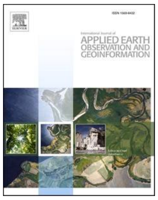
International Journal of Applied Earth Observations and Geoinformation 109 (2022) 102769

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images

Chenxiao Zhang [a], Yukang Feng [a,*], Lei Hu [a], Deodato Tapete [b], Li Pan [a], Zheheng Liang [c], Francesca Cigna [d], Peng Yue [a,e]

[a] *Wuhan University, School of Remote Sensing and Information Engineering, 129 Luoyu Road, Wuhan, Hubei 430079, China*
[b] *Italian Space Agency (ASI), Via del Politecnico snc, 00133 Rome, Italy*
[c] *South Digital Technology Co., Ltd. 4/F, Surveying Building, No.24-26 Ke Yun Road, Tian He District, Guangzhou, Guangdong 510665, China*
[d] *National Research Council (CNR), Institute of Atmospheric Sciences and Climate (ISAC), Via del Fosso del Cavaliere 100, 00133 Rome, Italy*
[e] *Wuhan University, Hubei Province Engineering Center for Intelligent Geoprocessing (HPECIG), 129 Luoyu Road, Wuhan, Hubei 430079, China*

ARTICLE INFO

Keywords:
Heterogeneous change detection
Feature alignment
Siamese network
Domain adaptation
Image fusion
Feature transformation
Satellite imagery

ABSTRACT

Heterogeneous remote sensing source-based change detection with optical and SAR data and their combined all-time and all-weather observation capability provides a reliable and promising solution for a wide range of applications. State-of-the-art supervised methods typically take a two-stage strategy that suffers from the loss of original image features and the introduction of noise on the transferred images. This paper proposes a domain adaptation-based multi-source change detection network (DA-MSCDNet) suitable to process heterogeneous optical and SAR images. DA-MSCDNet employs feature-level transformation to align inconsistent deep feature spaces in heterogeneous data. Feature space transformation and change detection are bridged within the network to encourage task communication. Experiments are conducted on two public datasets based on Sentinel-1A and Landsat-8 imagery acquired over the Sacramento, Yuba, and Sutter Counties (California, USA), and QuickBird-2 and TerraSAR-X imagery over Gloucester (UK), as well as one new large-scale dataset of Sentinel-2 and COSMO-SkyMed imagery over Wuhan (China). Compared with other six supervised and unsupervised approaches, the proposed method achieves the highest performance with an average precision of 80.81%, recall of 84.39%, mIOU of 73.67% and F1 score of 82.58%, beating the state-of-the-art method with 5.42% improvements on F1 score and 10 times efficiency on training time cost on the large-scale change detection task.

1. Introduction

Change detection (CD) in satellite images plays a key role in various applications, e.g. geohazard monitoring (Lv, Z.Y. et al., 2018), and building damage assessment (Zheng et al., 2021). Currently, most CD studies exploit a single data source, mainly optical images. However, limited by the satellite revisit time and the influence of complex weather conditions (e.g. cloudy and rainy days), CD on optical images can be challenging or, in some cases, even unfeasible. Therefore, to acquire data under the above scenarios, it is crucial to introduce active sensing systems, e.g. Synthetic Aperture Radar (SAR). SAR can depict ground surface backscattering information in all-weather conditions and proves effective and promising in CD (Cigna et al., 2013; Cigna and Tapete, 2018). However, compared with optical, SAR images lack of spectral information and might be more challenging for an image analyst to process.

To effectively complement the information of heterogeneous remote sensing sources, CD in SAR and optical images has increasingly attracted researchers' interest. Among the current studies, unsupervised methods greatly depend on hand-crafted features, and experimental parameters need to be carefully configured to maintain good results. In recent years, inspired by the deep learning-based image style transfer methods in the computer vision community, image transformation techniques have been introduced in the heterogeneous CD in a two-stage manner: an image style transfer architecture is firstly used to transform images in one domain to another, and then CD is carried out on the transformed images. Such methods explicitly transform the image style to match the other one, thus alleviating the difficulty of changed pixel discrimination. However, two major inherent drawbacks are found: the loss of raw image information and the unexpectedly introduced image noise.

* Corresponding author.
E-mail address: fengyukang2016@whu.edu.cn (Y. Feng).

https://doi.org/10.1016/j.jag.2022.102769
Received 24 January 2022; Received in revised form 18 March 2022; Accepted 1 April 2022

State-of-the-art methods apply a two-step strategy: “image style transfer – change detection on the transferred images”. Although the style transfer network can transfer the SAR images into optical images that are visibly similar to the real optical images, it is difficult to deal with the strong speckle noise existing in SAR images, information of SAR images may be discarded and external noises may be introduced in the generated optical images during the transfer process, which finally degrade the CD task performance. In this paper, we innovate by developing a domain adaptation-based method for SAR and optical image CD, capable to effectively alleviate the loss of raw image features in an end-to-end manner. Major contributions of the work are three-fold: 1) domain adaptation constraints are applied to align heterogeneous data into a common space at deep feature level rather than image level, thus alleviating the information loss; 2) deep heterogeneous feature alignment and change map reconstruction are bridged together into a unified architecture in an end-to-end manner thus avoiding the unexpected introduced noises; 3) three experimental datasets are used to test the method in two different CD scenarios: mapping changes due to flooding on one side, and urbanization and infrastructure construction on the other, thus assessing the performance over a large variety of land surface change types and potential applications.

The paper is organized as follows: State-of-the-art is summarized in Section II; Section III illustrates the proposed method; Section Ⅳ describes the experimental datasets and discusses the benchmark comparisons; main conclusions and an outlook into future research lines are provided in Section Ⅴ.

2. State-of-the-art

Pixel-based and object-based methods dominate the traditional CD methods in homogeneous remote sensing images. The former use threshold values on each pixel, to determine whether it changes or not over time. Typical methods include change vector analysis, difference value, and wavelet transformation. The latter first apply image segmentation to acquire objects in different shapes and sizes, then object-wise comparison is conducted to detect changed areas. Geometric and texture features of objects are typically used for comparison. In the past decade, deep learning-based CD methods became mainstream methods. They make use of the powerful feature extraction ability of neural networks to obtain deep image difference features of homogeneous data to generate CD results. For example, Lv, N. et al. (2018) propose a stacked self-encoder-based method to extract image features based on which a K-means clustering is conducted to produce change maps. Peng et al. (2019) propose an improved UNET++ architecture to realize CD by channel-wise stacking the bi-temporal images as a single image. Daudt et al. (2018) apply a Siamese network structure enhanced by jump connections to improve CD performance. By exploring the channel-spatial interactions among deep features through attention mechanisms, Zhang et al. (2020) propose an image fusion network for binary CD in high resolution optical satellite images. The above methods are proposed for homogeneous CD assuming the pre- and post-change images are in the same feature space. Due to their different imaging mechanisms, SAR and optical images have great differences in feature spatial distribution, thus perfectly aligned deep features are difficult to obtain directly. Therefore, direct application of the above methods on the SAR and optical remote sensing image pairs is unfeasible.

In terms of multi-source image CD researches, Mubea and Menz (2012) explore Support Vector Machine (SVM) and Maximum Likelihood (ML) for post-classification CD. Qin et al. (2013) perform Principal Component Analysis (PCA) on stacked dual-phase heterogeneous images and using the eigenvalues for image segmentation. De Giorgi et al. (2021) apply a supervised post-classification comparison method and a data fusion approach on bi-temporal COSMO-SkyMed SAR and Pléiades optical image pairs to identify land cover transitions during a post-hurricane recovery phase. Traditional image analysis methods have severe instability problems when applied for large-area CD tasks due to their limited perception field and weak pattern recognition ability.

In recent years, some researches focused on homogenization methods for heterogeneous images, which consist in the transformation of images in different spaces into the same feature space. Liu et al. (2016) propose a symmetric convolutional coupling network (SCCN) to extract the common space features in optical and SAR images, then a bi-temporal image difference map is acquired through pixel-wise Euclidean distance calculation to produce the final change map. Similarly, Liu et al. (2017) propose a heterogeneous CD method with pixel transformation, using self-organized mapping to unify the two data feature spaces, and then fuzzy clustering to detect the changed regions with pixel-level difference discrimination. By using different layers of the VGG16 (Simonyan and Zisserman, 2014) to extract style and content information, deep homogenous feature fusion is achieved with iterative image style transfer (Jiang et al., 2020), then a SVM is used for final change area detection. Niu et al. (2018) develop a conditional Generative Adversarial Network (cGAN) to convert SAR and optical satellite images into the same image type, and then conduct CD based on a difference map acquired from the transformed image. Saha et al. (2020) adopt the CycleGAN (Zhu et al., 2017) to realize bi-directional optical and SAR data transformation, then the transformed optical-like SAR features are forwarded to the depth change vector analysis for unsupervised CD in SAR image pairs. Similar to Saha et al. (2020), Li et al. (2021) also use a CycleGAN framework for image style transformation between SAR and optical images. An improved UNet++ framework is applied for the sequential supervised CD task. Inspired by the object-level comparison, a patch-based network (SiamCRNN) is proposed for supervised CD in optical satellite images and LiDAR data (Chen et al., 2019). Specifically, a deep Siamese CNN is firstly used for deep feature acquisition, then LSTM is adopted to discriminate changed pixels.

To conclude, state-of-the-art methods carry out style transformation on heterogeneous images to obtain visually similar image pairs based on which supervised or unsupervised CD is conducted in the following step. On one hand, a finely transformed image retaining its original features is hard to acquire, considering the inevitably lost information during the adversarial training process (e.g., a smooth house roof in the optical image may be transformed to a rough surface in the SAR scene, thus implying texture loss). On the other hand, noise is introduced into the transformed images during the image reconstruction process, which further increases the difficulty of the sequent CD task.

3. Method

A domain adaptation-based multi-source image CD network (DA-MSCDNet) is proposed for heterogeneous SAR and optical image CD. DA-MSCDNet consists of three parts (Fig. 1): (i) a pseudo-Siamese structure for heterogeneous image feature extraction, (ii) a domain adaptation-based feature consistency constraint block, and (iii) a multi-scale decoder for change map reconstruction.

A registered heterogeneous bi-temporal image pair (i.e., optical and SAR image) is firstly inputted into the pseudo-Siamese structure for deep feature extraction. Then the extracted features are aligned with each other enforced by the domain adaptation constraint block. During the extraction process, the difference of dual-phase features at each layer is calculated and connected to the subsequent multi-scale decoder with the skip-connection concept. The multi-scale decoder finally outputs the CD results based on the acquired feature difference maps.

3.1. Pseudo-Siamese feature extraction

A pseudo-Siamese structure implemented with ResNet34 (He et al., 2016) is applied as the feature extraction backbone of DA-MSCDNet. Networks for CD in homogeneous images usually apply Siamese structures with shared weights assuming features in bi-temporal images are in the same feature distribution. For the task of heterogeneous image CD, the distribution spaces of features extracted from of input multi-

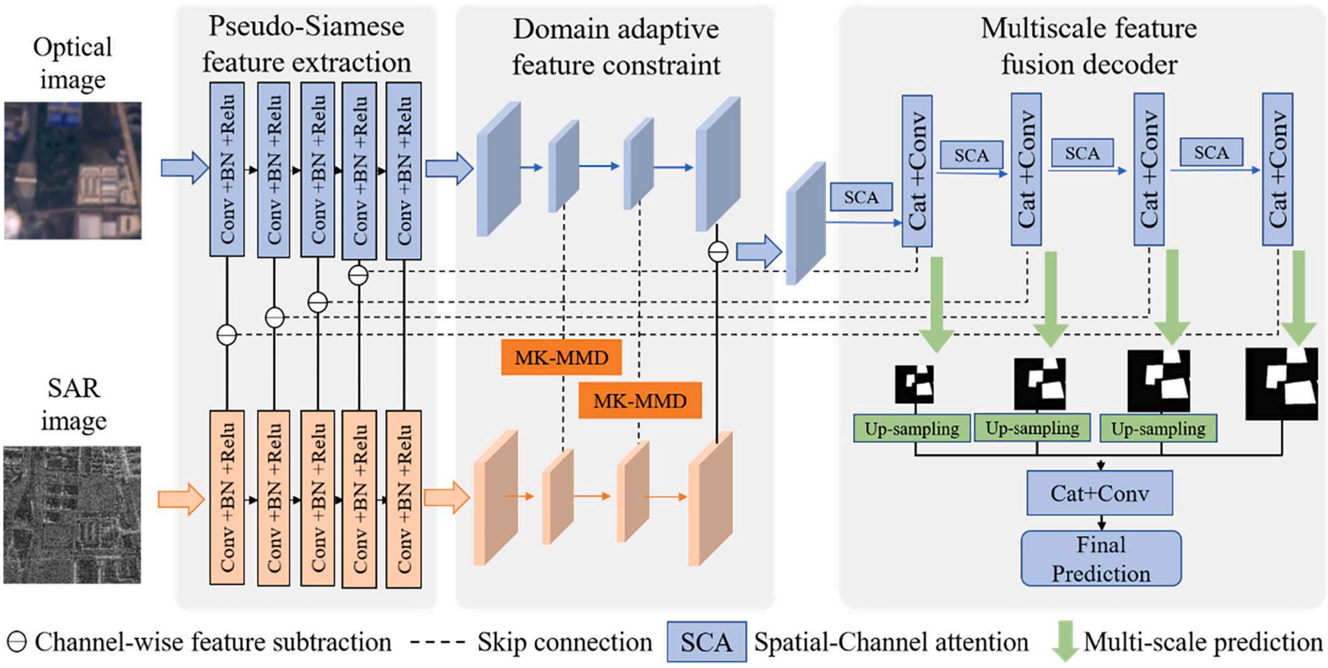

Optical image
SAR image
Pseudo-Siamese feature extraction
Conv +BN +Relu
Domain adaptive feature constraint
MK-MMD
MK-MMD
Multiscale feature fusion decoder
SCA
Cat +Conv
Up-sampling
Cat+Conv
Final Prediction
⊖ Channel-wise feature subtraction
---- Skip connection
SCA Spatial-Channel attention
Multi-scale prediction

Fig. 1. Overview of DA-MSCDNet and methodological workflow.

source data are different. Using a structure with shared weights is not suitable. Therefore, DA-MSCDNet adopts a two-branch network using a pseudo-Siamese structure with non-shared weights for raw image feature extraction. The first branch accepts an optical image as input and the second branch accepts a SAR image as input. Each branch outputs features in different sizes. High-level features of the two branches are fed into the sequential domain adaptive blocks for heterogeneous feature alignment, respectively. Low-level and mid-level features of the two branches are fed into the decoder block to help change map reconstruction by providing multi-scale raw image features. In such a way, fine-grained change maps can be acquired by retaining image features in different sizes. Sizes of input image sizes, the extracted low-, mid-, and high-level features are 256 × 256, 256 × 256, 128 × 128, 64 × 64, 32 × 32, 16 × 16, respectively.

3.2. Domain adaptation-based feature constraints

The aim of domain adaptation is to apply the knowledge learned from one or more domains to another one via mapping them into a uniform feature space. Specifically, labeled source domain and unlabeled target domain samples are integrated to train a model, thus significantly improving its performance on target domain data. Due to their respective imaging mechanisms, SAR and optical images have great differences in feature distribution space. The comparability of change information extracted from dual-phase features using the pseudo-Siamese network is relatively poor, therefore degrading the CD performances. Domain adaptation constraints can ensure that the extracted dual-phase features are in the unified feature space and can improve the results of heterogeneous image CD. Inspired by the idea that Maximum Mean Discrepancy (MMD, Gretton et al., 2012) can effectively measure the distribution distance of heterogeneous domains, we introduce domain distribution constraints between SAR and optical deep features in the CD task to encourage the bi-temporal image feature distribution alignment. Such that, features extracted from heterogeneous images are more comparable. Specifically, we propose a domain constraint block adopted by multi-layer domain adaptation (Long et al., 2015). Distribution discrepancies of SAR and optical images are computed as follows:

$$\mathrm{d}(I^{opt}, I^{sar}) = \| \frac{1}{n\_opt} \sum\nolimits_{i=1}^{n\_opt} f\left(\mathrm{I}_i^{opt}\right) - \frac{1}{n\_sar} \sum\nolimits_{j=1}^{n\_sar} f\left(\mathrm{I}_j^{sar}\right) \|_{\mathscr{H}} \tag{1}$$

where I^{opt} and I^{sar} are the heterogeneous optical and SAR image features extracted from the pseudo-Siamese convolutional network, respectively. $f(\hat{A}\cdot)$ is the feature mapping kernel function, and $\| \bullet \|_{\mathscr{H}}$ is the computation in Hilbert space. n_opt is the number of pixels in the optical image domain. I_i^{opt} is the ith pixel feature set of optical images. n_sar is the number of pixels in the SAR image domain. I_j^{sar} is the j th pixel feature set of SAR images. MMD represents distance between distributions as distances between mean embeddings of features. Therefore, if the distributions of SAR and optical image features tend to the same, MMD would approach zero.

Compared with the limited expression ability of the single fixed kernel, the multi-kernel method can greatly improve the domain consistency. Accordingly, in the domain constraint block, MK-MMD is utilized to finely measure the deep SAR and optical features. MK-MMD ($d_K(I^{opt}, I^{sar})$) is defined as follows:

$$\begin{cases} d_K(I^{opt}, I^{sar}) = \| \frac{1}{n\_pre} \sum\nolimits_{i=1}^{n\_opt} f\left(\mathrm{I}_i^{opt}\right) - \frac{1}{n\_pst} \sum\nolimits_{j=1}^{n\_sar} f\left(\mathrm{I}_j^{sar}\right) \|_{\mathscr{H}_K} \\ K = \sum\nolimits_{i=0}^{n\_k} \mu_i k_i : \sum\nolimits_{i=0}^{n\_k} \mu_i = 1, \mu_i \geq 0, \forall i \end{cases} \tag{2}$$

where $\mathscr{H}_K$ is the reproducing kernel Hilbert space (RKHS) based on multiple kernel functions K, $\| \bullet \|_{\mathscr{H}_K}$ is the distance between the two domain features in unified Hilbert space, n_k is the number of kernels, k_i is the ith kernel function and μ_i indicates the weight relation of multiple kernels in MK-MMD. K is the weighted combination of multiple kernel function k_i. The weighted summation kernel K is used to generate a Hilbert space $\| \bullet \|_{\mathscr{H}_K}$ that is similar to $\| \bullet \|_{\mathscr{H}}$ in eq. (1). We measure the distribution distances between the two datasets based on the generated Hilbert space $\| \bullet \|_{\mathscr{H}_K}$.

Enforced by the multi-layer domain constraints, the pseudo-Siamese extraction structure is regulated to produce deep features that are finely aligned with each other in each layer. The loss function of multi-layer domain constraints is defined as follows:

$$\mathrm{L}_{MK-MMD} = \frac{1}{n\_l}\sum_{i=1}^{n\_l} d_K^i(I^{opt}, I^{sar}) \tag{3}$$

where $d_K^i(I^{opt}, I^{sar})$ indicates the MK-MMD of the i th layer. n_l is the number of deep layers that need to be aligned. An average domain distance of all layers is finally calculated as the final domain consistency loss.

It should be noted that, to alleviate the heavy computation of MK-MMD, a down-sampling layer is utilized to sample large feature maps into small ones (from 16 × 16 to 4 × 4) for efficient domain adaptation in the network. Additionally, two domain constrain blocks are exploited. This is motivated by the idea that the first-round domain constraint can roughly align heterogeneous features. By exploring a new common space based on the roughly aligned feature maps for a second-round domain constraint, the domain consistency of heterogeneous features can be further enhanced.

3.3. Multi-scale decoder

After the domain adaptation constraints, aligned features are up-sampled to the size of the last feature extraction layer. To acquire image difference feature maps, we subtract the aligned optical image deep features to the SAR image deep features. In addition, position attention module (Fu et al, 2019) is introduced in the network to enhance the image difference feature maps' representation capability. The optimized difference feature map is used as the input of multi-scale decoder. It should be noted that subtracting optical features to SAR features is also applicable. To fully utilize features in different levels, image difference feature maps obtained by each feature extraction layer are also computed as follows:

$$\begin{cases} d_i = f_i^{opt} - f_i^{sar} \\ u_i = Conv^{3\times3}[up(u_{i-1}); d_i] \end{cases} \tag{4}$$

where f_i^{opt} and f_i^{sar} are the i th layer feature block of optical and SAR image branches, respectively. d_i is the feature difference map of the i th layer feature block, u_i and u_{i-1} are the i th and $i-1$ th layers after each up-sampling operation in the multi-scale decoder. $Conv^{3\times3}$ is the convolutional layer with a kernel of 3 × 3, $[;]$ is the channel-wise concatenation operation, up is the up-sampling operation.

Since the combined bi-temporal image features have large redundancies, feature maps that are highly related to the CD task need to be augmented, while those task-irrelevant feature maps need to be muted. Accordingly, a spatial channel dual-attention mechanism (SCA) is introduced for feature refinement in both channel and spatial dimension. SCA firstly computes the spatial and channel attention maps separately (Fig. 2b). ⊕ represents the element-wise summation. The two attention maps are multiplied with original features to acquire both channel and spatial-wisely refined features. SCA modifies the sequential stack of spatial attention and channel attention refinement (Fig, 2a, Woo et al., 2018) to the parallel mode, such that the two attention modules can directly refine the input feature blocks and improve the dual-attention efficiency.

SAM refines features across the spatial dimension, it performs average and maximum pooling in the spatial dimension, respectively (Fig. 2c). The extracted features are then stacked in the channel-wise, as follows:

$$M_s(F) = \sigma\left(Conv^{7\times7}([AvgPool(F); MaxPool(F)])\right) \otimes F \tag{5}$$

where $M_s(F)$ is the spatial attention refined features, $Conv^{7\times7}$ is the convolution operation with a kernel of 7 × 7, σ is the sigmoid function, ⊗ denotes the element-wise multiplication, F indicates the input features.

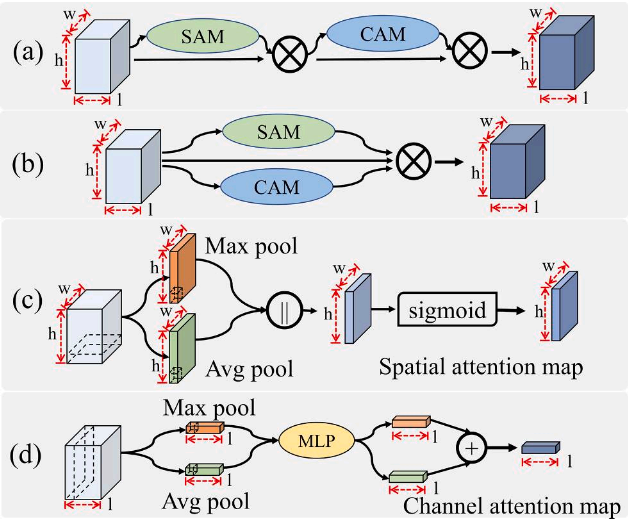


Fig. 2. Structure of (a) Convolutional block attention module (CBAM). (b) spatial channel attention (SCA) module. (c) spatial attention module (SAM). (d) channel attention module (CAM).

Global average pooling ($AvgPool$) that catches the smoothing features and global maximum pooling ($MaxPool$) that catches the sharp features are parallelly performed on the channel dimension of original features (Fig. 2d). After obtaining the one-dimensional vector, a shared fully connected network is used to reweight each feature channel. The channel attention map is produced by summing each element along the two one-dimensional vectors, as follows:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \otimes F \tag{6}$$

where $M_c(F)$ is the channel attention refined features, MLP is multi-layer perception which is a weight-shared full connection layer.

In the decoding part (Fig. 1), to fully utilize the intermediate features refined by SCA, a convolutional layer is attached to each feature block to produce a preliminary change map in each different scale. Finally, the predicted multi-scale change maps are up-sampled to the original image size for further change image fusion. The final change map is computed as follows:

$$cm = \sigma\left(Conv^{3\times3}\left(\sum_{i=1}^{n} up(Conv_i(F_i))\right)\right) \tag{7}$$

where cm indicates the final predicted change map, F_i is the i th feature block refined by SCA, n is the number of multi-scale feature blocks, $Conv_i$ is convolution operation corresponding to F_i, $Conv^{3\times3}$ is the final convolutional layer, σ is the sigmoid activation function.

3.4. Loss function

The loss function of DA-MSCDNet combines: a domain consistency loss for domain adaptation constraint (illustrated in Eq. (3)), and a binary segmentation loss of the reconstructed change maps. Binary Cross Entropy (BCE) and Dice loss (Sørensen, 1948) are used for segmentation measurements of the CD results. BCE loss is a binary classification cross-entropy loss, suitable for the binary CD task. Dice loss is a measurement of the similarity of two samples, effective in dealing with the sample imbalance problem. The losses are defined as follows:

$$\begin{cases} L_{dice} = 1 - \frac{2|P\bigcap L|}{|P| + |L|} \\ L_{bce} = -[l_i log p_i + (1 - l_i)\log(1 - p_i)] \end{cases} \tag{8}$$

where P and L are predictions and ground truth labels, respectively. ∩ indicates the intersection of the two samples, $|\bullet|$ is the per-pixel summation, p_i is the predicted pixel value and l_i is the pixel label. Total loss of DA-MSCDNet is summed together by the two sub-losses as follows:

$$\mathrm{L} = L_{dice} + L_{bce} + \lambda L_{MK-MMD} \tag{9}$$

where λ controls the weight of domain distribution constraint. λ is depended by the distribution similarity between optical and SAR images which is difficult to be manually quantified. If the two images are very similar, the term λ would approach zero. Otherwise, λ would be large. Through our extensive comparison experiments, λ with the value of 0.1 achieves the best performances on the first two small datasets. On the third dataset, the best performance is obtained with λ of 0.02.

4. Dataset and experiments

In this section, we first introduce the experimental datasets. Then we describe the results, assess the performance and prove the effectiveness of DA-MSCDNet compared with other well-established CD methods.

4.1. Datasets and evaluation criterions

4.1.1. Dataset description

Three multi-source datasets comprising pre-change optical (red–green–blue, RGB) and post-change SAR images are used:

1) The first dataset (Fig. 3) covers the Sacramento, Yuba and Sutter Counties in California, USA (Luppino et al., 2019), and includes a Landsat-8 optical image acquired in January 2017, and a Sentinel-1A VV-polarized SAR image acquired in February 2017 (https://sites.google.com/view/luppino/data). Image size and spatial resolution of the pair are 3500 × 2000 and 15 m, respectively. The ratio of changed and unchanged pixels on the dataset is 1:23. The ground-truth maps are produced based on two other Sentinel-1 images acquired during the same period.
2) The second dataset (Fig. 4) covers the city of Gloucester in UK (Mignotte, 2020), and includes: a QuickBird-2 optical image acquired in July 2006, and a TerraSAR-X StripMap HH-polarized SAR image taken in July 2007 (https://www.iro.umontreal.ca/~mignotte/ResearchMaterial/index.html#M3CD). Image size and spatial resolution of the pair are 2325 × 4135 and 0.65 m, respectively. The ratio of changed and unchanged pixels is 1:7. The ground-truth maps are created manually by experts with prior information.
3) The third dataset is collected over the urban area of Wuhan, China (Fig. 5), and includes: a Sentinel-2 optical image acquired in March 2017 with a resolution of 10 m, and a SAR amplitude product derived from an HH-polarized image acquired by the COSMO-SkyMed constellation (Caltagirone et al., 2014) in StripMap HIMAGE mode in March 2020 with a ground resolution of 3 m (https://github.com/GeoZcx/A-Domain-Adaption-Neural-Network-for-Change-Detection-with-Heterogeneous-Optical-and-SAR-Remote-Sens). The change label is obtained by manual region of interest (ROI) labeling using ENVI software. The size of the original image is 11,216 × 13,693 and pre-processing of this dataset includes radiometric and geometric correction, clipping and log transformation, to allow the statistical distribution characteristics of the SAR image to be similar to an RGB image (Zhan et al., 2018). Bilinear interpolation up-sampling is conducted on the low-resolution optical image to make its spatial resolution the same as the SAR image. The ratio of changed and unchanged pixels is 1:5.

4.1.2. Evaluation criterions

To quantitatively assess the performance of DA-MSCDNet, four metrics are defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

$$mIOU = \frac{TP}{TP + FN + FP} \tag{13}$$

where TP is short for True Positive samples corresponding to correctly predicted changed pixels, FP for False Positive samples corresponding to incorrectly predicted changed pixels, TN for True Negative samples corresponding to correctly predicted unchanged pixels, and FN for False Negative samples corresponding to incorrectly predicted unchanged pixels. Precision indicates the proportion of correctly predicted changed pixels to the overall predicted changed pixels. Recall quantifies the proportion of correctly predicted changed pixels to the actual changed pixels. F1 score provides a weighted measurement of

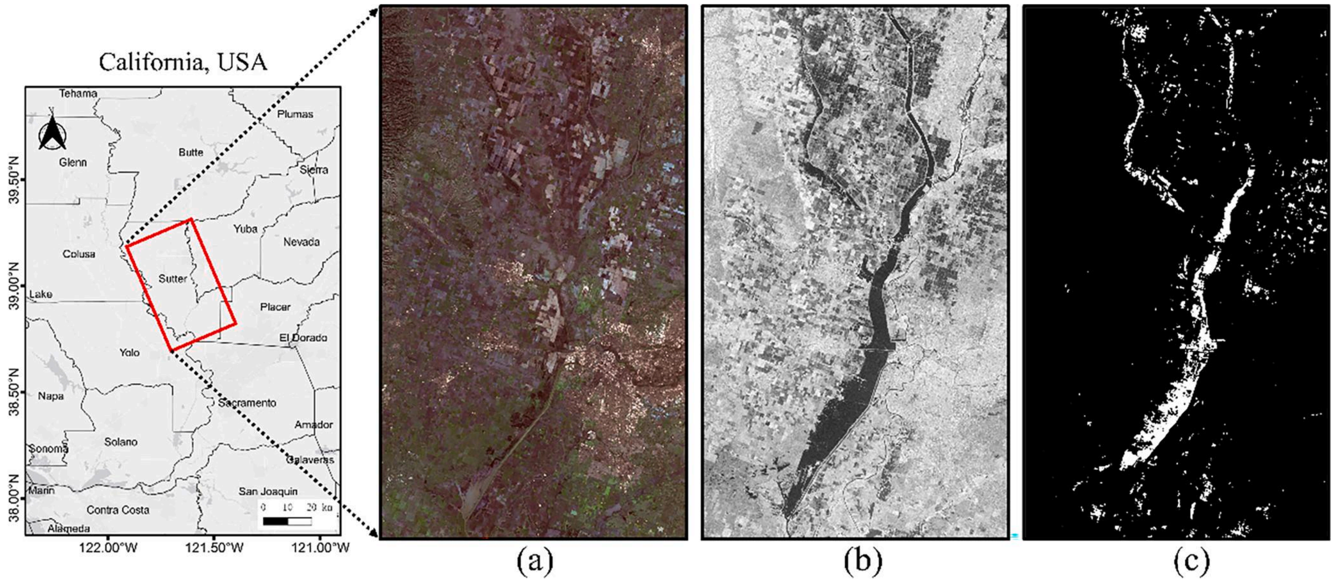


Fig. 3. California dataset: (a) Landsat-8 optical and (b) Sentinel-1A SAR images, and (c) ground-truth. Landsat-8 image courtesy of the U.S. Geological Survey. Contains Copernicus Sentinel-1 data 2017.

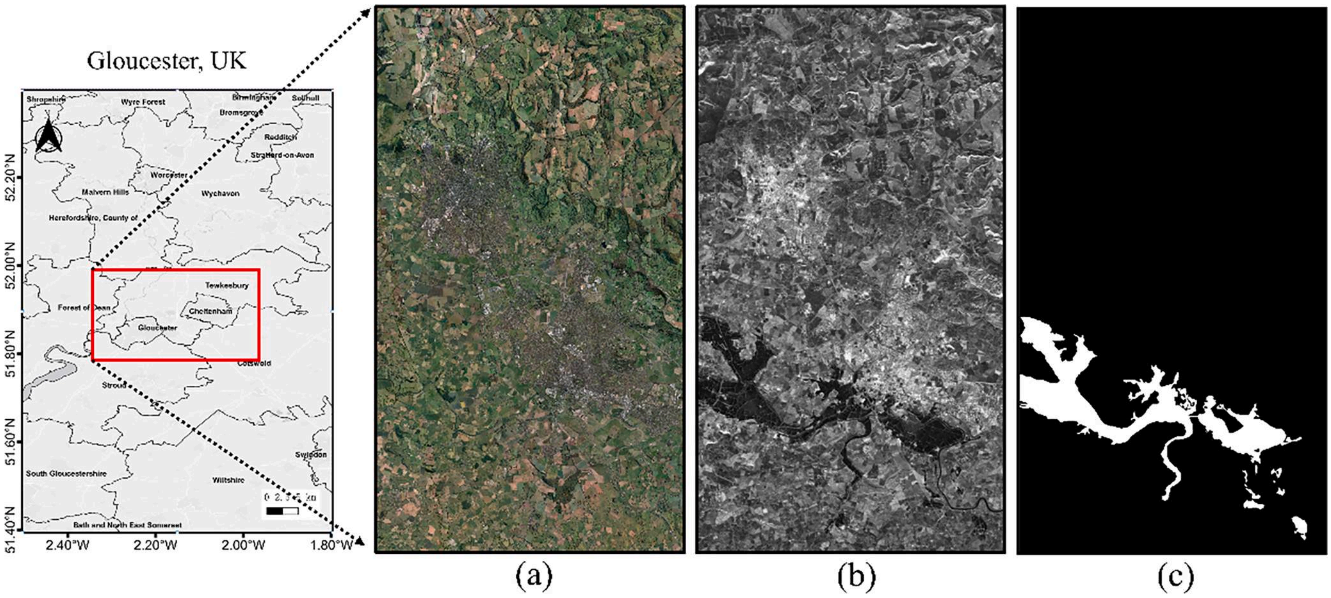


Fig. 4. Gloucester dataset: (a) QuickBird-2 optical and (b) TerraSAR-X SAR images, and (c) ground-truth. QuickBird-2 Product ©European Space Agency 2021. TerraSAR-X Product ©German Aerospace Centre. All Rights Reserved.

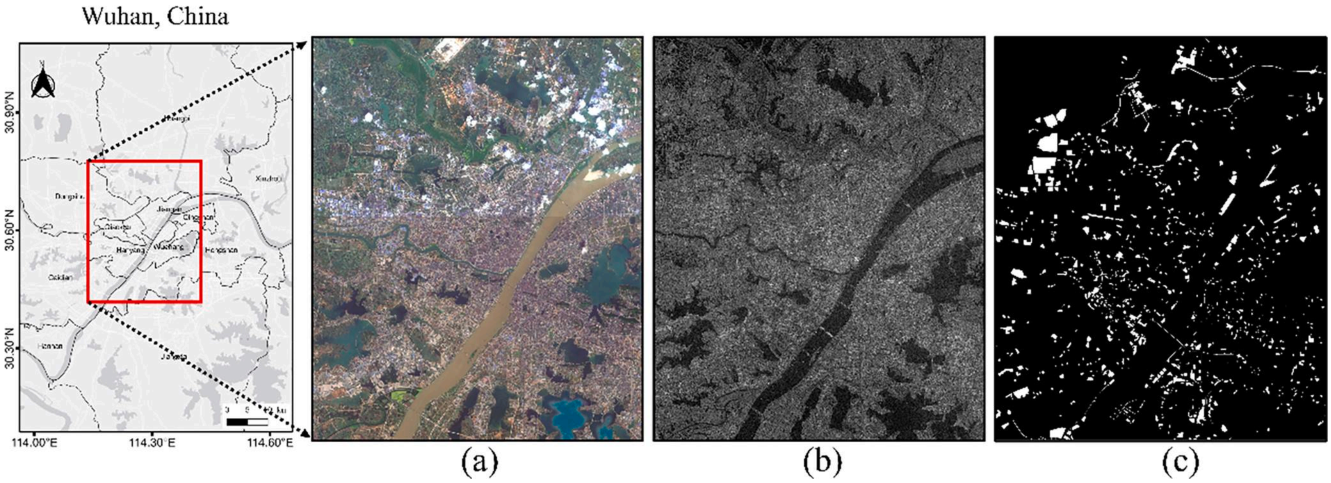


Fig. 5. Wuhan dataset: (a) Sentinel-2 optical and (b) COSMO-SkyMed SAR images, and (c) ground-truth. Contains Copernicus Sentinel-2 data 2017. Original COSMO-SkyMed® Product ©Italian Space Agency 2020. All Rights Reserved.

recall and precision, and can measure model performance in a more balanced way. Finally, mean Intersection-Over-Union (mIOU) is commonly used for semantic segmentation performance evaluation, and firstly computes the IOU for each class and then their average.

4.2. *Experimental setting*

The feature extraction part of the proposed network uses ResNet34 with non-shared weight as the backbone. To retain a high feature map size, the first down-sampling layer of ResNet34 is removed, feature maps after the first convolutional block are in size of 256 × 256. Four down-sampling operations are conducted after the following convolutional operations. After feature extraction, the highest-level feature maps are in size of 16 × 16. A 1 × 1 convolution layer is used to incorporate the two domain consistency constraint layers to calculate the domain distribution distance of per-pixel feature set. For the domain consistency constraint layer, the weighted average of multi-kernel and multi-layer domain distances is calculated, and the weight of calculated values for each layer is set to 0.5. Kernel size of convolution layers in multi-scale decoder is set to 3 × 3.

Since the three datasets have different number of images and a sufficient number of training samples is required for model training, we use different dataset split ratios regarding the three datasets. The first two small public datasets are divided into training, validation, and test datasets with a ratio of 8:2:1 considering its limited number of image pairs. The third manually created large-scale dataset is divided using a ratio of 5:1:1 considering its large available number of training samples. The input image sizes of the three datasets are set to 256 × 256 by considering both a moderate reception field and the limited GPU memory.

In this experiment, PyTorch is used for model building and training on a Tesla P40 GPU with 24 GB memory. Training epoch and batch size are set to 100 and 8, respectively. Adam optimizer is used for model training. Initial learning rate is set to 0.0005. The learning rate is reduced by 20% for each 5 training epochs. λ in Eq. (9) is set to 0.1 through extensive experimental tests. The same experimental parameters are configured in the other supervised deep learning models for a fair comparison. It should be noted that image operations including horizontal and vertical flip, image rotation are adopted for data augmentation.

4.3. Experimental results and discussion

Six deep learning-based CD methods are used to assess the performance of DA-MSCDNet through benchmark comparison: SCCN (Liu et al., 2016) and DHFF (Jiang et al., 2020) in an unsupervised manner, plus FC-EF (Daudt et al., 2018), FC-Siam-conc (Daudt et al., 2018), DTCDN (Li et al., 2021) and MSCDNet (without domain adaptation layer), in a supervised manner. It should be noted that the original DTCDN use NICE-GAN for image style transfer. But unfortunately, NiceGAN is hard to train and the transferred image quality is very bad, which can hardly be used for the subsequent CD in our experiments. Therefore, we use an alternative architecture CycleGAN with similar performance to NiceGAN in DTCDN and compare its performance with our method.

4.3.1. Visual comparison

Fig. 6 shows some sample CD results for two areas within the California dataset obtained by DA-MSCDNet and the six benchmark methods. The resulting changing areas are mainly water bodies of which shape and extent varied significantly during the flood season. As indicated in the blue rectangles in Fig. 6c-d,m-n, the results of unsupervised methods have a larger number of omitted detections across the flooded areas compared with ground truth map (Fig. 6j,t). Meanwhile, there is also a large number of falsely detected changed regions on the unflooded areas. Unsupervised methods tend to discriminate the flooded areas from the unflooded areas using only the SAR images, few optical image information is incorporated, suffering the salt and pepper noise problem. By contrast, supervised methods show better performances by effectively combine the optical and SAR image information, capturing the majority of the changed areas. FC-EF performs the worst among the supervised methods with broken object boundaries and low internal compactness on changed regions. This is because FC-EF integrates the two image channels into a single image without considering the object changes on the temporal dimension. Accordingly, by separately taking each image as network input, FC-Siam-conc gains much improvements on large changed area detection than FC-EF. DTCDN(CycleGAN) shows a high quality change map with fine object boundaries and high internal compactness (Fig. 6q). However, the change results heavialy depend on the transfered optical images. Some minor unchanged pixels (red rectangles in Fig. 6q) are miss-classified as changed pixels. MSCDNet (Fig. 6r) further improves the performance on changed regions with more accurate boundaries by taking an end-to-end manner. But it directly compares the two heterogeneous features without aligning its distributions. When dealing with unflooded regions in different colors on the optical images, MSCDNet would mistake this phenomenon of same objects with different colors as temporal changes (red rectangles in Fig. 6h). Further improving the performance by taking domain distribution constraints into the CD structure, results of DA-MSCDNet (Fig. 6i, s) show high detection accuracy on both changed and unchanged areas.

Sample CD results for the Gloucester dataset are shown in Fig. 7. Major change objects in this dataset are large-scale water bodies (e.g., ponds and reservoirs), which are less challenging to detect with respect to the small-scale change features included in the California dataset. Similar to the results acquired in the California dataset, SCCN and DHFF do not perform well in this case with most changed areas not being

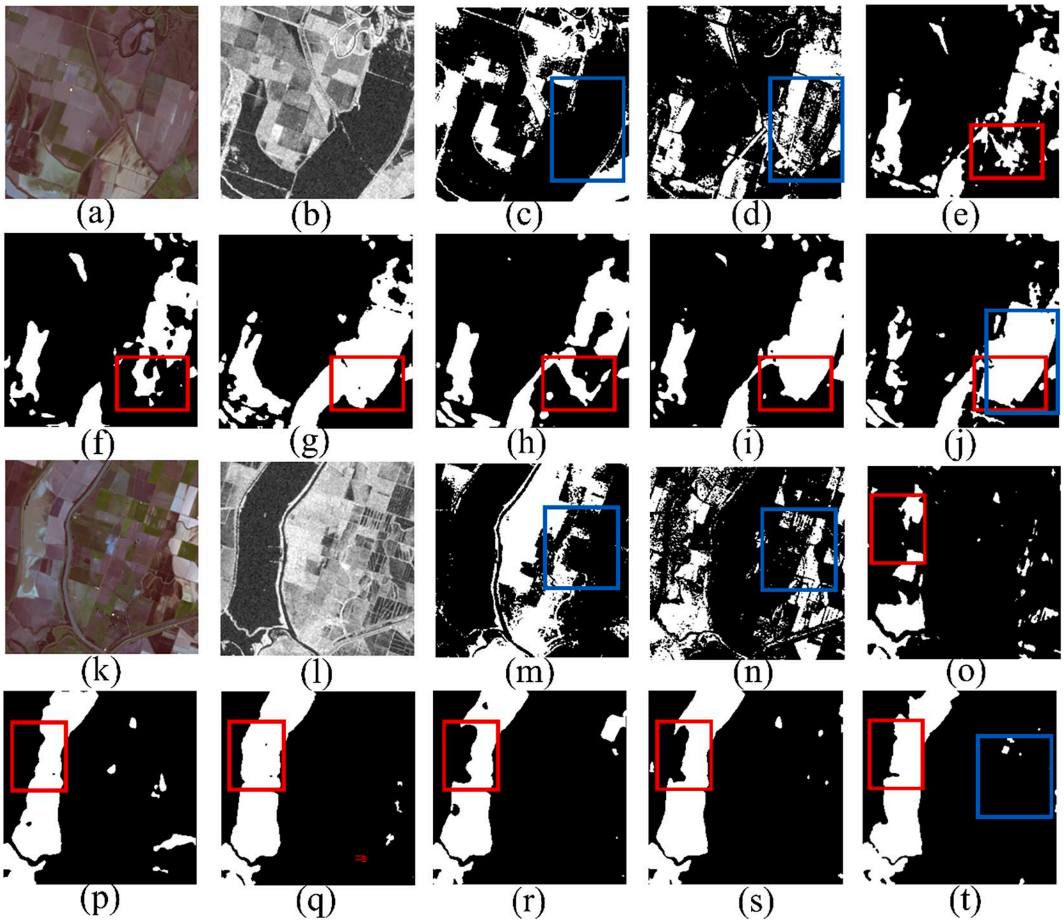


Fig. 6. Sample CD results for the California dataset: (a, k): optical image. (b, l): SAR image. (c, m): SCCN. (d, n): DHFF. (e, o): FC-EF. (f, p): FC-Siam-conc. (g, q): DTCDN(CycleGAN). (h, r): MSCDNet. (i, s): DA-MSCDNet. (j, t): ground-truth maps.

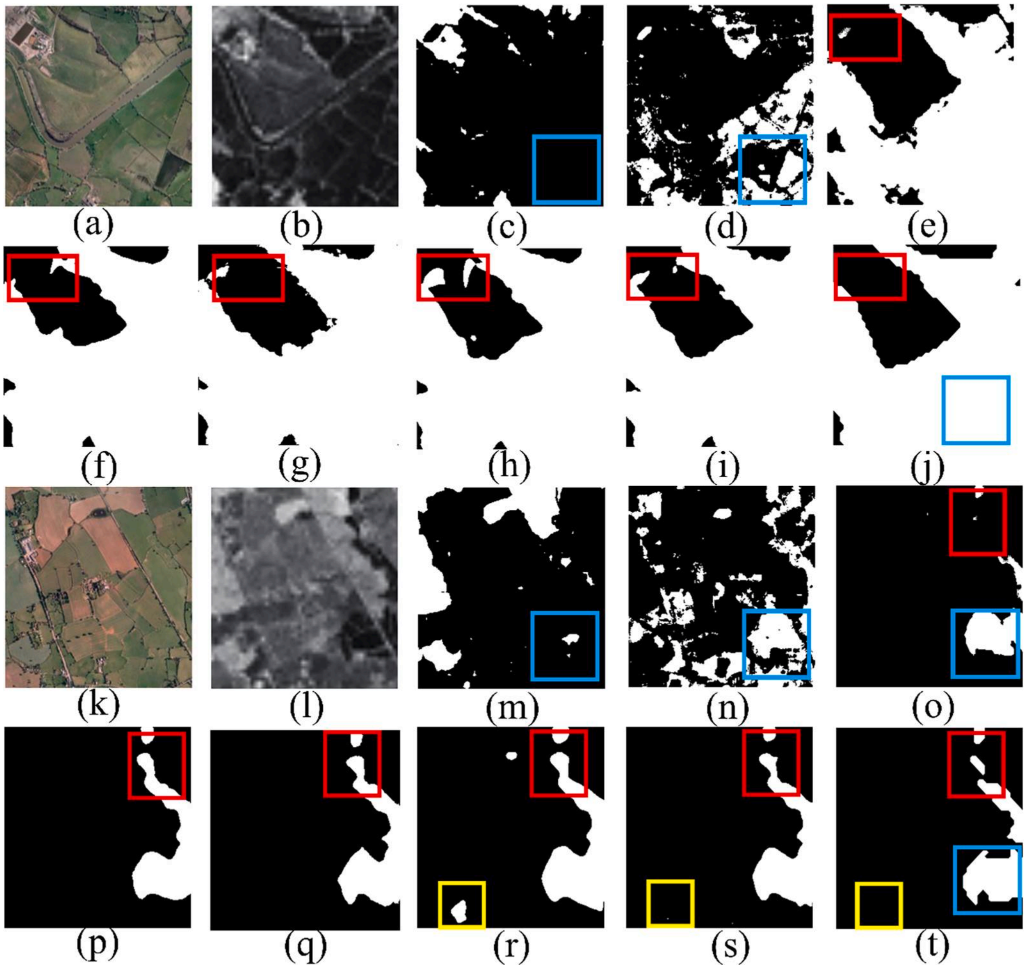


Fig. 7. Sample CD results for the Gloucester dataset: (a, k): optical image. (b, l): SAR image. (c, m): SCCN. (d, n): DHFF. (e, o): FC-EF. (f, p): FC-Siam-conc. (g, q): DTCDN(CycleGAN). (h, r): MSCDNet. (i, s): DA-MSCDNet. (j, t): ground-truth maps.

detected (Fig. 7c-d, m-n). The CD result of DHFF (blue rectangle in Fig. 7n), in particular, reveals a lot of false detection regions, which indicates its poor resilience to SAR image noises. On the other hand, FC-EF, FC-Siam-conc, DTCDN(CycleGAN), and MSCDNet provide significant improvements, with most changed areas being detected. However, all the supervised methods show a low robustness to the small unchanged areas, as shown in the red rectangles in Fig. 7e-i, o-s. Specifically, visual comparison of the two yellow squares in Fig. 7r,s reveals that DA-MSCDNet (Fig. 7s) demonstrates a lower number of false alerts over MSCDNet (Fig. 7r).

Some sample CD results obtained for the Wuhan dataset are shown in Fig. 8. In this area, CD is much more challenging, since the overall spatial range of this dataset is larger, and the changed ground objects are more complex than those within the other two datasets. Due to its weak global reception ability on large changed regions, SCCN and DHFF have difficulties in achieving good recall and high precision in detecting changed pixels (Fig. 8c-d, m-n). The unsupervised methods suffer severe salt-and-pepper noise problem resulting in a lot of false alerts. Supervised methods (i.e., FC-EF and FC-Siam-conc) significantly improve the CD accuracy in a wide range of change areas (Fig. 8e-f, o-p). However, the broken boundary of change areas and the large areas of undetected changed pixels demonstrate its insufficient exploration of temporal change relations between bi-temporal images. Though the recall of large changed areas is increased, DTCDN(CycleGAN) can not transfer SAR images into high quality optical images, resulting in change maps with irregular shapes. Comparatively, MSCDNet and DA-MSCDNet (Fig. 8h-i, r-s) both produce highly accurate change maps with continuous and smooth object boundary and high internal completeness of large changed areas. DA-MSCDNet shows higher resilience to false alerts than MSCDNet on some building areas.

4.3.2. *Quantitative comparison*

Training hyperparameters have direct impacts on the performance of supervise methods. To fully evaluate the robustness of the proposed method and the other four supervised methods, experiments with different hyperparameter configurations are carried out and their performance scores are averaged for quantitative comparison. Specifically, batch size of 8, 10 and 12, learning rate of 0.0005, 0.0002 and 0.0003, and network parameter initialization of xavier_normal and kaiming are tested in the experiments.

Quantitative assessment of CD performance of the California dataset is shown in Table 1. DHFF achieves an F1 score of 44.41% and mIOU of 54.50%. The poorest performance is achieved by SCCN with an F1 score of only 46.31% and mIOU of 38.84%. As visually compared in the previous section, supervised methods have much higher accuracy than unsupervised ones. FC-EF achieves an F1 score of 76.68% and mIOU of 66.89%. Showing very similar performances, FC-Siam-conc achieves an F1 score of 78.26% and mIOU 68.64% which are slightly better than FC-EF, which can validate the applicability of Siamese network in dealing with bi-temporal image change tasks. For the state-of-the-art supervised method, DTCDN(CycleGAN) achieves an F1 score of 80.67% and mIOU of 71.28%, which indicates the effectiveness of the image style transfer process. Instead of pursuing image style similarity, MSCDNet applies multi-scale decoding for fine-grain change map reconstruction and

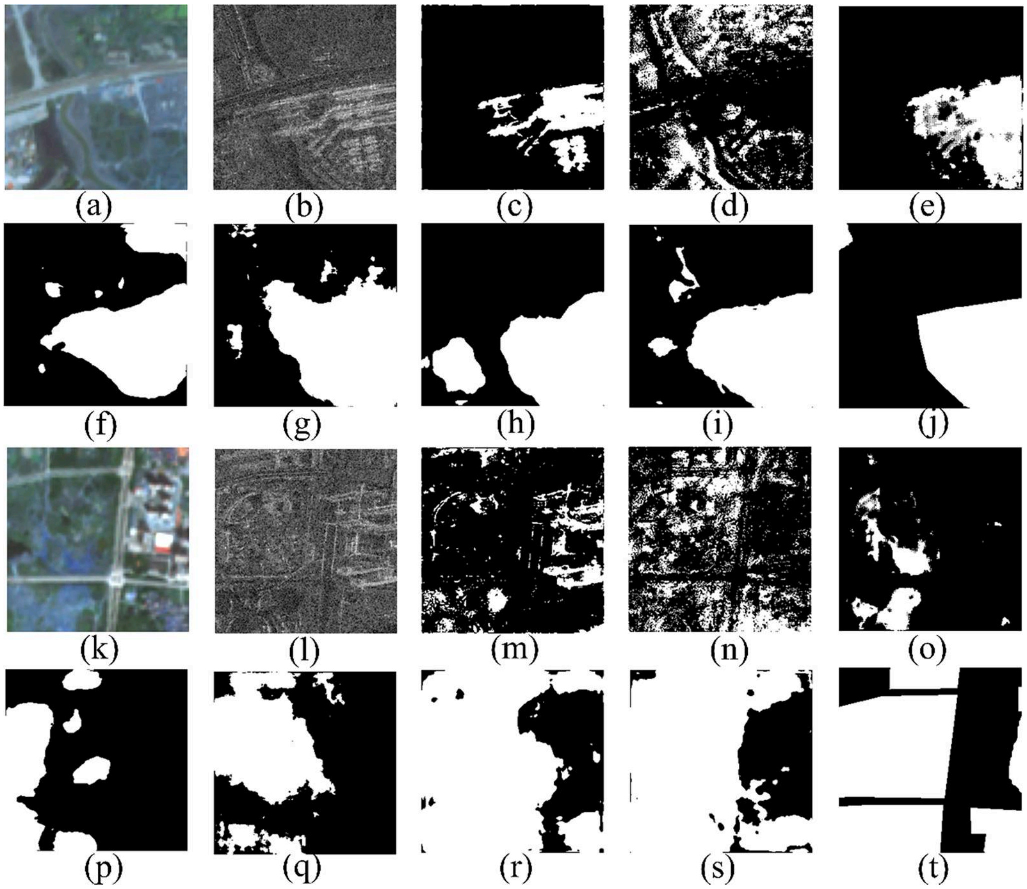


Fig. 8. Sample CD results for the Wuhan dataset: (a, k): optical image. (b, l): SAR image. (c, m): SCCN. (d, n): DHFF. (e, o): FC-EF. (f, p): FC-Siam-conc. (g, q): DTCDN (CycleGAN). (h, r): MSCDNet. (i, s): DA-MSCDNet. (j, t): ground-truth maps.

Table 1
Quantitative assessment results on the California dataset.

Method	Precision (%)	Recall (%)	mIOU (%)	F1 (%)
SCCN	49.92	49.69	38.84	46.31
DHFF	55.75	73.45	44.41	54.50
FC-EF	73.16	81.92	66.89	76.68
FC-Siam-conc	77.21	80.02	68.64	78.26
DTCDN(CycleGAN)	80.29	82.06	71.28	80.67
MSCDNet	79.89	84.22	72.25	81.54
DA-MSCDNet	78.89	85.38	72.74	82.17

further increases the CD performance with improvement on F1 score of 0.87% and mIOU of 0.97%. Benefiting from the domain distribution constraints, DA-MSCDNet further improves the performance over MSCDNet by 0.63% and 0.49% for F1 score and mIOU, respectively.

Table 2 summarizes the quantitative CD results for the Gloucester dataset. SCCN achieves the lowest performance. Differently from the results in the California dataset, DHFF performs much better than SCCN with improvements of 6.16% on F1 score. FC-EF, FC-Siam-conc and MSCDNet show much improved values for all metrics and their gaps are very small, which is a similar finding to the results in the California dataset. Interestingly, DTCDN(CycleGAN) achieves an F1 score of 91.57%, mIOU of 85.32%, which is slightly lower than FC-Siam-conc. Because the spatial resolution of the Gloucester is much higher (i.e., 0.65 m) than the California dataset (i.e., 15 m), the difficulty of CycleGAN to produce high-quality high-resolution optical images is much higher. The poorly transferred optical images on the Gloucester dataset further affect its CD performance. Comparatively, DA-MSCDNet achieves the best performance with the highest F1 score of 93.86%, mIOU of 88.88%, recall of 95.88%, and precision of 92.04%.

Table 3 shows the results for the Wuhan dataset. SCCN achieves an F1 score of 50.01% and mIOU of 40.57%, which is close to DHFF (F1 of 47.62%, mIOU of 36.53%). The typical two supervised methods perform much better than the unsupervised ones: FC-EF (F1 of 57.71% and mIOU of 48.65%), FC-Siam-conc (F1 of 57.74% and mIOU of 48.85%). State-of-the-art method DTCDN(CycleGAN) outperforms FC-Siam-conc on this moderate spatial resolution dataset with significant improvements on F1 score of 8.55% and mIOU of 6.07%. The end-to-end structure

Table 2
Quantitative assessment results on the Gloucester dataset.

Method	Precision (%)	Recall (%)	mIOU (%)	F1 (%)
SCCN	47.13	43.91	38.98	45.02
DHFF	52.94	58.17	41.37	51.18
FC-EF	89.08	91.01	82.35	89.52
FC-Siam-conc	91.99	94.54	87.17	92.75
DTCDN(CycleGAN)	88.83	95.17	85.32	91.57
MSCDNet	91.60	93.31	86.24	92.17
DA-MSCDNet	92.04	95.88	88.88	93.86

Table 3
Quantitative assessment results on the Wuhan dataset.

Method	Precision (%)	Recall (%)	mIOU (%)	F1 (%)
SCCN	50.14	50.11	40.57	50.01
DHFF	48.21	47.07	36.53	47.62
FC-EF	61.6	56.59	48.65	57.71
FC-Siam-conc	61.59	58.24	48.85	57.74
DTCDN(CycleGAN)	67.42	65.36	54.92	66.29
MSCDNet	68.67	74.24	57.43	70.01
DA-MSCDNet	71.51	71.91	59.91	71.71

MSCDNet also beats DTCDN(CycleGAN) by improving F1 score with 3.72% and mIOU with 2.51%. The best performance is achieved by DA-MSCDNet with respect to the other six methods in terms of precision, mIOU, and F1. However, the recall of DA-MSCDNet is slightly lower than MSCDNet which is mainly caused by a lower recall on the changed pixels.

4.3.3. Time efficiency comparison

To fairly compare the time efficiencies of DA-MSCDNet and the other five supervised methods (FC-EF, FC-Siam-conc, DTCDN(CycleGAN), MSCDNet and DA-MSCDNet), training and predicting time costs on a single training epoch are provided in Table 4. The experiments are implemented with PyTorch and carried out on a single GPU (NVIDIA Tesla P40 with 24 GB RAM). Training batch size is set to 8 and the total training image is set to 200. As shown in Table 4, training time cost on a single training epoch of FC-EF, FC-Siam-conc, DTCDN(CycleGAN), MSCDNet, and DA-MSCDNet are 5.98, 7.12, 10.98, 35.82 and 26.19 s, respectively. The required predicting time on a single optical and SAR image pair with the size of 256 × 256 are 0.065, 0.037, 0.055, 0.061 and 0.058 s, respectively. It should be noted that DTCDN(CycleGAN) takes a two-stage manner, transferring SAR into optical images is required to be finished before the supervised CD task. It takes CycleGAN 266.15 s to train an epoch in the first stage. Therefore, the total training time cost of the state-of-the-art method is 277.13 s. Besides, transferring SAR into optical images is also required during change map predicting, which takes DTCDN(CycleGAN) about 0.218 s. Comparatively, DA-MSCDNet shows significantly improved training and predicting time efficiency, by benefiting from its end-to-end structure.

4.3.4. Discussion

Through the experiments and a comprehensive benchmark comparison (both visual and quantitative), we can draw the following observations:

(i) Unsupervised methods have generally low detection accuracy for some regions that can be easily confused as changes. For instance, this is observed in areas of the California dataset where unsupervised methods get confused in dense agricultural plots, and they cannot separate well the signals from flooded and unflooded vegetation/bare soil. For the two-stage methods SCCN and cGAN, the image transformation stage impacts on the subsequent CD task, which brings difficulties in obtaining good generalization and high precision in detecting changed areas. Moreover, the traditional classification and segmentation methods used in the second stage further degrade the detection performance compared to the supervised methods.

(ii) For supervised methods, FC-EF, FC-Siam-conc, and MSCDNet can effectively deal with the detection of large changed areas. However, producing robust change maps with high boundary continuity and internal completeness is still challenging, which is mainly due to the mismatch of deep feature spaces in the heterogeneous images. By introducing domain consistent constraints into the CD network, the proposed DA-MSCDNet can effectively explore the common-space distributed SAR and optical feature set with the guidance of CD task, thus achieving enhanced CD results.

5. Conclusion

In this paper, we propose a domain adaptation neural network for change detection in heterogeneous optical and SAR remote sensing images, called DA-MSCDNet. Features among heterogeneous images are firstly extracted through a pseudo-Siamese structure with non-shared weights. Then a domain adaptation constraint is imposed on the extracted heterogeneous deep features to align them into a common deep feature space. Aligned deep features are fed into a multi-scale decoder to produce the final change map. A comprehensive experimental comparison shows that the proposed DA-MSCDNet achieves the best performance over the other six established unsupervised and supervised methods on two public datasets and a new large-scale dataset that, as a whole, provided a well assorted sample of satellite optical and SAR input data, change patterns, types and spatial scales to robustly test the proposed method. The significant improvements in precision, recall, mIOU and F1 score of DA-MSCDNet on the three datasets prove the enhanced performance of the proposed method. Given that the tests were conducted in two different scenarios, i.e. changes due to flooding on one side and urbanization and infrastructure construction on the other, DA-MSCDNet demonstrates promising effectiveness for different CD applications that are characterized by a different range of land surface change types.

Table 4
Training and predicting time cost comparison.

Method	Training(s/epoch)	Predicting (s/pair)
FC-EF	5.98	0.065
FC-Siam-conc	7.12	0.037
DTCDN(CycleGAN)	266.15 + 10.98	0.218 + 0.055
MSCDNet	35.82	0.061
DA-MSCDNet	26.19	0.058

Since the image resolution of multi-source heterogeneous image data is often different (depending on satellite sensor viewing and acquisition modes), and the current CD methods require the input data of the same size for pixel-level image segmentation, the future research priority will be to develop approaches capable of efficiently handling input data with different resolutions. This priority is particularly relevant in light of the enhanced image acquisition capabilities of current and upcoming satellite missions, and the increasing demand for very high-resolution imaging for land and environmental applications.

In addition, semantic CD proved promising for detailing the specific change trends of ground objects, meaning that the results of heterogeneous image CD will not be limited to categories of change and nonchange, hence providing enhanced insights into changing urban and rural landscapes. Therefore, future efforts will be put on this research aspect.

CRediT authorship contribution statement

Chenxiao Zhang: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Funding acquisition. **Yukang Feng:** Methodology, Software, Validation, Investigation, Resources, Writing – original draft. **Lei Hu:** Validation, Investigation, Visualization, Writing – review & editing. **Deodato Tapete:** Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Li Pan:** Data curation. **Zheheng Liang:** Resources. **Francesca Cigna:** Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Peng Yue:** Conceptualization, Supervision, Formal analysis, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was supported by China National Postdoctoral Program for Innovative Talents (No. BX2021223), China Postdoctoral Science Foundation (No. 2021 M702510), Hubei Provincial Natural Science Foundation of China (No. 2020CFA001). Input data include a derived product processed by D. Tapete from an original COSMO-SkyMed®

Product, © Italian Space Agency (ASI), delivered under a license to use by ASI.

References

Caltagirone, F., Capuzi, A., Coletta, A., De Luca, G.F., Scorzafava, E., Leonardi, R., Rivola, S., Fagioli, S., Angino, G., LAbbate, M., Piemontese, M., Zampolini Faustini, E., Torre, A., De Libero, C., Esposito, P.G., 2014. The COSMO-SkyMed dual use earth observation program: Development, qualification, and results of the commissioning of the overall constellation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (7), 2754–2762.

Chen, H., Wu, C., Du, B., Zhang, L., Wang, L., 2019. Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network. IEEE Trans. Geosci. Remote Sens. 58, 2848–2864. https://doi.org/10.1109/TGRS.2019.2956756.

Cigna, F., Tapete, D., 2018. Tracking human-induced landscape disturbance at the nasca lines UNESCO world heritage site in Peru with COSMO-SkyMed InSAR. Remote Sensing 10, 572. https://doi.org/10.3390/rs10040572.

Cigna, F., Tapete, D., Lasaponara, R., Masini, N., 2013. Amplitude change detection with ENVISAT ASAR to image the cultural landscape of the Nasca region, Peru. Archaeological Prospection 20, 117–131. https://doi.org/10.1002/arp.1451.

Daudt, R.C., Le Saux, B., Boulch, A., 2018. October. Fully convolutional siamese networks for change detection. In: In 2018 25th IEEE International Conference on Image Processing, pp. 4063–4067. https://doi.org/10.1109/ICIP.2018.8451652.

De Giorgi, A., Solarna, D., Moser, G., Tapete, D., Cigna, F., Boni, G., Rudari, R., Serpico, S.B., Pisani, A.R., Montuori, A., Zoffoli, S., 2021. Monitoring the Recovery after 2016 Hurricane Matthew in Haiti via Markovian Multitemporal Region-Based Modeling. Remote Sensing 13, 3509. https://doi.org/10.3390/rs13173509.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146–3154.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., Sriperumbudur, B.K., 2012. Optimal kernel choice for large-scale two-sample tests. In Advances in neural information processing systems 1205–1213.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: In Proceedings of the *IEEE conference on computer vision and pattern recognition*, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Jiang, X., Li, G., Liu, Y., Zhang, X.P., He, Y., 2020. Change detection in heterogeneous optical and SAR remote sensing images via deep homogeneous feature fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1551–1566. https://doi.org/10.1109/JSTARS.2020.2983993.

Li, X., Du, Z., Huang, Y., Tan, Z., 2021. A deep translation (GAN) based change detection network for optical and SAR remote sensing images. ISPRS J. Photogramm. Remote Sens. 179, 14–34. https://doi.org/10.1016/j.isprsjprs.2021.07.007.

Liu, J., Gong, M., Qin, K., Zhang, P., 2016. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans. Neural Networks Learn. Syst. 29, 545–559. https://doi.org/10.1109/TNNLS.2016.2636227.

Liu, Z.G., Zhang, L., Li, G., He, Y., 2017. July. Change detection in heterogeneous remote sensing images based on the fusion of pixel transformation. In: In 2017 20th International Conference on Information Fusion, pp. 1–6. https://doi.org/10.23919/ICIF.2017.8009656.

Long, M., Cao, Y., Wang, J. and Jordan, M., 2015, June. Learning transferable features with deep adaptation networks. *In International conference on machine learning*, pp. 97-105. Available: https://arxiv.org/abs/1502.02791.

Luppino, L.T., Bianchi, F.M., Moser, G., Anfinsen, S.N., 2019. Unsupervised image regression for heterogeneous change detection. IEEE Trans. Geosci. Remote Sens. 57, 9960–9975. https://doi.org/10.1109/TGRS.2019.2930348.

Lv, N., Chen, C., Qiu, T., Sangaiah, A.K., 2018a. Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images. IEEE Trans. Ind. Inf. 14, 5530–5538. https://doi.org/10.1109/TII.2018.2873492.

Lv, Z.Y., Shi, W., Zhang, X., Benediktsson, J.A., 2018b. Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (5), 1520–1532.

Mignotte, M., 2020. A fractal projection and Markovian segmentation-based approach for multimodal change detection. IEEE Trans. Geosci. Remote Sens. 58 (11), 8046–8058.

Mubea, K., Menz, G., 2012. Monitoring Land-Use Change in Nakuru (Kenya) Using Multi-Sensor Satellite Data. ARS 01 (03), 74–84.

Niu, X., Gong, M., Zhan, T., Yang, Y., 2018. A conditional adversarial network for change detection in heterogeneous images. IEEE Geosci. Remote Sens. Lett. 16 (1), 45–49.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing 11, 1382. https://doi.org/10.3390/rs11111382.

Qin, Y., Niu, Z., Chen, F., Li, B., Ban, Y., 2013. Object-based land cover change detection for cross-sensor images. Int. J. Remote Sens. 34, 6723–6737. https://doi.org/10.1080/01431161.2013.805282.

Saha, S., Bovolo, F., Bruzzone, L., 2020. Building change detection in VHR SAR images via unsupervised deep transcoding. IEEE Trans. Geosci. Remote Sens. 59, 1917–1929. https://doi.org/10.1109/TGRS.2020.3000296.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab 5, 1–34.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. Available In Proceedings of the European conference on computer vision 3–19. https://arxiv.org/abs/1807.06521.

Zhan, T., Gong, M., Jiang, X., Li, S., 2018. Log-based transformation feature learning for change detection in heterogeneous images. IEEE Geosci. Remote Sens. Lett. 15, 1352–1356. https://doi.org/10.1109/LGRS.2018.2843385.

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J. Photogramm. Remote Sens. 166, 183–200. https://doi.org/10.1016/j.isprsjprs.2020.06.003.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., Zhang, L., 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. Remote Sens. Environ. 265, 112636.

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *In Proceedings of the IEEE international conference on computer vision*, pp. 2223-2232, Available: https://arxiv.org/abs/1703.10593.