote sensing scene classification between different image sources, and paves the way for a generalized image classification model for multi-modal earth observation data.

1. INTRODUCTION

With the rapid growth of remote sensing satellites, massive multi-source remote sensing data will show an explosive trend of growth. Remote sensing is facing the “big data” challenge. How to mine and extract remote sensing image data quickly and effectively is particularly important. Traditional remote sensing image analysis is based on pixel, such as pixel-level segmentation. Then, several studies focused on object-level classification and segmentation. In this paper, we focused on the scene classification, which aims to automatically assign a semantic label to each scene image, is important for remote sensing image interpretation. As a remote sensing image patch with certain conceptual semantics, scene has become the basic unit of massive remote sensing image classification, which makes it possible to quickly interpret and analyse large-scale remote sensing images. Scene classification is important for remote sensing image interpretation, and has widespread applications, such as change detection (Chen, 2006), urban planning and image retrieval. However, remote-sensing data are often multimodal, e.g., optical (multi- and hyperspectral), and synthetic aperture radar (SAR) sensors, where the imaging geometries and content are completely different. Therefore, the great variations in the spatial arrangements and structural patterns make scene classification a considerably challenging task.

Deep learning has become a research hotspot in the field of remote sensing scene classification, and dataset is an important driving force to promote its development. Most of the remote sensing scene classification datasets are optical remote sensing images, such as UCMerced_Landuse (Yang, 2010), RSSCN7 (Zou, 2015), Aerial Image dataset (AID) (Xia, 2017), etc., and few of them are based on radar images (Hou et al., 2020). As long as there has a large number of optical remote sensing dataset, this paper will investigate whether the knowledge and features from the optical images can be transfer to SAR data, which has fewer labels.

The classification model trained from one data source, for example optical remote sensing images, cannot always transferred successfully to other data source, such as SAR images, owing to the domain difference caused by different sensors. To tackle this problem, the state-of-the-art studies apply domain adaptation (DA), as one of the transfer learning techniques, to solve this problem. It aims to adapt the knowledge learned from one domain, called the source domain, and apply it to another related domain, called the target domain. By reducing the difference of data and feature distribution between the source domain and the target domain, the model trained in the source domain can be transferred and work well in the target domain. The main challenge of the DA problem is that a significant variation exists between the source and the target data distribution; therefore, traditional supervised models trained using the source data without any adaptation will more likely fail in the target domain. The studies of domain adaptation have made great progress in the task of scene classification using daily images. These methods can be divided into three categories, discrepancy-based methods, adversarial-based methods and other methods (Li,2021). Whether the algorithms can be applied for optical and SAR images, which has the huge differences between them, still remain unknown. Although the joint research of optical-SAR data has achieved fruitful results, most of them focus on optical and SAR image registration, SAR-to-optical

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-785-2021 | © Author(s) 2021. CC BY 4.0 License.

image translation (Wang, 2019) and multimodal data fusion (Zhu, 2017).

In this paper, we proposed a multi-modal remote sensing scene classification dataset (MRSSC) based on Tiangong-2, a manned space laboratory launched in September 15，2016. Tiangong-2 can acquire optical and SAR images at the same time, which provides valuable data for scene classification based on multi-modal remote sensing images. There are some multimodal datasets, for example, SARptical dataset (wang, 2017) is a dataset for SAR and optical image matching in dense urban areas and WHU-SEN-City dataset (wang, 2019) covers 32 Chinese cities for SAR-to-optical image translation studies. These optical SAR datasets have only one single scene type, mostly urban, and appear in pairs. While our dataset contains wide variety of scene categories, and the optical image and SAR data, though taken from similar imaging attitude, they are not strictly aligned in pairs, which is more flexible for further DA application.

We also evaluate our dataset using eight state-of-the-art domain adaptation methods to establish a baseline for future research. This dataset will fill the gap between remote sensing scene classification between different image sources, and paves the way for a generalized image classification model for multi-modal earth observation data.

2. MRSSC DATASET

2.1 Introduction to Data Sources

Tiangong-2 is a space laboratory equipped with Wide-band Imaging Spectrometer (WIS), Interferometric Imaging Radar Altimeter (InIRA) and so on. Optical and SAR imagery viewing the same place can be taken by them respectively. The imagery taken by Tiangon-2 have the characteristics of multi-temporal, multi-spatial resolution and multi-modal.

Wide-band Imaging Spectrometer is the first time to achieve the spectral bandwidth of 2.5nm in combination of wideband multispectral imager, which can obtain high signal-to-noise ratio images (Gu, 2019). The signal-to-noise ratio of visible near infrared and short wavelength infrared channels is greater than 800 (20% ground albedo), and the average temperature detection sensitivity of thermal infrared channels is less than 20mk (300K blackbody). The wide band imager adopts push broom and multi view stitching imaging technology, which can obtain 300 km clipping images in 42 ° field of view. Data index of Wide-band Imaging Spectrometer are shown in Table 1.

Index	Visible Near Infrared	Short Wavelength Infrared	Thermal Infrared
spectral range（μm）	0.4~1.0	1~1.7	8~10
numbers of channels	14	2	2
Channel range（μm）	V1： 0.970~0.990 V2： 0.930~0.950 V3： 0.895~0.915 V4： 0.845~0.885 V5： 0.810~0.830 V6： 0.740~0.760 V7： 0.6775~0.6875 V8： 0.655~0.675 V9： 0.610~0.630 V10： 0.555~0.575 V11： 0.510~0.530 V12： 0.480~0.500 V13： 0.433~0.453 V14： 0.403~0.423	S1：1.23～1.25 S2：1.63～1.65	T1：8.125～8.825 T2：8.925～9.275
spatial resolution（m）	100	200	400
field（°）	42	42	42
swath（km）	300	300	300
accuracy of absolute radiation calibration	10%	10%	2K

Table 1. Data index of wide-band imaging spectrometer

The Interferometric Imaging Radar Altimeter (InIRA) is the first microwave remote sensor with wide swath, accurate measurement of ocean topological height and three-dimensional land and sea morphology, which adopts the technologies of small incidence angle short baseline interference, aperture synthesis and height tracking. The spatial resolution of 3D imaging microwave altimeter interferometry observation: the ocean is 10km × 10km, and the altimetry accuracy is better than 8.2cm; the land is 200m × 200m, and the altimetry accuracy is better than 10m; the spatial resolution of 2D imaging is 40m × 40m. Data index of InIRA are shown in Table 2.

Index	Interferometric Imaging Radar Altimeter
working frequency	13.58 GHz
work bandwidth	40 MHz
certainty of backscatter sounding	≤ 2.0 dB
two-dimensional image（convention）	swath 30km，spatial resolution 40 m×40 m
two-dimensional image（high resolution）	swath 5km，spatial resolution 30 m×30 m
DEM（marine）	swath 30km，spatial resolution 10km×10km， relative height measurement accuracy is better than 20cm

DEM（land）	swath 30km，spatial resolution 200m×200m，elevation accuracy is better than 10m

Table 2. Data index of interferometric imaging radar altimeter

To summarize, the optical images and SAR data from Tiangong-2 have the following advantages for scene classification:

Firstly, both the optical images and SAR data are from the same platform, which we can easily get enough multimodal data, especially the same area data taken at the same time.

Secondly, both WIS and InIRA have relatively wide field of view compared to most remote sensing satellite, and thus every image cover a wide variety of typical scenes which make it valuable for scene understanding.

Moreover, compared to the orbit of remote sensing satellite, Tiangong-2 has an orbit of low earth, and this give a unique chance that the data is acquired at different times of the day. This will naturally enhance the time distribution of the data sample.

2.2 MRSSC dataset

2.2.1 Data Acquisition and Processing

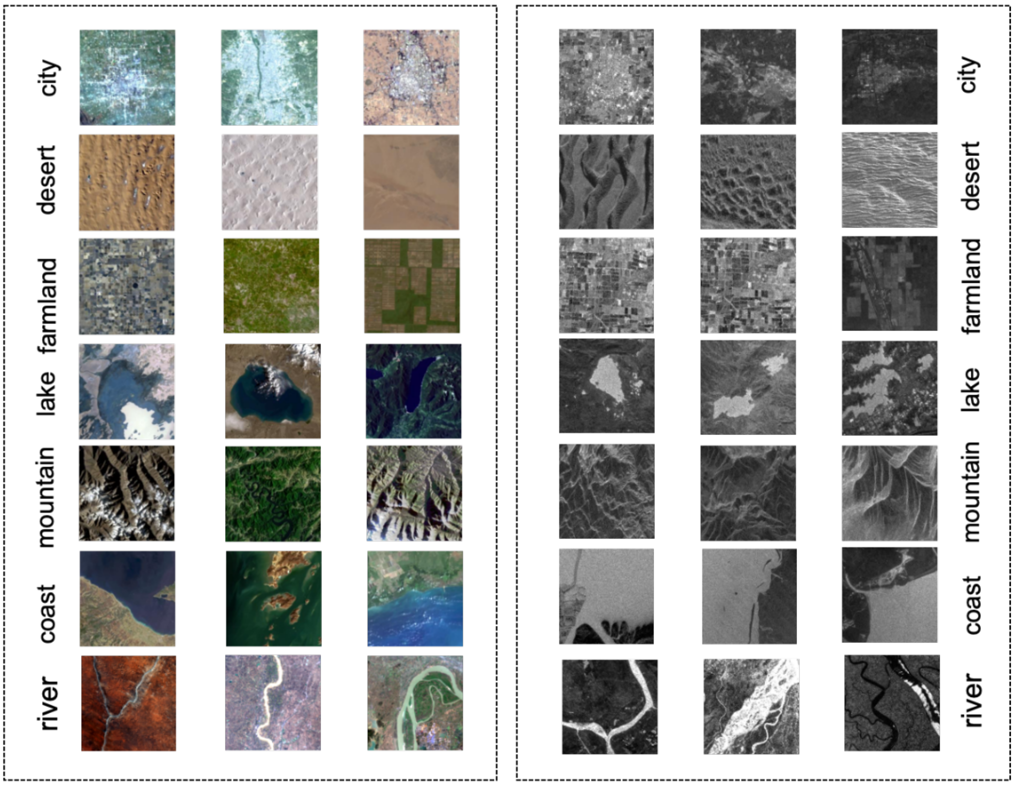
The data products of Wide-band Imaging Spectrometer (WIS) and Interferometric Imaging Radar Altimeter (InIRA) are used as the data sources of optic and SAR images respectively. Among them, the Level 2 product of WIS has been processed by field of view splicing, inter band registration, nonuniformity correction, radiometric correction, sensor correction and geometric correction, with a spatial resolution of 100 meters and 14 bands. The Level 2 product of InIRA has been processed by the following procedures: imaging processing, azimuth multi view processing, radiometric correction and geometric correction to form a two-dimensional image product with map projection，with a spatial resolution of 40 meters.

2.2.2 Remote Sensing Scene Category Selection

According to the characteristics of data, seven types of scenes are selected, namely city, desert, farmland, mountain, lake, coast and river. Optical and SAR data with good imaging conditions and rich distribution of typical objects are selected for clipping. For optical data, images with no cloud or cloud coverage less than <20% are selected. The image scene presents a high degree of diversity and complex heterogeneity. Example images of MRSSC is shown in Figure 1.

In order to ensure the content of image reflect the main scene lable, the main object is located in the middle of the image, and taking up more than 50% pixels of the whole image.

In addition, the data distribution of optical and SAR images should be as balanced as possible, and the areas in the same scene category should have high consistency. Considering the information content of the scene, spatial resolution and adaptability of the algorithm, the image size in MRSSC is 256 × 256 pixels for both optical and SAR images. The GSD is different (100 m and 40 m for optical and SAR, resp.), and thus the dataset contains scenes in different scales. For example, for optical images, we select larger cities, while for SAR images, we select smaller cities. The number of images of each type is shown in Figure 2.



(a) Optical Images (b) SAR Images

Figure 1. Examples images of MRSSC dataset

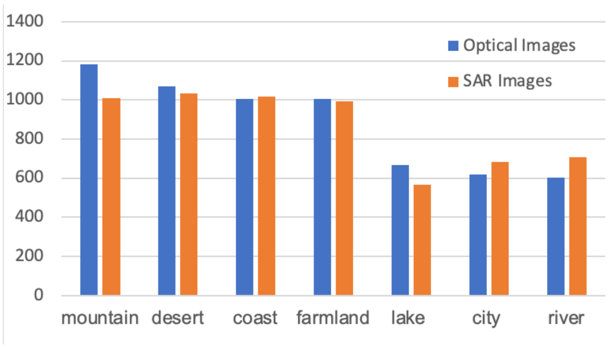


Figure 2. Number of images per class

2.2.3 Data Clipping and Band Selection

ENVI5.3 is used for data clipping. The optical image selection bands of 8,10,12 to synthesize true color images. The SAR image is a single band image, so it can be directly cropped. Both of optical and SAR images were saved .JPG format.

2.3 Characteristics of MRSSC Dataset

Multi-modal remote sensing scene classification dataset (MRSSC) consists more than 12000 images of seven typical scenes, such as city, farmland, mountain, desert, coast, lake and river, specific details can be found in Table 3.

Name of dataset	Multimodal remote sensing scene classification（MRSSC）
scenes	city, desert, farmland, mountain, lake, coast and river
Spatial resolution	100m for Optical; 40m for SAR
Acquisition time	0:00 - 24:00
Scene size	256×256 pixels
Data size	206MB
Data format	.jpg
Data download	http://msadc.cn

Table 3. Multimodal remote sensing scene classification

Compared with the existing multi-modal remote sensing scene classification dataset, MRSSC has the following characteristics:

1) It is the first optical-SAR remote sensing scene classification dataset based on Tiangong-2. Due to the different imaging mechanism, the color and texture information of optical and SAR data are significantly different. They have strong complementarity, and the data distribution has large domain differences, which is challenging for the research of scene understanding for multi-modal remote sensing images.

2) The dataset involves multiple seasons, different weather and different imaging times (different solar elevation angles). Therefore, our dataset contains abundant data distribution, which hopefully can effectively reduce the over fitting of the model and improve the robustness of the model.

3) It has high intra-class diversity and inter-class similarity. High intra-class diversity means that the appearances of samples belonging to the same class are various. For instance, for the category of river, there are many types of rivers with large appearance difference in the dataset, as shown in Figure 3. High inter-class similarity refers to that some samples from different classes have a very similar appearance. For example, in city and farmland scenes include both construction and crop planting area, as shown in Figure 4. This high inter-class similarity reflects the true data distribution in the remote sensing scenes, and will increase the difficulties for scene classification.

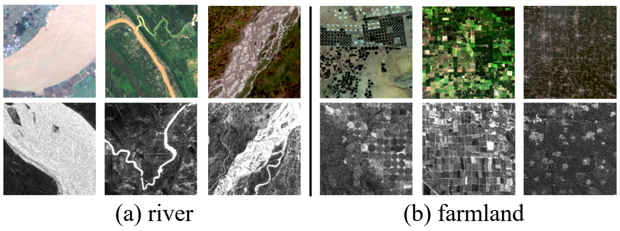


Figure 3. Large diversity within one category. (a) Multiscale images of the same scene. (b) Different styles of the same scene.

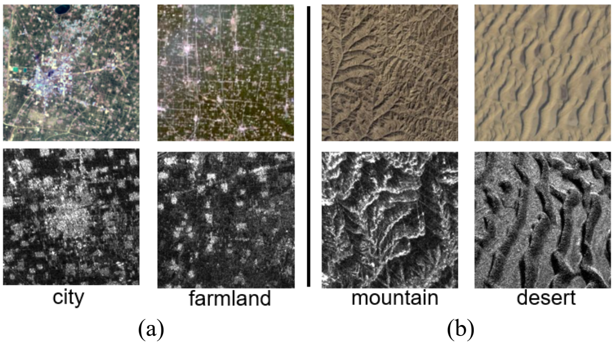


Figure 4. Small interclass distance. (a) Similar objects between different scenes. (b) Similar textures between different scenes.

3. SCENE CLASSIFICATION RESULTS BY BASELINE DOMAIN ADAPTATION METHODS

3.1 Domain Adaptation Baseline Algorithms

To test the effectiveness of the proposed MRSSC dataset for scene classification, experiments are carried out using eight baseline domain adaptation methods, which can be divided into three main categories: discrepancy-based, adversarial-based and others.

Discrepancy-based methods. These DA methods aim to find domain-invariant features in some network layers through all kinds of discrepancy metric minimization techniques. DDC (Tzeng et al., 2014) adds adaptive metric MMD (Gretton et al., 2008) to the penultimate layer of the classification network. DAN (Long et al., 2015) made improvements on the basis of DDC, added three adaptive layers, and adopted multi-core MMD metrics with better characterization capabilities. JAN (Long et al., 2017) aligns the multi-layer joint distribution. MDD (Zhang et al., 2019) defines Margin Disparity Discrepancy with rigorous generalization bounds.

Adversarial-based methods. Inspired by the two-player game, this type of method uses a discriminator to distinguish which domains the data comes from, and at the same time learns a feature extractor that can confuse the discriminator. DANN (Ganin et al.,2015) adds an adversarial mechanism to the network for the first time. CDAN (Long et al., 2018) introduces conditioning target predictions to achieve discriminative adversarial adaptation.

Other methods. AFN (Xu et al., 2019) uses Stepwise Adaptive Feature Norm in order to learn task-specific features with large norms in a progressive manner. MCC (Jin et al., 2020) proposes a new loss function to minimizes the class confusion in the target predictions for Versatile Domain Adaptation (VDA).

The source domain is set as optical data and the target domain is set as SAR data. The rest of experiments in this section will investigate the following questions: 1) Can baseline DA methods increase the scene classification accuracy compared to source only approach, which directly test the classification model trained by optical images on SAR data? 2) Which types of DA methods perform better in the case the of large appearance difference between optical and SAR data?

3.2 Experimental Settings

The MRSSC dataset is divided into three parts: the source domain, the target domain and the test set. There is no data intersection between the target domain and the test set. The source domain includes optical images and the target domain includes SAR images. The images in source domain contain labels for training, while the images in target domain do not have labels and will be used for domain adaption. The number of categories in the three parts is in Table 4.

	Source (optical)	Target (SAR)	Test (SAR)
city	621	584	100
coast	1005	916	100
desert	1069	935	100
farmland	1004	895	100
lake	667	466	100
mountain	1184	910	100
river	605	606	100

Table 4. Division of MRSSC dataset

For a fair comparison, the model training and testing settings for eight baseline DA methods are the same and shown in Table 5.

GPU	Tesla T4
backbone	ResNet-50
epochs	10
mini batch	32
optimizer	SGD (momentum=0.9, weight decay=0.001)

Table 5. Experimental settings

We use DALIB (Jiang et al., 2020), a transfer learning library, developed by Tsinghua University to implement the methods, and the backbone is ResNet-50 pre-trained on ImageNet. In the training phase, the source and target images are randomly cropped to 224×224, and perform random horizontal flip as the input of the network. In the test phase, the test set images are cropped from centre to 224×224 for predictive input.

3.3 Results

3.3.1 Overall Accuracy

The labelled source domain data and the unlabelled target domain data are used to train the network, and then the test set is used to test the classification accuracy of the network. The evaluation index used is overall accuracy (OA):

$$OA = \frac{1}{N}\sum_{i=1}^{r} x_{ii} \tag{1}$$

where N = number of test images, which is 700 in this paper
i = the index of each class
r = number of classes, which is 7 in this paper
x_{ii} = number of correct predictions in each class

In Table 6, the overall accuracies of the eight baseline DA methods are summarized, which Last OA is the overall accuracy of the test set after 10 epochs of training, and Best OA is the performance of the best model obtained during the 10 epochs of training. Source Only refers to the ResNet-50 classification network, which only uses source domain data for training, and directly test on SAR without DA. From the Best OA results, it can be seen that compared with Source Only, all the DA methods have improved the scene classification accuracy by average of 31.3%. Among them, CDAN has the best performance, with an overall accuracy of 84.4%. Although the existing DA method can improve the classification results, but the best test result does not exceed 90%. This demonstrates that future work can be done to further increase the scene classification accuracy.

Method	Last OA (%)	Best OA (%)
Source Only	46.0	47.9
DDC	63.7	66.6
DAN	69.1	75.0
JAN	79.1	82.7
MDD	81.9	82.9
DANN	79.9	84.0
CDAN	84.4	84.4
AFN	78.7	79.0
MCC	77.7	79.0

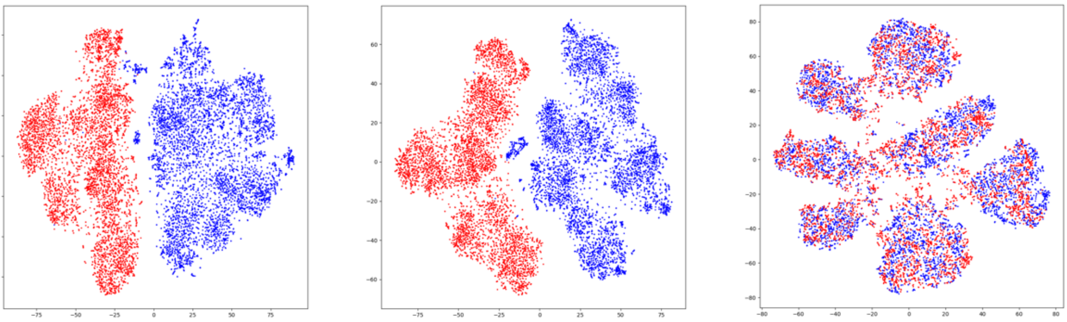
Table 6. Classification accuracy of each algorithm

3.3.2 t-SNE Analysis

In order to analyse the features distribution of the source domain and target domain images extracted by the network, we visualize t-SNE (Van der Maaten et al., 2008) embeddings of the features from ResNet-50 without training, ResNet-50 trained on source domain (Source Only) and CDAN respectively as shown in Figure 5. Note that points of different colours indicate different domains, and when the points of the same class are gathered into a cluster, it means that there is better separability. It can be seen from Figure 5(a) that the source domain and target domain data are distributed in two separate areas. Figure 5(b) shows that the ResNet-50 network trained only with source domain data cannot align the source domain and target domain features well, and at the same time, the target domain data distribution does not form a good category boundary. From Figure 5(c), it can be seen that the network trained with CDAN can very well align the features from source domain and target domain. Meanwhile, CDAN make inter-class separated and intra-class clustered tightly. This visualization shows that the network trained by the DA method can improve the performance of cross-modal scene classification on the proposed dataset MRSSC.

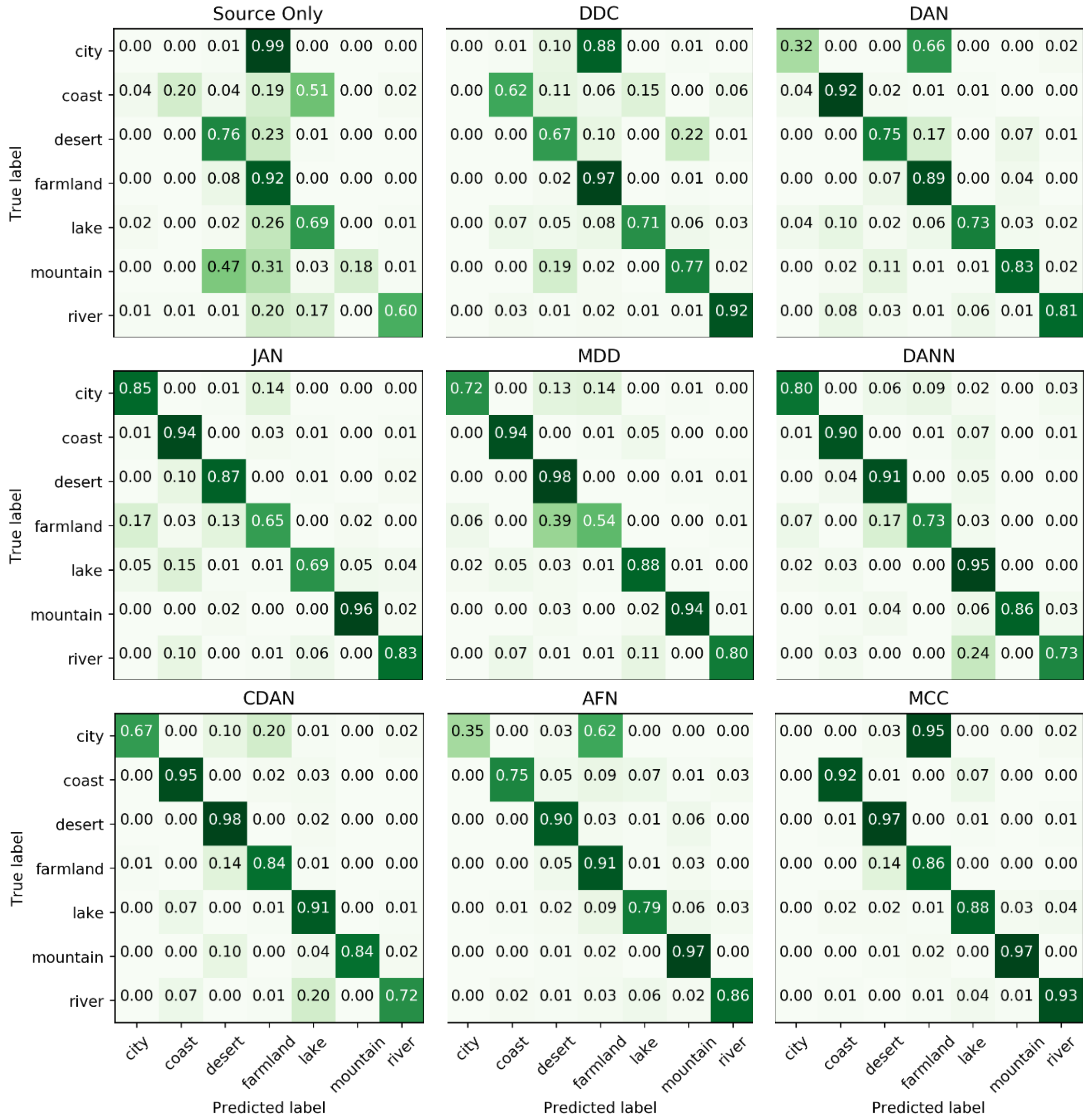
3.3.3 Scene Classification Confusion Matrix

Figure 6 shows the confusion matrix obtained from the test set prediction results, which can be used to study the distribution of each class in the MRSSC cross-modal classification experiment



(a) ResNet-50 without training (b) Source Only (c) CDAN

Figure 5. The t-SNE visualizations of (a) ResNet-50 without training, (b) Source Only and (c) CDAN, where blues points indicate source domain data and red points indicate target domain data



Source Only

	city	coast	desert	farmland	lake	mountain	river
city	0.00	0.00	0.01	0.99	0.00	0.00	0.00
coast	0.04	0.20	0.04	0.19	0.51	0.00	0.02
desert	0.00	0.00	0.76	0.23	0.01	0.00	0.00
farmland	0.00	0.00	0.08	0.92	0.00	0.00	0.00
lake	0.02	0.00	0.02	0.26	0.69	0.00	0.01
mountain	0.00	0.00	0.47	0.31	0.03	0.18	0.01
river	0.01	0.01	0.01	0.20	0.17	0.00	0.60

DDC

	city	coast	desert	farmland	lake	mountain	river
city	0.00	0.01	0.10	0.88	0.00	0.01	0.00
coast	0.00	0.62	0.11	0.06	0.15	0.00	0.06
desert	0.00	0.00	0.67	0.10	0.00	0.22	0.01
farmland	0.00	0.00	0.02	0.97	0.00	0.01	0.00
lake	0.00	0.07	0.05	0.08	0.71	0.06	0.03
mountain	0.00	0.00	0.19	0.02	0.00	0.77	0.02
river	0.00	0.03	0.01	0.02	0.01	0.01	0.92

DAN

	city	coast	desert	farmland	lake	mountain	river
city	0.32	0.00	0.00	0.66	0.00	0.00	0.02
coast	0.04	0.92	0.02	0.01	0.01	0.00	0.00
desert	0.00	0.00	0.75	0.17	0.00	0.07	0.01
farmland	0.00	0.00	0.07	0.89	0.00	0.04	0.00
lake	0.04	0.10	0.02	0.06	0.73	0.03	0.02
mountain	0.00	0.02	0.11	0.01	0.01	0.83	0.02
river	0.00	0.08	0.03	0.01	0.06	0.01	0.81

JAN

	city	coast	desert	farmland	lake	mountain	river
city	0.85	0.00	0.01	0.14	0.00	0.00	0.00
coast	0.01	0.94	0.00	0.03	0.01	0.00	0.01
desert	0.00	0.10	0.87	0.00	0.01	0.00	0.02
farmland	0.17	0.03	0.13	0.65	0.00	0.02	0.00
lake	0.05	0.15	0.01	0.01	0.69	0.05	0.04
mountain	0.00	0.00	0.02	0.00	0.00	0.96	0.02
river	0.00	0.10	0.00	0.01	0.06	0.00	0.83

MDD

	city	coast	desert	farmland	lake	mountain	river
city	0.72	0.00	0.13	0.14	0.00	0.01	0.00
coast	0.00	0.94	0.00	0.01	0.05	0.00	0.00
desert	0.00	0.00	0.98	0.00	0.00	0.01	0.01
farmland	0.06	0.00	0.39	0.54	0.00	0.00	0.01
lake	0.02	0.05	0.03	0.01	0.88	0.01	0.00
mountain	0.00	0.00	0.03	0.00	0.02	0.94	0.01
river	0.00	0.07	0.01	0.01	0.11	0.00	0.80

DANN

	city	coast	desert	farmland	lake	mountain	river
city	0.80	0.00	0.06	0.09	0.02	0.00	0.03
coast	0.01	0.90	0.00	0.01	0.07	0.00	0.01
desert	0.00	0.04	0.91	0.00	0.05	0.00	0.00
farmland	0.07	0.00	0.17	0.73	0.03	0.00	0.00
lake	0.02	0.03	0.00	0.00	0.95	0.00	0.00
mountain	0.00	0.01	0.04	0.00	0.06	0.86	0.03
river	0.00	0.03	0.00	0.00	0.24	0.00	0.73

CDAN

	city	coast	desert	farmland	lake	mountain	river
city	0.67	0.00	0.10	0.20	0.01	0.00	0.02
coast	0.00	0.95	0.00	0.02	0.03	0.00	0.00
desert	0.00	0.00	0.98	0.00	0.02	0.00	0.00
farmland	0.01	0.00	0.14	0.84	0.01	0.00	0.00
lake	0.00	0.07	0.00	0.01	0.91	0.00	0.01
mountain	0.00	0.00	0.10	0.00	0.04	0.84	0.02
river	0.00	0.07	0.00	0.01	0.20	0.00	0.72

AFN

	city	coast	desert	farmland	lake	mountain	river
city	0.35	0.00	0.03	0.62	0.00	0.00	0.00
coast	0.00	0.75	0.05	0.09	0.07	0.01	0.03
desert	0.00	0.00	0.90	0.03	0.01	0.06	0.00
farmland	0.00	0.00	0.05	0.91	0.01	0.03	0.00
lake	0.00	0.01	0.02	0.09	0.79	0.06	0.03
mountain	0.00	0.00	0.01	0.02	0.00	0.97	0.00
river	0.00	0.02	0.01	0.03	0.06	0.02	0.86

MCC

	city	coast	desert	farmland	lake	mountain	river
city	0.00	0.00	0.03	0.95	0.00	0.00	0.02
coast	0.00	0.92	0.01	0.00	0.07	0.00	0.00
desert	0.00	0.01	0.97	0.00	0.01	0.00	0.01
farmland	0.00	0.00	0.14	0.86	0.00	0.00	0.00
lake	0.00	0.02	0.02	0.01	0.88	0.03	0.04
mountain	0.00	0.00	0.01	0.02	0.00	0.97	0.00
river	0.00	0.01	0.00	0.01	0.04	0.01	0.93

Figure 6. Confusion matrix of nine classification algorithms

It can be seen in Figure 6, in the confusion matrices of DDC, DAN, AFN and MCC, city is easily to be mis-predicted as farmland. As shown in Figure 4(a), sometimes, there are contain similar objects in city and farmland. From the Table 4, it can be seen that the number of farmland is larger than the number of city, which may be the reason for the low accuracy of city class. While for JAN and MDD this dis-prediction phenomenon does not happen.

In comparison, this dis-prediction phenomenon does not happen for the two adversarial-based methods, DANN and CDAN. These experimental results demonstrate that the adversarial-based algorithms are able to distinguish the appearance difference between city and farmland. However, they sometimes misclassify rivers into lakes. In addition, they may also misclassify desert and mountain.

4. CONCLUSIONS

This paper proposed a multi-modal scene classification dataset, MRSSC. The dataset involves different regions, different seasons, different weather and different imaging time (different solar elevation angles). The dataset contains 12167 images of seven typical scenes. In addition, we have evaluated the new dataset using state-of-the-art domain adaptation methods to establish a baseline for future research. For the first time, we verified the effectiveness of the domain-adaptive algorithm in optical and SAR. Although the optical and SAR data are very different due to their different imaging mechanisms, the experimental results demonstrate that the domain adaptive algorithm can reduce the difference in data distribution between different domains, improve the accuracy of scene classification.

The MRSSC dataset also remain several open questions and research directions for further works including how to deal with the imbalance class distribution and how to cope with the large inter-class similarity problem. We do hope that the dataset will inspire innovative research ideas and algorithms in multi-modal remote sensing image classification.

ACKNOWLEDGEMENTS

The work is supported by the Open Research Fund of National Earth Observation Data Center (No.NODAOP2020012).

REFERENCES

Chen, X. L., Zhao, H. M., Li, P. X., & Yin, Z. Y. (2006). Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote sensing of environment*, *104*(2), 133-146.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2015). Domain-Adversarial Training of Neural Networks. *arXiv e-prints*, arXiv-1505.

Gu, Y., Gao, M., & Zhao, G. (Eds.). (2018). *Proceedings of the Tiangong-2 Remote Sensing Application Conference: Technology, Method and Application* (Vol. 541). Springer.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2008). A Kernel Method for the Two-Sample Problem. *Journal of Machine Learning Research*, *1*, 1-10.

Hou, X., Ao, W., Song, Q., Lai, J., Wang, H., & Xu, F. (2020). FUSAR-Ship: building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition. *Science China Information Sciences*, *63*(4), 1-19.

Jiang, J., Fu, B. Long, M., 2020. Transfer-Learning-library. github.com/thuml/Transfer-Learning-Library.

Jin, Y., Wang, X., Long, M., & Wang, J. (2020, August). Minimum Class Confusion for Versatile Domain Adaptation. In *European Conference on Computer Vision* (pp. 464-480). Springer, Cham.

Li, S., Xie, B., Lin, Q., Liu, C. H., Huang, G., & Wang, G. (2021). Generalized Domain Conditioned Adaptation Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015, June). Learning transferable features with deep adaptation networks. In *International conference on machine learning* (pp. 97-105). PMLR.

Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018, January). Conditional Adversarial Domain Adaptation. In *NeurIPS*.

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017, July). Deep transfer learning with joint adaptation networks. In *International conference on machine learning* (pp. 2208-2217). PMLR.

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019). SEN12MS-a Curated Dataset of Georeferenced Multi-Spectral SENTINEL-1/2 Imagery for Deep Learning and Data Fusion. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 153-160.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., & Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Wang, L., Xu, X., Yu, Y., Yang, R., Gui, R., Xu, Z., & Pu, F. (2019). SAR-to-optical image translation using supervised cycle-consistent adversarial networks. *IEEE Access*, *7*, 129136-129149.

Wang, Y., Zhu, X. X., Zeisl, B., & Pollefeys, M. (2017). Fusing Meter-Resolution 4-D InSAR Point Clouds and Optical Images for Semantic Urban Infrastructure Monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(1), 14-26.

Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., ... & Lu, X. (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 3965-3981.

Xu, R., Li, G., Yang, J., & Lin, L. (2019). Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1426-1435).

Yang, Y., & Newsam, S. (2010, November). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 270-279).

Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019, May). Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning* (pp. 7404-7413). PMLR.

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8-36.

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters, 12*(11), 2321-2325.