

ORIGINAL ARTICLE

Open Access



Cross-temporal high spatial resolution urban scene classification and change detection based on a class-weighted deep adaptation network

Sunan Shi¹, Yanfei Zhong^{1,2*}, Yinhe Liu¹, Liangpei Zhang¹ and Deren Li¹

Abstract

Multi-temporal urban scene classification and change analysis based on high resolution (HR) remote sensing imagery can provide reliable time-series information for the semantic interpretation of urban land use and the transitional relationships, which is important information for urban planning and sustainable development. However, there are still some difficulties encountered when applying the existing multi-temporal scene classification methods to analyze urban development in China due to the complex urban structure and scene shape. The main reasons for this can be summarized as follows: 1) the multi-temporal data labeling workload caused by the differences of the data distributions among multi-temporal images; and 2) the lack of practical socio-geographical urban unit boundaries resulting from the uniform grid based segmentation. In this paper, a multi-temporal scene classification framework based on a class-weighted deep adaptation network (CWDAN) is proposed. In the CWDAN framework, multi-temporal OpenStreetMap (OSM) road networks are introduced for the scene segmentation at the land parcel level, to build clear and meaningful geographic boundaries for the scene units. The problem of large scale difference of parcels is solved by area-weighted voting (AWV). In order to solve the problems of the high workload of multi-temporal data labeling in the cross-temporal scene classification task, a gradient reversal layer (GRL) is used in the proposed CWDAN to obtain deep features with invariance relative to the shift between the domains. A class-weighted fully connected layer is used to solve the problem of unbalanced proportion of different urban scene classes. Post-classification is finally performed to obtain the scene change information. Experiments with tri-temporal datasets in Chinese areas demonstrated that the proposed framework can obtain a significantly improved performance in the cross-temporal scene classification and change analysis task.

Keywords Cross-temporal scene classification, Urban change analysis, Spatial-temporal-aligned deep adaptation network, High resolution remote sensing imagery, Deep adaptation network, OpenStreetMap road network

1 Introduction

As global urbanization has accelerated in recent decades, the environmental, traffic, and demographic issues that arise in the urbanization process have become of great concern to the international community (Grimm et al., 2008). Accurate judgment of urban development modes and in-depth analysis of the current urban problems can be considered as one way to achieve sustainable urban development. Multi-temporal classification and change detection based on remote sensing imagery is a commonly used method in urban development analysis.

*Correspondence:

Yanfei Zhong
zhongyanfei@whu.edu.cn

¹ State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, P.R. China

² Hubei Provincial Engineering Research Center of Natural Resources Remote Sensing Monitoring, Wuhan University, Wuhan, P.R. China

As high resolution (HR) Earth observation technology is developing rapidly, a large quantity of HR images can now be obtained. The plentiful contextual information and spatial details provided by the HR imagery can help us to obtain richer land surface distribution and change information (Bouziani et al., 2010) in specific urban tasks. There has been a long history of pixel-based (Polykretis et al., 2020; Johnson and Kasischke, 2010) and object-oriented (Shi et al., 2020; Hussain et al., 2013). multi-temporal classification and change detection research. However, what we need in practical urban applications is the urban land-use pattern and change information, such as the transition from farmland to residential and the expansion of industrial areas. Due to the diversity of objects and the complex spatial distribution, the traditional pixel-based and object-oriented methods are unable to obtain the high-level semantic information expressed by HR images from the low-level data features, which poses a great challenge for practical urban analysis.

With the aim of better solving the problems of the traditional methods and achieve an understanding of the high-level semantic information, the concept of the “scene” has been proposed (Zhao et al., 2013). Urban scenes refer to the basic urban units with clear geographical boundaries and socio-economic attributes, which can express the high-level semantic land-use information. The spatial aggregation of different urban land-use patterns can create different urban scenes, which can further reflect the urban functional distribution behind the urban scenes (Zhao et al., 2013). On account of the abundant and detailed structural and spatial information provided by HR imagery (Zhu et al., 2016), multi-temporal scene classification and change detection can allow us to obtain multi-temporal urban land-use distribution and change information at a higher semantic level (Ni and Wu, 2019). This can help us to accurately grasp the spatial patterns and rules of urban land expansion, reasonably analyze urban development patterns, and correctly guide the sustainable use of urban land-use resources.

Based on the research in the field of natural image processing, the methods of remote sensing scene classification have been continuously developed from low-level feature based methods (Oliva and Torralba, 2001) and mid-level feature based methods (Zhu et al., 2016) to deep learning based methods (Yong Wang et al., 2019). However, methods at the first two stages are based on manual features. Therefore, these methods can obtain good application results only in certain cities, but cannot achieve satisfactory feature adaptability. Differing from the methods based on manually designed features, as an end-to-end framework, convolutional neural networks (CNNs) can realize automatic deep feature extraction

and scene class prediction (Wu et al., 2017). Due to the data-driven characteristic of deep learning, the CNN-based multi-temporal methods all require a large amount of data labeling for the model training on multi-temporal images, which is time-consuming and does not fully utilize the prior knowledge of the existing labeled datasets (Ru et al., 2021; Hu et al., 2015). The deep learning model needs to follow the assumption that the training and test sets should be independently and identically distributed. Therefore, if we do not perform additional multi-temporal data labeling and use the well-trained model of one time phase for multi-temporal scene classification, the model generalization ability will be insufficient due to the differences in data distribution and diverse distribution patterns within the scene among multi-temporal images. In addition, the existing methods of multi-temporal urban scene classification and change detection all use a uniform grid for uniform scene patch segmentation, which makes the mapping results have a serious mosaic phenomenon and can not match with the actual urban application.

For the purpose of making the most of the previously labeled data while ensuring a satisfactory accuracy for the model in the multi-temporal scene classification and change detection tasks, transfer learning can be introduced to reduce the workload of multi-temporal data labeling. The idea of transfer learning is defined as the ability to transfer the knowledge obtained in the previous task (source domain) to a new task (target domain), enabling satisfactory learning results for the target domain (Mansour et al., 2009; Sinno Jialin Pan and Yang, 2010; Jiang and Zhai, 2007; Audebert et al., 2017; Gopalan et al., 2011). Domain adaptation is commonly used in transfer learning strategy (Matasci et al., 2015). So far, cross-temporal classification and change detection based on domain adaptation have not been carried out at the scene level, although there is some research basis at the re-pixel level (Hu et al., 2015; Demir et al., 2013).

In this paper, a multi-temporal scene classification framework based on a class-weighted deep adaptation network (CWDAN) is proposed. In this framework, multi-temporal OpenStreetMap (OSM) road networks and uniform grid sampling are introduced for mapping at the land parcel level and model training at the patch level, respectively. In order to reduce the workload of multi-temporal data labeling, a domain adversarial neural network (DANN) with a gradient reversal layer (GRL) is introduced so that the labeled source scene patches and the unlabeled target scene patches are used as the network inputs for cross-temporal scene classification to obtain deep features with invariance relative to the shift between the domains. The problem of large scale difference of parcels is solved by the area-weighted voting

(AWV) strategy instead of the traditional equal-weighted voting (EWV) strategy. A class-weighted (CW) fully connected layer is used to solve the problem of unbalanced proportion of different urban scene classes during different time phases. Post-classification is finally applied to obtain the multi-temporal scene change information.

The main contributions of this paper are as follows:

- 1) A multi-temporal scene classification framework based on a class-weighted deep adaptation network (CWDAN) is proposed to reduce the multi-temporal data labeling workload. In the proposed framework, the idea of domain adaptation is introduced. The training data of only one time phase are labeled, and cross-temporal scene classification is achieved by feature alignment. Compared with the existing multi-temporal scene classification change detection frameworks, the proposed framework can drastically reduce the multi-temporal data labeling workload, as well as improve the model training efficiency. Tri-temporal scene classification and change detection mapping in study areas of Wuhan urban center and the Wan Chai of Hong Kong was undertaken to verify the effectiveness of the proposed CWDAN framework.
- 2) As urban areas develop, the proportion of urban scene classes contained in the cross-temporal images will change. When the source domain contains a different proportion of categories than the target domain, it has a negative impact on the model training. Thus, the original DANN network is improved by adding the calculation of class weights in the fully connected layer, which means that, for a category with a low proportion in the target domain, its weight will be increased to balance the classification effect, and vice versa. The CWDAN model can solve the negative transfer problem caused by the category imbalance and can significantly improve the cross-temporal classification accuracy.
- 3) Multi-temporal land parcel level scene classification and change detection maps are acquired. To solve the problem of the mosaic phenomenon caused by uniform patch-based scene segmentation, OSM road network data are added to achieve land parcel level scene segmentation. The uniform patch-level scene segmentation is also performed as the input of the model training and prediction, which is finally mapped to the parcel level through the area-weighted voting (AWV) strategy for scene classification mapping and post-classification change detection. The results of multi-temporal scene classification and change detection based on land parcels can better reflect the urban layout and urban change informa-

tion, thus providing favorable support for practical urban application analysis.

1.1 Related work

1.1.1 Scene classification and change detection based on bi-temporal labels

Scene classification methods based on low- and mid-level features have a history of more than ten years, and then few change detection methods based on multi-temporal scene classification came into being. A bag-of-visual-words (BOVW) model based bi-temporal scene classification and change detection framework was proposed by Wu et al. (2016). In this method, the BOVW model is used to encode different kinds of low-level features. Support vector machine (SVM) is then applied to classify the bi-temporal images separately, and post-classification change detection is finally performed to obtain the scene change information. This work laid a good foundation for the subsequent research on multi-temporal scene classification and change detection. However, for multi-temporal cases, this method does not consider the temporal correlation. To solve this problem, a scene change detection framework which introduced kernel slow feature analysis (KSFA) was proposed by Wu et al. (2017). This method also applies the BOVW model to perform scene classification, but a histogram intersection kernel (HIK)-based KSFA module is added in the classification process to obtain the change probabilities of the bi-temporal scene images. After obtaining the bi-temporal classification results and the temporal correlation established by KSFA, post-classification with Bayesian theory is utilized to determine the “from-to” change information of the scene pairs for change detection. This framework integrates the temporal correlation of bi-temporal images, and can achieve a superior classification and change detection accuracy. However, both of the above methods are based on manually designed mid-level features in the feature extraction stage.

In recent years, CNNs have been widely used and have been successfully applied in many areas of remote sensing, including image classification (Shin et al., 2016), target detection (Gopalan et al., 2011), semantic segmentation (Girshick et al., 2013; Chen et al., 2014), and scene classification tasks (Cheng et al., 2018; Limin Wang et al., 2016). At the same time, a scene classification and change detection framework based on a CNN has also been developed by Yong Wang et al. (2019). This framework is based on bi-temporal images, and the CNN is applied to extract spectral-spatial features. Nonlinear high-dimensional transformation is then conducted based on the deep features of bi-temporal images, and deep canonical correlation analysis (DCCA) is applied

to extract the change information. Ru et al. (2021) proposed the CorrFusion module to perform the bi-temporal scene classification and scene change detection task. This method also uses a CNN to extract the bi-temporal scene features, while the temporal correlation is calculated through the CorrFusion module. Cross-temporal fusion is then finally performed based on the extracted features and the temporal correlation calculation results. All the existing methods require bi-temporal classification labels to train models for multi-temporal classification and change detection, and there is a large demand for manual labeling.

1.2 Scene classification and change detection based on single temporal labels

Domain adaptation has been widely applied in remote sensing image classification (Long et al., 2015; Long Liu et al., 2018) and change detection (Pan et al., 2011). For example, Wei Liu et al. (2019) applied an unsupervised land-use change detection method, which detects the unchanged regions and transfers the unchanged samples of the source domain (historical data) to the target domain (new data), to implement land-use change updating. Demir et al. (2013) and Hu et al. (2015) both applied a deep CNN to realize transfer between different classification datasets. However, multi-temporal classification and change detection methods based on domain adaptation have not been developed at the scene level up to now.

2 Methods

In order to reduce the workload of multi-temporal data labeling and to address the needs of urban development analysis in practical city planning, the CWDAN framework is proposed. Firstly, tri-temporal OSM road network data and uniform grid sampling were used to divide the scene units into land parcels and patches for the final multi-temporal scene classification and change detection mapping and model training. To reduce the cost of tri-temporal data labeling, only one time phase of the image was labeled to construct the training dataset, and the proposed CWDAN framework was then used to obtain the cross-temporal scene classification results. After obtaining the tri-temporal scene results for image patches, Area-weighted voting (AWV) was applied to generate the scene classification maps at the land parcel level. Finally post-classification was applied to obtain the final multi-temporal scene change maps.

2.1 Training datasets and multi-level scene segmentation

2.1.1 Class system for cross-temporal scene classification

For the urban scene classification standards and mapping rules used in this study, we synthesized the “Urban Land

Classification and Planning Construction Land Standard (GB 50137-2011)” to form the parcel-level scene classification system of Residential area (Re), Educational and research area (Er), Sports area (Sa), Commercial service facilities (Cs), Industrial area (Ia), Transportation facilities (Tf), Park (Pa), Square (Sq), Green buffer (Gb), Water (Wa), Agricultural and forest (Af), Vacant area (Va), Other construction land (Oc) and Mixed land (MI). In this study, functional areas of urban scenes that cannot be directly identified by remote sensing images, including “Oc” and “MI”, were labeled for future research, but were not involved in the accuracy assessment of this study.

2.2 Multi-temporal road network processing

For the original downloaded OSM road network data, there was relatively serious data redundancy and topological relationship errors. Therefore, it was necessary to manually correct the original road network data. Due to the unavailability of the historical OSM road network data, production of the historical road network data based on the obtained road network data was also needed. Firstly, for the original road network data, the topological errors were corrected by extending or cropping to eliminate the hanging nodes in the road networks. Some overlapping redundant road networks were also manually removed. Secondly, based on the corrected road networks and the historical HR image data, road network correction was executed through manually adding or removing the road network data to obtain the historical road network. For a multi-temporal study, the road networks will change over time. For example, some minor roads may disappear, major roads may increase in size, and some road sections may be merged. In order to ensure that the numbers of land parcels obtained from each temporal division were consistent and the accuracy of the road network was guaranteed, a “union” operation that combines all the multi-temporal road network data was performed so that the intersecting parts of the multi-temporal road networks can independently form multiple component elements, that is, a parcel may be divided into multiple parcels due to partial changes in its contents. Finally, the complete road network is obtained for multi-temporal research.

In accordance with the attributes detailed in the OSM road network dataset and the road widths specified in the “Code for Transport Planning on Urban Road (GB50220-95),” which was effective until 2019, we established various buffer radii for distinct road classes (Table 1), to generate the buffer results for the final road networks. To address the misalignment issues caused by image registration errors, the buffer radii were reduced by 10 meters to ensure that the acquired parcels are completed. The road network buffers were then erased from the overall

Table 1 Buffer zone settings for the different road classes

Urban road hierarchy	Road class	Buffer zone setting (meters)
Branch road	footway	5
	living_street	5
	pedestrian	5
	residential	5
	service	5
	steps	5
	tertiary	5
	tertiary_link	5
Secondary road	motorway	15
	motorway_link	15
	primary	15
	primary_link	15
	road	15
	secondary	15
	secondary_link	15
Main road	trunk	25
	trunk_link	25
unclassified	unclassified	15

surface region of the study areas, to obtain the parcel-level scene segmentation results.

2.3 Cross-temporal scene classification and change detection based on a class-weighted deep adaptation network (CWDAN)

In the cross-temporal scene classification task, we used the labeled data of only one time phase (source domain) to train the network and implement the classification task for the remaining time phases (target domain). However, as a result of the different light situations, sensor conditions, and radiation, there was a discrepancy in the data distributions among the different time phase images. Furthermore, imbalance between the categories of the target domain and source domain can also reduce the classification accuracy.

To solve the above-mentioned problem, a class-weighted DANN-based cross-temporal scene classification model is proposed in this paper. We use M to represent the scene patch images segmented by uniform grid sampling, and N to represent their labels. The source time phase images can then be described as $S = \{X_l^s, Y_l^s\}$, and the target time phase images can be described as $T = \{X_l^t, Y_l^t; X_u^t\}$. The superscripts of X and Y are s or t , indicating source domain or target domain, respectively; and the subscripts of X and Y are u or l , indicating labeled or unlabeled, respectively. It should be noted that the target domain description consists of

two parts, where $T_u = \{X_u^t\}$ is the unlabeled patch data to be processed by the proposed model to obtain the corresponding image labels, and $T_l = \{X_l^t, Y_l^t\}$ is the partially labeled patch data for the evaluation of the algorithm accuracy at the patch level. Finally, the multi-temporal scene classification framework based on a class-weighted deep adaptation network (CWDAN) is summarized as Fig. 1.

The tri-temporal OSM road network data and uniform grid sampling are respectively used to achieve parcel-level and patch-level scene segmentation. For the parcel level, all of the time phase images are labeled for the final scene classification mapping accuracy evaluation. For the patch level, the CWDAN is trained by the labeled scene patches of the first time phase. The trained model is then used to obtain the scene labels for the remaining time phase images. After obtaining the multi-temporal scene classification results at the patch level, parcel-patch mapping with the AWV strategy is applied to obtain the final parcel-level scene classification results. Post-classification change detection is then used to acquire the final change information. The overall accuracy of the scene classification and change detection is evaluated by the number of land parcels correctly classified as a percentage of the total number of land parcels.

2.4 The class-weighted DANN

To solve the redundant data labeling task of multi-temporal data classification, we use the idea of transfer learning and use the DANN as the benchmark network. DANN is tasked with accomplishing the following objectives. 1) Accurate Source Domain Classification. The initial task involves achieving precise classification for the source domain dataset while minimizing classification errors in the existing images. 2) Domain Confusion and Adaptation. The subsequent task focuses on inducing confusion between the source domain dataset and the target domain dataset. This aims to maximize the error in domain classification, thereby facilitating the interchange of characteristics between these domains. Moreover, the process extends to confusing the target domain dataset with the source domain dataset to further enhance adaptation. In the training process, image classifier and domain classifier confront each other using the a GRL to achieve the balance between image classification loss and domain classification loss. The loss of traditional DANN can be defined as Eq. (1).

$$L_{TD} = L_{cls}(F(x_l^s), y_l^s) + \frac{1}{2} (L_{adv}(Dd(grl(F(x_l^s))), 1), L_{adv}(Dd(grl(F(x_u^t))), 0)) \quad (1)$$

where L_{cls} represents the classification loss of the source domain and calculated by Cross Entropy. L_{adv} is

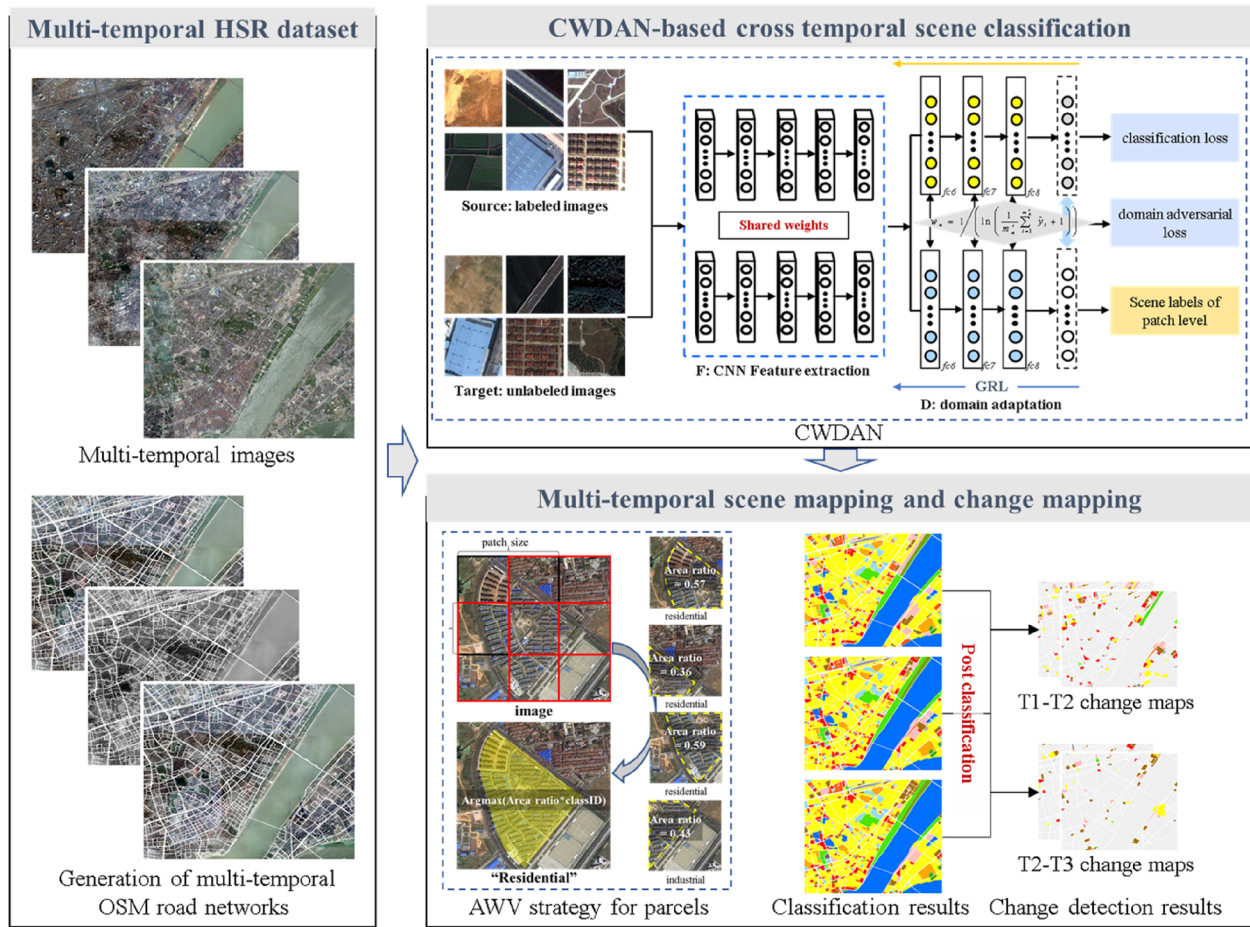


Fig. 1 Flowchart of the urban scene analysis framework based on the proposed CWDAN

adversarial domain confusion loss calculated by Cross Entropy. F is the feature extractor. D_d is a domain discriminator and grl is the gradient reversal layer.

When directly classifying images in a new time phase (target domain), the different proportion of scene categories in the source domain and the target domain can have a great impact on the classification accuracy. Therefore, to solve this problem, a new weight parameter w_a as shown in Eq. (2) is introduced, where $\hat{y}_i$ is the predicted probability for the scene patch images with label a , and m_a^t represents the number of scene patch images labeled as a in the target domain. For a category that has a high proportion in both the source domain and the target domain, the probability of being correctly predicted in the target domain is high, and the number of correct predictions increases accordingly. However, for a category that has a low proportion in the source domain, the probability of being correctly predicted in the target domain is low due to the difficulties for model learning. Thus, the weights for categories with low proportions will be higher to balance the

model's attention to these categories and improve the classification accuracy for the target domain.

$$w_a = 1 / \left(\ln \left(\frac{1}{m_a^t} \sum_{i=1}^{m_a^t} \hat{y}_i + 1 \right) \right) \quad (2)$$

Equation (3) shows the final loss function of the proposed CWDAN model, which consists of two parts, namely, the class-weighted classification loss and the domain-adversarial loss. By minimizing the classification loss and the domain-adversarial loss, optimal classification results can be obtained. The key structure of the proposed CWDAN model is shown in Fig. 2.

$$L = w_a L_{cb}(F(x_i^s), y_i^s) + \frac{1}{2} (L_{adv}(D_d(grl(F(x_i^s))), 1), L_{adv}(D_d(grl(F(x_i^t))), 0)) \quad (3)$$

2.5 Multi-level parcel-patch category mapping system

In order to allow the multi-temporal data to be input into the deep learning network for training and classification,

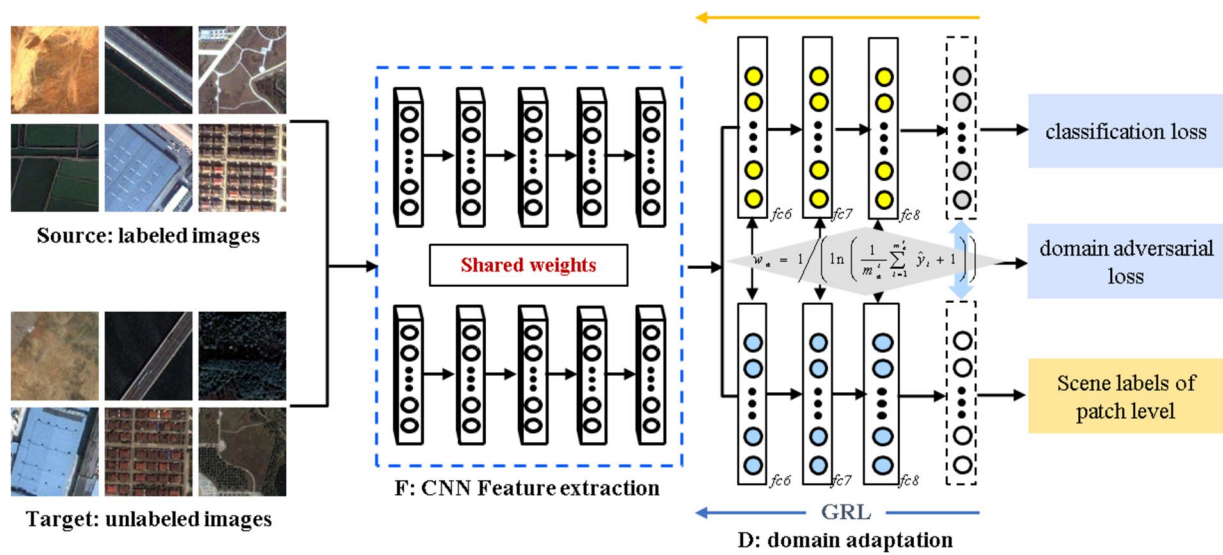


Fig. 2 The structure of the proposed CWDAN model

we performed scene segmentation on the original images using uniform grid sampling to obtain the patches to be input into the network, and the network then output the scene category labels for each patch of the target images. Utilizing patches derived from uniform grid sampling for model prediction and mapping can induce a mosaic phenomenon. Therefore, to acquire parcel-level mapping results that contain true geographical boundaries, it is imperative to establish a mapping relationship between the patch level and the parcel level. Traditional approaches employ an EWV strategy. However, when the parcel shape is elongated or irregular, portions of the patches belonging to other parcels may increase, interfering with the voting results of the current parcel.

Consequently, an area-weighted voting strategy is introduced in this study, realizing a multi-level parcel-patch category mapping system to ameliorate this issue.

As shown in Fig. 3, in order to form the final mapping results at the parcel level, AWV effectively considers the scale difference and shape diversity of different parcels by voting according to the parcel area to which each patch belongs so as to obtain more robust mapping accuracy. The association between parcel and patch is managed through a one-to-many voting mechanism. Only those patches that overlap with the current parcel participate in the voting process for the corresponding parcel. And these patches should collectively cover all areas of the current parcel.

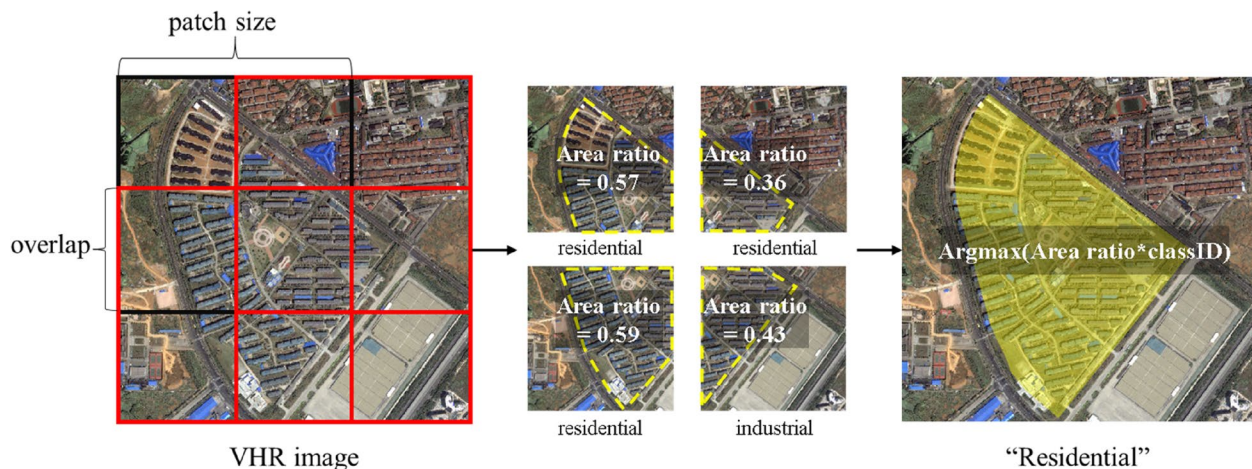


Fig. 3 The AWW strategy

2.6 Scene change detection based on post-classification comparison

After obtaining the parcel-level multi-temporal classification results, post-classification is finally applied to obtain the multi-class scene change information. Differing from the traditional change detection accuracy evaluation method of using pixel-based indices, we treat the multi-class change information as different categories, and the accuracy is evaluated by the number of parcels that obtain the correct change information.

Finally, through the multi-temporal scene classification and change detection mapping, the time-series urban land-use distribution and change pattern analysis is obtained by calculating the parcel number of each category in the different time phases and combining this with the change detection results.

3 Experiments and results

In this section, we describe the results obtained for: 1) dataset construction used to validate the proposed method; 2) accuracy evaluation and mapping for multi-temporal scene classification; 3) accuracy evaluation and mapping for multi-temporal scene change detection.

3.1 Data collection

In this study, Chinese urban areas with diverse urban structures, city scales, and development models were chosen as the study areas, focusing on Wuhan City and Hong Kong Special Administrative Region. Wuhan is the capital of Hubei province, and plays a core role in the Yangtze River Economic Belt. It is also a strategic pivot point for the “Rise of Central China” plan (Wenmei Li et al., 2020). Hong Kong is a special administrative region of China of high prosperity, which is located in southern China and east of the mouth of the Pearl River. Hong Kong is the third-largest financial center in the world, and is ranked as the third-largest first-tier city in the world by the GaWC Research Network (Shang and Cha, 2013). By selecting the above study areas, the differences of urbanization development in different cities can be revealed.

This study was carried out based on tri-temporal HR images. For Wuhan city center, the image acquisition dates were June 2000, June 2010, and May 2017. The image size of these three images is 15424×12834 pixels and the spatial resolution is 0.5 m. For the Wan Chai area of Hong Kong, the image acquisition dates were June 2000, June 2010, and January 2017. The image size of these three images is 8192×7680 pixels and the spatial resolution is 0.5 m. The multi-temporal OSM road network data of the two study areas were obtained from both the OSM website (the road network for 2017) and by manual annotation and correction (the road networks for 2000 and 2010). The

tri-temporal images and ground-truth maps at the parcel level are presented in Fig. 4.

In this study, the first time phase image of both datasets was used as the source domain image. The patches obtained based on the uniform grid sampling were used to construct the training datasets for the model training. The datasets used in this study were sliced using a 300×300 uniform grid with 150 pixel overlap. The number of patches obtained by the uniform grid sampling was 8555 for Wuhan center and 1784 for Wan Chai, separately. In order to simulate the situation that the open-source data generally cannot cover the whole study area, 80% of the patches of each class in the source domain images were labeled to form the final training data. 20% of the patches of each class in the images of three time phases were labeled for accuracy assessment at the patch level.

To express the multi-class change type in a simple way, the scene categories of the before and after phases were presented respectively in the changed area in two single figures. As can be seen from the ground-truth maps, the changes of Wuhan center mainly focus on the transition between residential and commercial. In 2017, a large number of built-up areas in Wuhan center were redeveloped. Meanwhile, the Wan Chai area of Hong Kong has seen little change. Some constructive activity appeared at the harbors.

3.2 Multi-temporal scene classification

The three constructed tri-temporal datasets were used to carry out the related experiments to evaluate the effectiveness of the CWDAN framework for the cross-temporal scene classification and change detection task. In the training procedure, the batch size and epochs were respectively set as 16 and 100. 80% of the patches on the first time phase are labeled for model training to simulate the incomplete coverage of historical products in practical applications. The ResNet-50 network was used as the backbone network and was pretrained using ImageNet. The experimental results obtained with other domain adaptation methods are also exhibited for comparison, i.e., DANN (Ganin and Lempitsky, 2015), deep adaptation networks (DAN) (Long et al., 2015), joint adaptation networks (JAN) (Long et al., 2017), the adversarial discriminative domain adaptation (ADDA) (Tzeng et al., 2017), conditional domain adversarial networks (CDAN) (M. S. Long et al., 2018), maximum classifier discrepancy for unsupervised domain adaptation (MCD) (Saito et al., 2018) margin disparity discrepancy (MDD) (Zhang et al., 2019) and minimum class confusion (MCC) (Jin et al., 2020).

We quantitatively analyze the results of the final multi-temporal scene classification at the parcel level in Table 2.

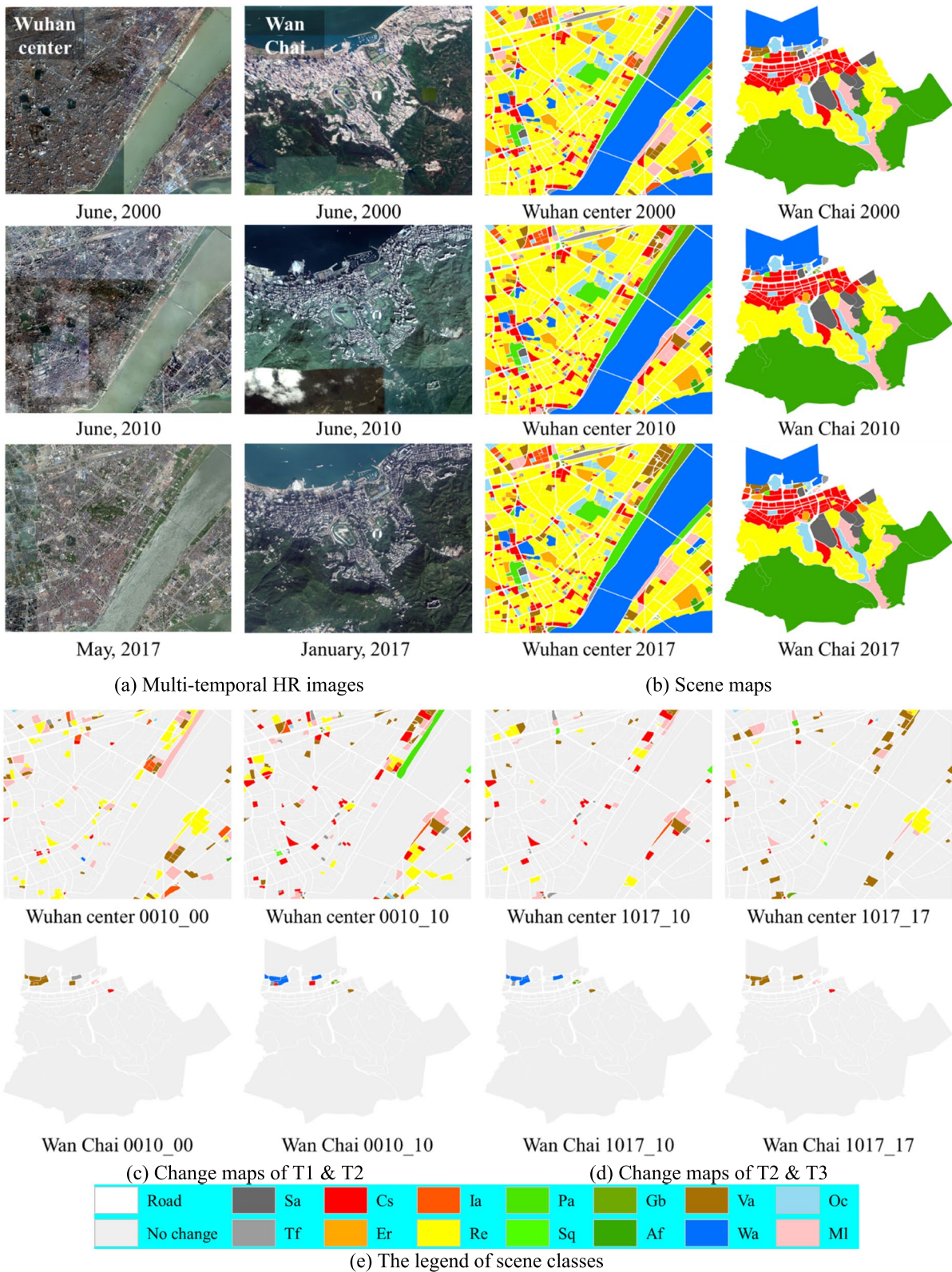


Fig. 4 The ground-truth maps at the parcel level for the three HR image datasets. In the classification maps, the numbers within each subtitle designate its mapping year. For the change maps, the first four numbers represent the years before and after the changes, and the last two numbers represent the year of the parcels shown in the current change map. **a** Multi-temporal HR images. **b** Scene maps. **c** Change maps of T1 & T2. **d** Change maps of T2 & T3. **e** The legend of scene classes

Compared to the other methods, the proposed CWDAN framework obtains the best classification accuracy on all new time phases for both datasets, which shows that the proposed CWDAN has good migration effect. The overall accuracy (OA) of the first-time phase is not optimal because the small amount of data brings the problem of codomain overfitting so that a small decrease in accuracy is acceptable. In addition, compared with DANN, the proposed CWDAN has a significant improvement in time consistency. The proposed AWV strategy works for most methods on most time phases of both datasets. In addition, the images of both the Wuhan center dataset and the Wan Chai dataset have different degrees of image stitching. However, the results obtained by the proposed framework show that these problems do not have a great impact on the final cross-temporal classification results, indicating that the proposed cross-temporal scene classification framework is insensitive to the problems of cross-sensor data and image stitching.

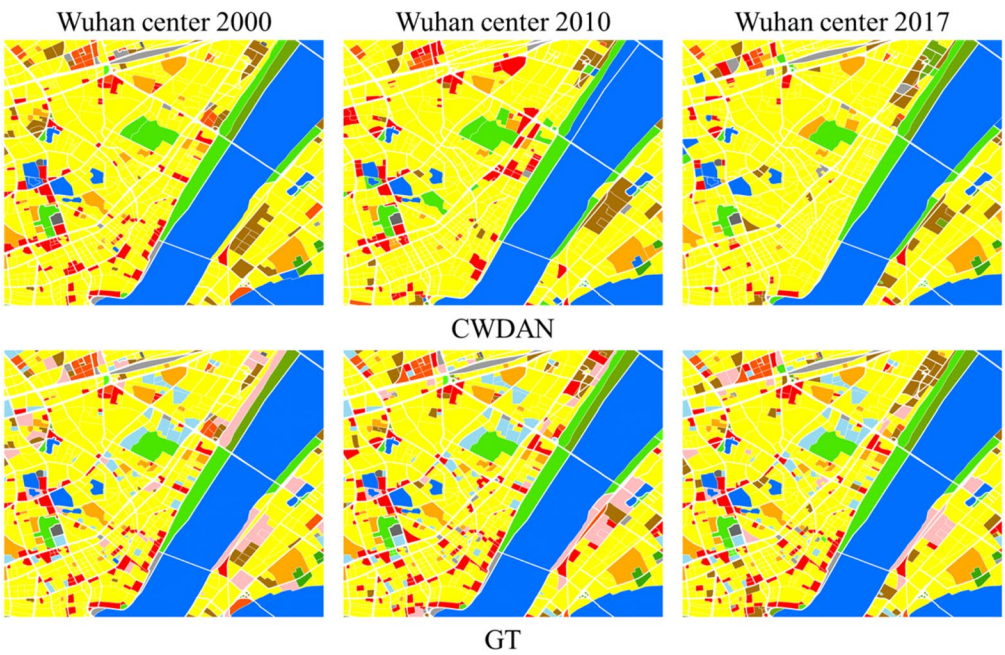
As can be seen from the multi-temporal mapping results (Fig. 5), the large areas of water and farmland regions are correctly classified in both datasets. The classification mapping results for the Wuhan center dataset show that, from 2000 to 2017, Wuhan city center was dominated by the residential and water categories, along with some commercial and green land categories. For

the Wan Chai dataset, the Wan Chai area of Hong Kong was dominated by the categories of commercial, residential, agricultural and forestry land, and water from 2000 to 2017. Among the different categories, agricultural and forestry land is mainly distributed in the southern part, while water is mainly distributed in the northern part. The residential and commercial areas are concentrated in the central region. The mapping results of the CWDAN-based multi-temporal scene classification can provide time-series urban land-use information, thus providing useful support for realistic urban development analysis.

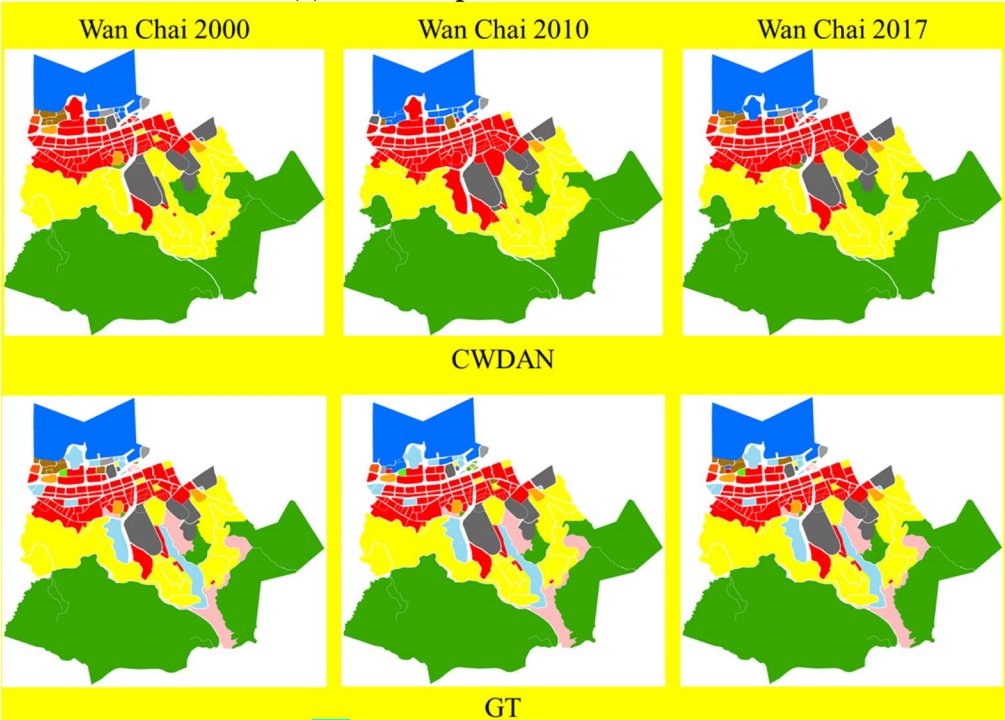
Figure 6 presents the confusion matrices for the scene classification results obtained using the CWDAN framework, where the columns show the classification results and the rows show the ground truth. For the Wuhan center dataset and Wan Chai dataset, the classification accuracy is highest for the residential and commercial categories, respectively. However, for the categories of sports, square, and transportation, the classification accuracy is relatively low. Due to the actual situation of the urban areas in general, the distributions of the different categories show great variance, which leads to the problem of sample imbalance in the training dataset. For example, as for Wuhan city center and Wan Chai area of Hong Kong, the residential and commercial categories have a larger coverage area, so that their sample numbers

Table 2 Accuracy of the multi-temporal scene classification at the parcel level with different comparison methods

Methods	EWV	AWV	Wuhan center (OA, %)			Wan Chai (OA, %)		
			T1	T2	T3	T1	T2	T3
Source_only	√		95.83	62.86	66.39	95.78	61.68	74.70
		√	96.25 (+0.42)	62.31 (-0.55)	66.67 (+0.28)	95.78 (+0.00)	61.68 (+0.00)	74.70 (+0.00)
DAN	√		93.33	64.92	66.80	96.39	73.05	81.33
		√	93.75 (+0.42)	65.61 (+0.69)	67.63 (+0.83)	96.39 (+0.00)	72.46 (-0.59)	80.72 (-0.61)
DANN	√		97.36	68.50	68.59	95.78	79.04	81.33
		√	97.64 (+0.28)	69.33 (+0.83)	69.68 (+1.09)	95.78 (+0.00)	77.84 (-1.20)	80.72 (-0.61)
JAN	√		95.28	68.36	69.41	96.99	65.27	78.92
		√	95.42 (+0.14)	69.88 (+1.52)	70.78 (+1.37)	96.99 (+0.00)	67.07 (+1.80)	78.92 (+0.00)
ADDA	√		82.50	66.16	66.94	92.17	76.65	75.30
		√	82.78 (+0.28)	66.71 (+0.55)	68.18 (+1.24)	92.17 (+0.00)	77.25 (+0.60)	77.11 (+1.81)
CDAN	√		94.72	68.23	68.45	96.99	77.84	81.33
		√	95.14 (+0.42)	68.09 (-0.14)	69.00 (+0.55)	96.99 (+0.00)	77.25 (-0.59)	81.33 (+0.00)
MCD	√		95.28	64.79	66.39	96.99	68.86	69.88
		√	95.69 (+0.41)	64.92 (+0.13)	67.49 (+1.10)	96.99 (+0.00)	68.26 (-0.60)	69.88 (+0.00)
MDD	√		95.83	68.64	65.57	97.59	73.65	78.31
		√	96.25 (+0.42)	69.74 (+1.10)	66.39 (+0.82)	97.59 (+0.00)	73.65 (+0.00)	78.31 (+0.00)
MCC	√		96.81	59.70	64.20	95.18	73.05	70.48
		√	96.67 (-0.14)	59.28 (-0.42)	63.92 (-0.28)	95.18 (+0.00)	73.05 (+0.00)	71.08 (+0.00)
CWDAN	√		96.53	71.53	71.33	96.99	80.24	81.93
(proposed)		√	96.81 (+0.28)	72.35 (+0.82)	73.39 (+2.06)	96.99 (+0.00)	80.84 (+0.60)	81.93 (+0.00)



(a) Scene maps for Wuhan center



(b) Scene maps for Wan Chai



(c) The legend of scene classes

Fig. 5 The CWDAN-based multi-temporal scene classification mapping results for the HR image datasets. **a** Scene maps for Wuhan center. **c** Scene maps for Wan Chai



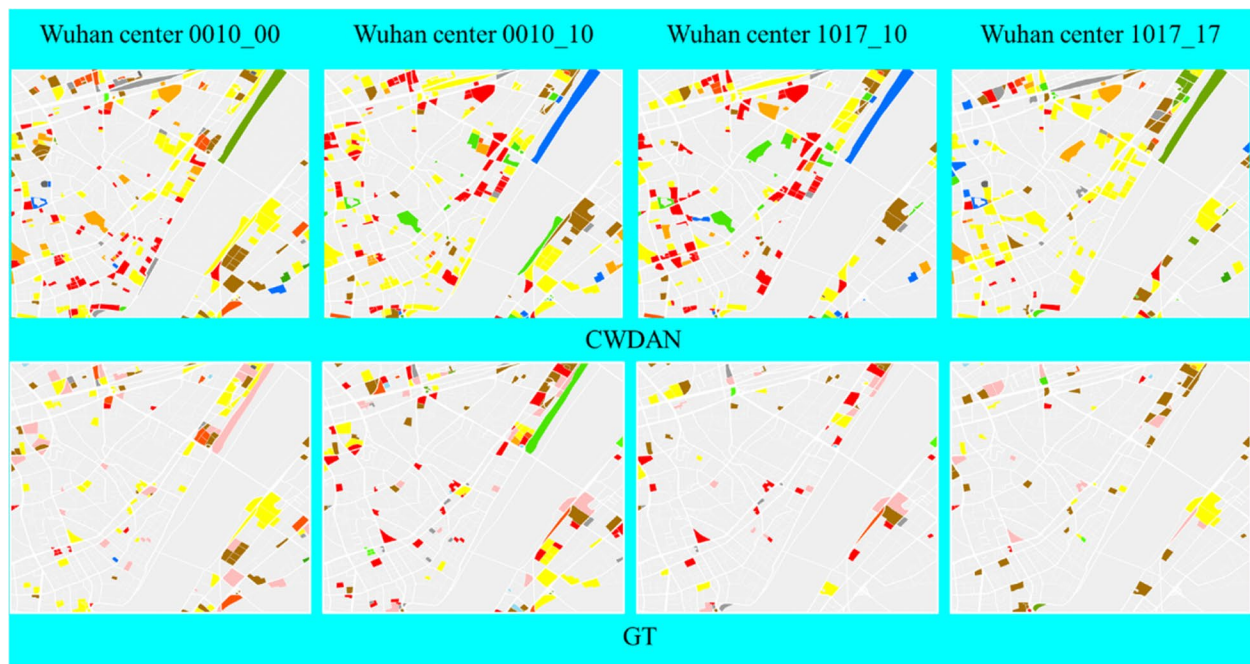
Fig. 6 The CWDAN-based multi-temporal scene classification confusion matrices for the two datasets

in the training dataset are sufficient. In contrast, the distribution of the sports, square, and transportation categories is limited, as a result of the small sample sizes of these scene categories. Thus, the categories with more training samples can obtain a better classification accuracy, but for categories with few samples, they achieve a poor accuracy, due to the limited training data. However, with regard to the overall classification results, the proposed CWDAN alleviates the negative effects of category imbalance to a certain extent and maintains a good accuracy.

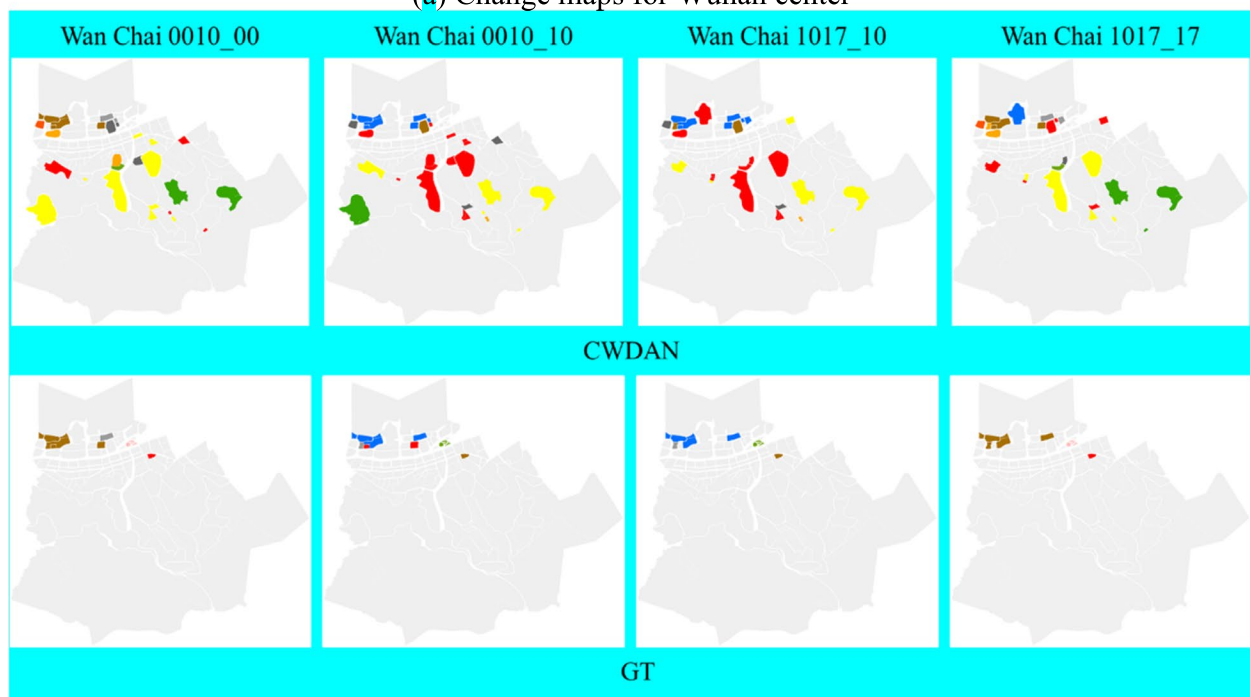
There is also a misclassification phenomenon in both datasets. For the Wuhan center dataset and the Wan Chai dataset, the decrease of the classification accuracy can be attributed to the misclassification between residential, commercial, and educational and research land. The commercial and residential categories are a combination of uniformly arranged buildings, while the school category is also a combination of uniformly arranged buildings, as well as green land, sports facilities, and so on. In addition, the dual use of commercial and residential buildings is also common. Thus, the similarity of the

Table 3 Accuracy of the multi-temporal scene change detection at the parcel level using different comparison methods with AWV

Methods	Wuhan center (OA, %)		Wan Chai (OA, %)	
	T1-T2 change	T2-T3 change	T1-T2 change	T2-T3 change
Source_only	62.61	60.31	62.05	51.81
DAN	65.62	63.11	72.89	65.06
DANN	69.91	64.94	77.71	72.29
JAN	70.34	64.94	67.47	60.84
ADDA	66.19	64.38	75.90	69.28
CDAN	68.19	64.52	77.71	74.10
MCD	65.47	60.31	68.07	60.24
MDD	69.48	63.11	74.10	66.87
MCC	59.74	56.10	72.29	57.83
CWDAN(proposed)	72.64	64.94	80.72	75.30



(a) Change maps for Wuhan center



(b) Change maps for Wan Chai

Fig. 7 The CWDAN-based multi-temporal scene change detection mapping. **b** Change maps for Wuhan center. **c** Change maps for Wan Chai

composition within the scene categories results in the misclassification problem.

3.3 Multi-temporal scene change detection

Finally, the multi-temporal scene change detection results were obtained by post-classification comparison. The overall accuracies are listed in Table 3. As presented, the

CWDAN framework obtains the best and the most robust change detection results by using the AWV strategy and the spatial-temporal aligned constraint. For the Wan Chai dataset, these areas were more stable during the study period, and the proportion of unchanged pairs is larger. Therefore, it is possible to achieve a better change detection accuracy than using the Wuhan center dataset.

Table 4 Ablation experiments of multi-temporal urban scene classification at the parcel level

Methods	CW	AWV	Wuhan center (OA, %)			Wan Chai (OA, %)		
			T1	T2	T3	T1	T2	T3
Source_only			95.83	62.86	66.39	95.78	61.68	74.70
DANN			97.36 (+1.53)	68.50 (+5.64)	68.59 (+2.20)	95.78 (+0.00)	79.04 (+17.36)	81.33 (+6.63)
CWDAN	✓		96.53 (+0.70)	71.53 (+8.67)	71.33 (+4.94)	96.99 (+1.21)	80.24 (+18.56)	81.93 (+7.23)
(proposed)	✓	✓	96.81 (+0.98)	72.35 (+9.49)	73.39 (+7.00)	96.99 (+1.21)	80.84 (+19.16)	81.93 (+7.23)

Figure 7 shows the final change maps. The main changes of Wuhan center were concentrated on the transition between commercial, residential and vacant lands, which reveals the changes within Wuhan center brought about by urbanization. Due to the accumulation of misclassification errors, the change results for Wuhan center also include the water changes, which may not match the actual situation. The Wan Chai area of Hong Kong has remained largely unchanged. The main changes of Wan Chai focus on the water due to the construction of harbors, which have been detected correctly by CWDAN. These findings are consistent with the development status of the two cities.

4 Discussion

4.1 Methods robustness analysis

We conducted further robustness analysis on the proposed CWDAN. The CWDAN is an enhancement upon the DANN. The CW strategy and AWV strategy proposed in this paper improve the accuracy of cross-time classification and change detection, and show higher classification accuracy in the third time phase (Table 4). As illustrated in Fig. 8, the shooting angle of image in the second time phase is very different from that of source domain image, and the shooting angle of image in the

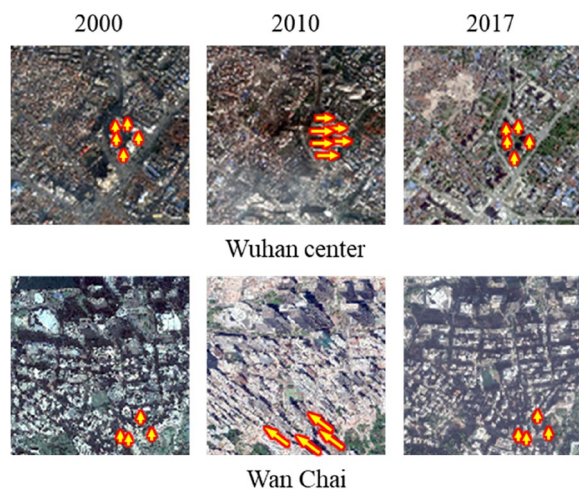
third time phase is closer to that of source domain image, so the classification effect of image in the third time phase is better. Better classification results can also lead to more reliable changes

4.2 OSM land-use product comparison

In order to verify the accuracy of the land-use mapping obtained in this study, we compared the results of the CWDAN-based land-use mapping with the OSM land-use product (Fig. 9). Since OSM only provides the historical land-use product from 2014 onwards, we only compared the 2017 mapping results for the Wuhan center dataset and WanChai dataset. From the visual representation, for the Wuhan center, the OSM land-use product only covers 20% to 30% of the study area, while our mapping results provide full coverage. In addition, a direct comparison of the areas covered by both OSM and our results shows that the results obtained by the CWDAN framework and the OSM product show a high degree of agreement. For the Wan Chai area of Hong Kong, the OSM land-use product covers a larger area. By comparing the mapping result and the land-use product, it is apparent that the main land-use categories, such as farmland, commercial, and residential, are accurately represented in our mapping results. Therefore, compared with the OSM land-use product, the CWDAN framework is able to obtain historical land-use mapping results, to compensate for the shortcomings of the OSM historical land-use product, as well as maintaining a satisfactory accuracy.

5 Conclusion

In this paper, a multi-temporal scene classification framework based on a class-weighted deep adaptation network has been proposed. This framework first uses OSM road network data and uniform grid sampling to segment the original multi-temporal HR images into the parcel level and patch level. The patch-level data are then input into the newly proposed CWDAN model, to carry out the cross-temporal scene classification task. The scene labels of the patch data are then mapped into the parcel level according to the AWV strategy. Finally, post-classification is applied to obtain the multi-temporal

**Fig. 8** Example of different imaging angles in different phases

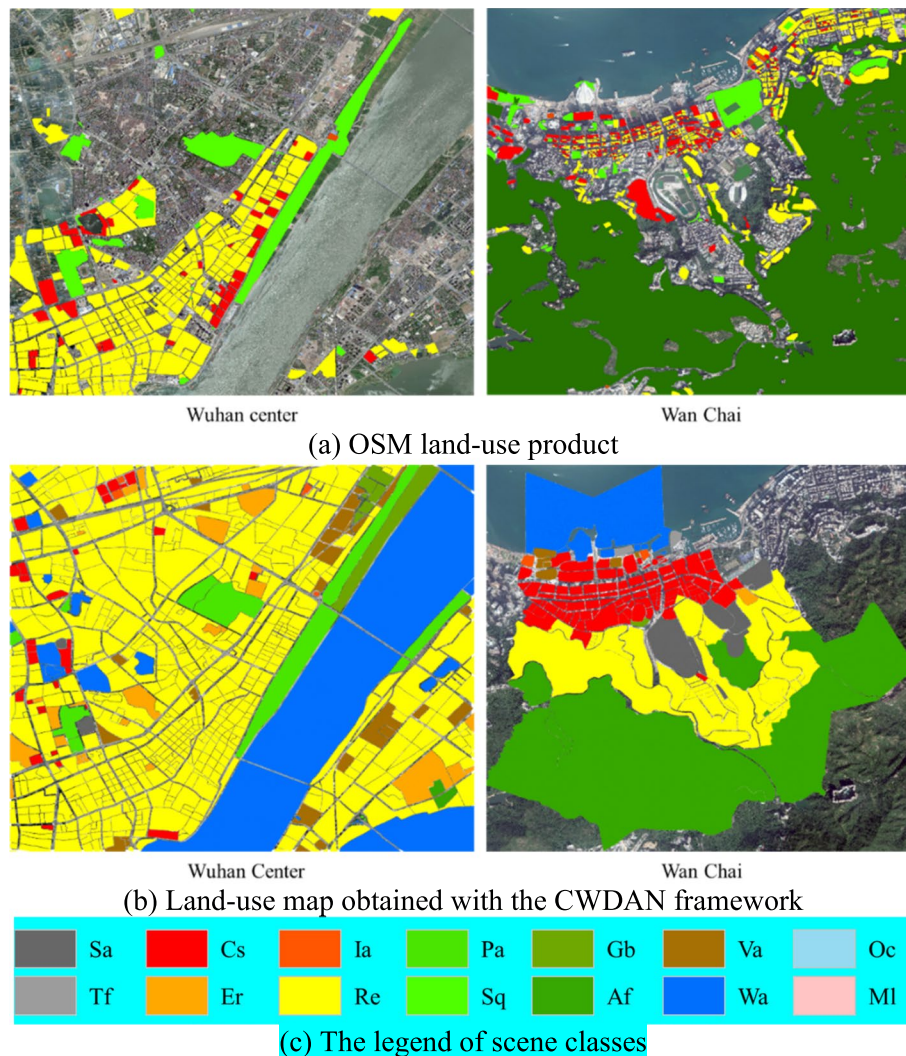


Fig. 9 Comparison of the OSM land-use product and the CWDAN-based land-use maps. **a** OSM land-use product. **b** Land-use map obtained with the CWDAN framework

change information. This work applies the idea of domain adaptation to multi-temporal scene classification, which can significantly reduce the data labeling workload. The newly introduced class-weighted fully connection layers of the network solve the problem of category imbalance within multi-temporal images to some degree. At the same time, the combination of the OSM network and uniform grid sampling not only provides meaningful geographic boundaries for the scene units, but also ensures that the data can be input into the deep learning network for efficient feature learning.

Land-use change is a visual representation of the urbanization process. Only single-phase annotation is used in this study to realize the continuation of the classification task on in long temporal sequence, which reduce

the training pressure of the scene classification task. Multi-temporal urban scene mapping can help to verify the consistency of the real urban construction and the urban planning, explain the urban development patterns, provide historical products for rational urban planning and assist the healthy and sustainable development of cities.

However, this work is still in the preliminary exploration stage of domain adaptation for multi-temporal urban scene classification and change detection. The framework proposed in this paper mainly focuses on the categories with imbalance proportions to enhance the robustness and transferability of deep features, yet the transferability in other areas of the whole study area can be further improved, and the generalization ability of

the model needs to be further enhanced. Moreover, this work only utilizes remote sensing imagery. In the future, it is possible to combine multi-modal data to strengthen the discriminative ability between easily confused categories such as residential and commercial areas. Lastly, the trend analysis of multi-temporal changes will be further researched to explore the relationship between the time span and the types and intensities of urban scene changes.

Acknowledgments

We thank very much for some members of RSIDEA for their help in this work, and the valuable comments from anonymous reviewers.

Authors' contributions

Conceptualization: YZ, LZ, DL; Writing—original draft preparation: SS; Writing—review and editing: YZ, SS, YL. All authors read and approved the final manuscript.

Funding

This work was supported by "the Fundamental Research Funds for the Central Universities" under Grant No. 2042023kfyzq04 and LIESMARS Special Research Funding.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to license reasons.

Declarations

Ethics approval consent to participate

Not applicable. All authors of this article declare they consent to participate.

Consent for publication

All authors of this article declare they consent for publication.

Competing interests

No potential conflict of interest was reported by the author(s).

Received: 10 August 2023 Revised: 25 October 2023 Accepted: 29 October 2023

Published online: 10 January 2024

References

- Audebert, N., Le Saux, B., & Lefevre, S. (2017). Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps. 1552–1560. <https://doi.org/10.1109/cvprw.2017.199>.
- Bouziani, M., Goiet, K., & He, D. C. (2010). Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge. *Isprs Journal of Photogrammetry & Remote Sensing*, 65(1), 143–153.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Computer Science*, 4, 357–361.
- Cheng, G., Yang, C., Yao, X., Guo, L., & Han, J. (2018). When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Transactions on Geoscience & Remote Sensing*, 2811–2821.
- Demir, B., Bovolo, F., & Bruzzone, L. (2013). Updating Land-Cover Maps by Classification of Image Time Series: A Novel Change-Detection-Driven Transfer Learning Approach. *IEEE Transactions on Geoscience & Remote Sensing*, 51(1), 300–312.
- Ganin, Y., & Lempitsky, V. (2015). 'Unsupervised domain adaptation by back-propagation' *International conference on machine learning*. PMLR, pp. 1180–1189.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.
- Gopalan, R., Li, R., & Chellappa, R. (2011). 'Domain adaptation for object recognition: An unsupervised approach' *IEEE International Conference on Computer Vision*.
- Grimm, N. B., Faeth, S. H., Golubiewski, N. E., Redman, C. L., Wu, J. G., Bai, X. M., et al. (2008). Global change and the ecology of cities. *SCIENCE*, 319(5864), 756–760. <https://doi.org/10.1126/science.1150195>
- Hu, F., Xia, G.-S., Hu, J., & Zhang, L. (2015). Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sensing*, 7, 14680–14707.
- Hussain, M., Chen, D., Cheng, A., Wei, H., & Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80, 91–106. <https://doi.org/10.1016/j.isprsjprs.2013.03.006>
- Jiang, J., & Zhai, C. X. (2007). 'Instance Weighting for Domain Adaptation in NLP' *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.
- Jin, Y., Wang, X., Long, M., & Wang, J. (2020). 'Minimum class confusion for versatile domain adaptation' *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, pp. 464–480.
- Johnson, R. D., & Kasischke, E. S. (2010). Change vector analysis: A technique for the multispectral monitoring of land cover and condition. *International Journal of Remote Sensing*, 19(3), 411–426. <https://doi.org/10.1080/014311698216062>
- Li, W., Wang, Z., Wang, Y., Wu, J., Wang, J., Jia, Y., et al. (2020). Classification of High-Spatial-Resolution Remote Sensing Scenes Method Using Transfer Learning and Deep Convolutional Neural Network. *IEEE Transactions on Geoscience & Remote Sensing*, 13, 1986–1995.
- Liu, L., Yang, L., & Zhu, B. (2018). Sparse Feature Space Representation: A Unified Framework for Semi-Supervised and Domain Adaptation Learning. *Knowledge-Based Systems*, S0950705118302405.
- Liu, W., Qin, R., & Su, F. (2019). Weakly supervised classification of time-series of very high resolution remote sensing images by transfer learning. *Remote Sensing Letters*, 10(7), 689–698. <https://doi.org/10.1080/2150704x.2019.1597295>
- Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). 'Learning transferable features with deep adaptation networks' *International conference on machine learning*. PMLR, pp. 97–105.
- Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). 'Deep transfer learning with joint adaptation networks' *International conference on machine learning*. PMLR, pp. 2208–2217.
- Long, M. S., Cao, Z. J., Wang, J. M., & Jordan, M. I. (2018). Conditional Adversarial Domain Adaptation. (Paper presented at the ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 (NIPS 2018))
- Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain Adaptation: Learning Bounds and Algorithms.
- Matasci, G., Volpi, M., Kanevski, M., Bruzzone, L., & Tuia, D. (2015). Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification. *IEEE Transactions on Geoscience & Remote Sensing*, 53(7), 3550–3564.
- Ni, K., & Wu, Y. (2019). Scene classification from remote sensing images using mid-level deep feature learning. *International Journal of Remote Sensing*, 41(4), 1415–1436. <https://doi.org/10.1080/01431161.2019.1667551>
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42, 145–175.
- Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2011). Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks*, 22(2), 199–210.
- Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
- Polykretis, C., Grillakis, M. G., & Alexakis, D. D. (2020). Exploring the Impact of Various Spectral Indices on Land Cover Change Detection Using Change Vector Analysis: A Case Study of Crete Island. *Greece. Remote Sensing*, 12(2), 319.

- Ru, L. X., Du, B., & Wu, C. (2021). Multi-Temporal Scene Classification and Scene Change Detection With Correlation Based Fusion. *Ieee Transactions on Image Processing*, 30, 1382–1394. <https://doi.org/10.1109/tip.2020.3039328>.
- Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. (Paper presented at the 2018 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)).
- Shang, Y. X., & Cha, K. J. (2013). The Rise of Central China Plan : Performance and Its Implications to Korea. *Chinese Studies*, 46, 335–356.
- Shi, S., Zhong, Y., Zhao, J., Lv, P., Liu, Y., Zhang, L. J. I. T. o. G., et al. (2020). Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery. 60, 1–16.
- Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298.
- Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial Discriminative Domain Adaptation. (Paper presented at the 30TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2017)).
- Wang, L., Guo, S., Huang, W., Xiong, Y., & Qiao, Y. (2016). Knowledge Guided Disambiguation for Large-Scale Scene Classification with Multi-Resolution CNNs. *IEEE Transactions on Image Processing*, 26(4), 2055–2068.
- Wang, Y., Du, B., Ru, L., Wu, C., & Luo, H. (2019). 'Scene Change Detection via Deep Convolution Canonical Correlation Analysis Neural Network' *IEEE International Geoscience and Remote Sensing Symposium*.
- Wu, C., Zhang, L., & Du, B. (2017). Kernel Slow Feature Analysis for Scene Change Detection. *IEEE Transactions on Geoscience & Remote Sensing*, 55(4), 2367–2384.
- Wu, C., Zhang, L., & Zhang, L. (2016). A scene change detection framework for multi-temporal very high resolution remote sensing images. *Signal Processing*, 124, 184–197. <https://doi.org/10.1016/j.sigpro.2015.09.020>
- Zhang, Y. C., Liu, T. L., Long, M. S., & Jordan, M. I. (2019). Bridging Theory and Algorithm for Domain Adaptation. (Paper presented at the INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 97)
- Zhao, B., Zhong, Y., & Zhang, L. (2013). Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. *Remote Sensing Letters*, 4, 1204–1213.
- Zhu, Q., Zhong, Y., Zhao, B., Xia, G. S., & Zhang, L. (2016). Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geoscience & Remote Sensing Letters*, 1–5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.