ISPRS Journal of Photogrammetry and Remote Sensing 179 (2021) 14–34

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A deep translation (GAN) based change detection network for optical and SAR remote sensing images

Xinghua Li [a], Zhengshun Du [a], Yanyuan Huang [a], Zhenyu Tan [b,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] College of Urban and Environmental Sciences, Northwest University, Xi'an 710127, China

ARTICLE INFO

Keywords:
Change detection
Deep translation
Depthwise separable convolution
GAN
Multi-scale loss
Optical and SAR images

ABSTRACT

With the development of space-based imaging technology, a larger and larger number of images with different modalities and resolutions are available. The optical images reflect the abundant spectral information and geometric shape of ground objects, whose qualities are degraded easily in poor atmospheric conditions. Although synthetic aperture radar (SAR) images cannot provide the spectral features of the region of interest (ROI), they can capture all-weather and all-time polarization information. In nature, optical and SAR images encapsulate lots of complementary information, which is of great significance for change detection (CD) in poor weather situations. However, due to the difference in imaging mechanisms of optical and SAR images, it is difficult to conduct their CD directly using the traditional difference or ratio algorithms. Most recent CD methods bring image translation to reduce their difference, but the results are obtained by ordinary algebraic methods and threshold segmentation with limited accuracy. Towards this end, this work proposes a deep translation based change detection network (DTCDN) for optical and SAR images. The deep translation firstly maps images from one domain (e.g., optical) to another domain (e.g., SAR) through a cyclic structure into the same feature space. With the similar characteristics after deep translation, they become comparable. Different from most previous researches, the translation results are imported to a supervised CD network that utilizes deep context features to separate the unchanged pixels and changed pixels. In the experiments, the proposed DTCDN was tested on four representative data sets from Gloucester, California, and Shuguang village. Compared with state-of-the-art methods, the effectiveness and robustness of the proposed method were confirmed.

1. Introduction

Remote sensing change detection (CD) is the process of identifying the object or phenomenon difference by multi-temporal images over the same geographical area (Ashbindu 1989). CD is a very popular and important topic in the field of remote sensing, and it plays a key role in practical applications such as disaster assessment (Ji et al. 2019; Sublime and Kalinicheva 2019), resource surveys (Khan et al. 2017; Lunetta et al. 2006) and urban planning (Jaturapitpornchai et al. 2019; Lyu et al. 2018; Nguyen and Han 2020).

Most of the attention has been paid to remote sensing CD from single data source, especially the optical remote sensing image. The optical images with various spatial and temporal resolutions can not only provide spectral information of ground features, but also reflect their texture and geometric shape, thereby ensuring the possibility and accuracy of CD. For example, Tian et al. (2014) realized the automatic CD of buildings through Kullback-Leibler divergence similarity measure using very high resolution (VHR) optical images. Additionally, an artificial neural network is proposed for extracting the characteristics of urban sprawl from Landsat TM images (Tong et al. 2010).

Recently, more and more kinds of data sources have been applied to CD. Synthetic aperture radar (SAR) images reflect the scattered information of the ground surface in all weather and all time. Therefore, the independence of SAR images from lighting conditions makes them advantageous for specific CD tasks (Geng et al. 2019). Saha et al. (2018) exploited the potential of deep encoding between VHR SAR images to identify post-earthquake destroyed buildings. Besides, the log cumulants and stacked autoencoder were used for fire damage assessment based on the Sentinel-1 and TerraSAR-X images (Planinšič and Gleich 2018). These studies have achieved effective CD of SAR images.

Although SAR images are better than optical images in poor atmospheric conditions, they lack rich spectral information and are affected

* Corresponding author.
E-mail address: tanzhenyu@nwu.edu.cn (Z. Tan).

https://doi.org/10.1016/j.isprsjprs.2021.07.007
Received 13 May 2021; Received in revised form 4 July 2021; Accepted 13 July 2021

by speckle noises. Optical and SAR images based CD become popular because they can provide complementary information. The combination of optical and SAR remote sensing images can timely and accurately reflect the situation before and after the event. Particularly, after a disaster, such as a flood or fire, affected by water vapor or smoke, the optical imaging cannot obtain clear information of the ground surface, while SAR can help us to respond rapidly to these situations. Optical and SAR remote sensing images CD is a difficult challenge because the imaging mechanisms, radiation characteristics, and geometric characteristics are inconsistent. To overcome these limitations, some CD models are proposed and can be divided into five categories:

(1) Parametric methods. The parametric methods usually utilize a mixture of *meta*-Gaussian or multivariate distributions to estimate the relationship between different sensor images. A relevant change indicator is extracted by solving these parameters. The representative methods include Kullback-Leibler-based change measures (Mercier et al. 2008), local joint distributions (Prendes et al. 2015), and Markov model for multimodal change detection (M3CD) (Touati et al. 2019). Based on the theory of copula, these methods use local statistics to extract accurate change maps from different images. However, they usually rely on a specific distribution to describe the relationship between a pair of remote sensing images. It is not easily generalizable when facing the rapid development of multimodal sensors.
(2) Non-parametric methods. Different from the parametric algorithms, this kind of method does not require the estimation of important parameters and can be used for any pairs of heterogeneous sensors. For example, the least-squares energy-based model employs an energy-based model to make each pixel pair of the two images satisfy a set of super constraints (Touati and Mignotte 2017). Although these methods can be applied to different remote sensing images, they are not as good as the methods based on parameters for a preassigned pair of remote sensing images.
(3) Invariant similarity measures based methods. This type of method emphasizes the similarity of image points between heterogeneous images. The similarity measures commonly use the correlation ratio or mutual information of the image patches as criteria. Then, the patches with a high similarity will be taken as the unchanged pixels, and the change pixels will represent the abnormal value of similarity. These methods include similarity map estimation (Alberga 2009), pixel pair algorithm (PP) (Ayhan and Kwan 2019; Kwan et al. 2019), imaging modality invariant (Touati et al. 2017, 2018), patch similarity graph matrix (PSGM) (Sun et al. 2020), adaptive local structure consistency (ALSC) (Lei et al. 2020) and improved nonlocal patch based graph (INLPG) (Sun et al. 2021a; Sun et al. 2021b). The change level can be measured by similarity graphs quantitatively, but it is difficult and complex when the ground features vary greatly or the image covers a very wide area.
(4) Classification-based methods. Compared to the previous three kinds of methods, they can not only recognize the changed areas, but also obtain the class information. They acquire the CD results by comparing the classification maps of heterogeneous remote sensing images, such as post-classification comparison (PC-CC) (Mubea and Menz 2012; Wan et al. 2019) and multi-temporal segmentation and mixed classification (MS-CC) (Zhou et al. 2008). However, the accuracy often depends on the effect of segmentation or classification so that it suffers from the accumulated classification errors.
(5) Conversion and projection based methods. The main idea is to convert the heterogeneous or multimodal images to the same feature space so that the images are more comparable in CD tasks. There are two main sub-categories, including conventional transformation and deep learning based transformation. Machine learning and statistical methods are the representative conventional transformation, such as homogeneous pixel transformation (HPT) (Liu et al. 2018b), recognition of cluster splits and merges (Luppino et al. 2017), and fractal projection and Markovian segmentation-based approach (FPMSMCD) (Mignotte 2020). The deep learning based transformation has been emerging recently. For example, symmetric convolutional coupling network (SCCN) (Liu et al. 2018a), conditional generative adversarial nets (cGAN) (Niu et al. 2019), X-Net and ACE-Net method using the cyclic adversarial network (Luppino et al. 2020). The deep learning models utilize depth characteristics to unify images from the original different domains into the same domain for comparison convenience. On the whole, due to easy generalization and explanation, the conversion and projection based methods are the promising direction of heterogeneous CD.

In fact, the conversion and projection based CD models not only depend on the quality of generated images, but also rest with the effective use of these translated images. After the image translation, most of the previous CD methods are based on the unsupervised idea, and the result generally adopts a traditional difference and threshold segmentation. However, the relationship between original images and generated images is still difficult to be described, and the result is easy to be affected by noises, which will decrease the detection accuracy of the change areas. In recent studies, it is aslo found that deep learning has nonlinear characterization and outstanding capabilities in extracting image features to improve CD (Chen and Shi 2020; Peng et al. 2019). Inspired by this, a novel deep translation based change detection network (DTCDN) for optical and SAR remote sensing images is proposed. Our DTCDN does not treat the translated images like the previous methods, but imports them into a supervised CD network with the reference images. The contributions of this research focus on the following three aspects:

(1) A general framework DTCDN is proposed to automatically extract the change regions between optical and SAR remote sensing images. The core idea is to first translate the optical and SAR remote sensing images into the same feature domain, and then input them into the CD network instead of directly using unsupervised methods. The CD network can efficiently extract and utilize deep image features to generate a CD binary map.
(2) An improved UNet++ was constructed for the CD task. The number of network parameters is reduced by depthwise separable convolution, and the training process is also improved. A multi-scale loss function is proposed, which is different from multiple side-outputs fusion (MSOF) (Peng et al. 2019). The multi-scale loss function takes features of different sizes into account to optimize global and local effects.
(3) This work proves the effectiveness of deep image translation in the optical and SAR remote sensing images CD. Through deep image translation, the unification between optical and SAR images is realized. Moreover, it is demonstrated that a better translation model can improve the generated images, thereby improving the outcomes of CD.

The rest of this paper is organized as follows. In Section 2, the study areas and data sets are described. Section 3 details the proposed method, followed by the experiments and results in Section 4. The discussion is presented in Section 5. Finally, our conclusion is drawn in Section 6.

2. Study areas and data description

In this paper, four pairs of optical and SAR remote sensing data sets from three different regions were used, including Gloucester in the UK (Longbotham et al. 2012; Mignotte 2020; Touati et al. 2019)., California in the USA (https://sites.google.com/view/luppino/data), and

Table 1
The detailed information of four optical and SAR data sets.

Name	Date	Size	Event	Spatial resolution	Sensor
Gloucester I	July 2006-July 2007	2325 × 4135	Flooding	0.65 m	QuickBird 2/TerraSAR-X
Gloucester II	Sept. 1999-Nov. 2000	1250 × 2600	Flooding	≈25 m	SPOT/ERS-1
California	Jan. 2017-Feb. 2017	2000 × 3500	Flooding	≈15 m	Landsat 8/ Sentinel 1-A
Shuguang	June 2008-Sept. 2012	419 × 342	Building Construction	8 m	Radasat-8/ Google Earth

Shuguang village in China (Mignotte 2020; Touati et al. 2019). Each dataset uses an optical image, a SAR image, and the corresponding ground truth for heterogenous CD. Table 1 reports their detailed information. All data sets have been resampled, coregistered and cropped to cover the same geographical areas.

The first Gloucester I data set consists of two images captured in Gloucester (UK) before and after flooding, as shown in Fig. 1 (Mignotte 2020; Touati et al. 2019). The optical image was acquired by Quickbird 2 in July 2006, while the SAR image was acquired by TerraSAR-X in July 2007. These two images cover the same area and their image size is 2325 × 4135 with a resolution of 0.65 m. This ground truth is generated manually by integrating some prior information and expert knowledge.

The Gloucester II data set was also collected from Gloucester and employed in the 2009–2010 IEEE GRSS Data Fusion Contest, as shown in Fig. 2 (Longbotham et al. 2012). Similarly, this dataset contains a flood, however, the sensors and time are different with Gloucester I. The optical image was acquired by SPOT in September 1999, while the SAR image was acquired by ERS-1 in November 2000. The spatial resolution is about 25 m and the ground truth is provided by the IEEE Data Fusion Technical Committee.

The third data set is a pair of multispectral and SAR images acquired in California (USA) before and after a flood, as shown in Fig. 3 (Luppino et al. 2019). The optical image was acquired by Landsat 8 on 5 January 2017 covering Sacramento County, Yuba County and Sutter County, California. The SAR image was taken by Sentinel 1-A and covers the same geographical area on 18 February 2017. Compared with the other three data sets, the SAR image is characterized by 3 bands which are recorded in polarisations VV and VH, and the ratio between the two intensities as a third channel. The two images' size is 2000 × 3500 with a resolution of about 15 m. The ground truth was obtained by two other single-polarization Sentinel-1 images at the same time.

The fourth data set contains two images acquired in Shuguang village of Dongying city (China), as shown in Fig. 4 (Mignotte 2020; Touati et al. 2019). The two images with the size of 419 × 342 show a building construction from farmland, and in the bottom right there is a river change. The SAR image was acquired by Radarsat-8 in June 2008, while the optical image was obtained by Google Earth in September 2012 with a resolution of 8 m. The ground truth was obtained manually by prior information that combines expert knowledge based on input images.

3. Methodology

The proposed DTCDN method for optical and SAR remote sensing images CD consists of two parts, a deep translation network and a CD network (Fig. 5). Due to the imaging difference between optical and SAR images, there will be remarkable errors if they are directly imported to the CD network for detecting changed pixels. As shown in Fig. 5, in order to reduce the errors, one of the heterogeneous images is first translated to the domain of the other image by the deep translation network. Here only shows one of the translation methods that optical image is converted to SAR image. And then the translated image and another reference image in the same domain are imported to the homogeneous CD network to generate the final change map. In the training phase,

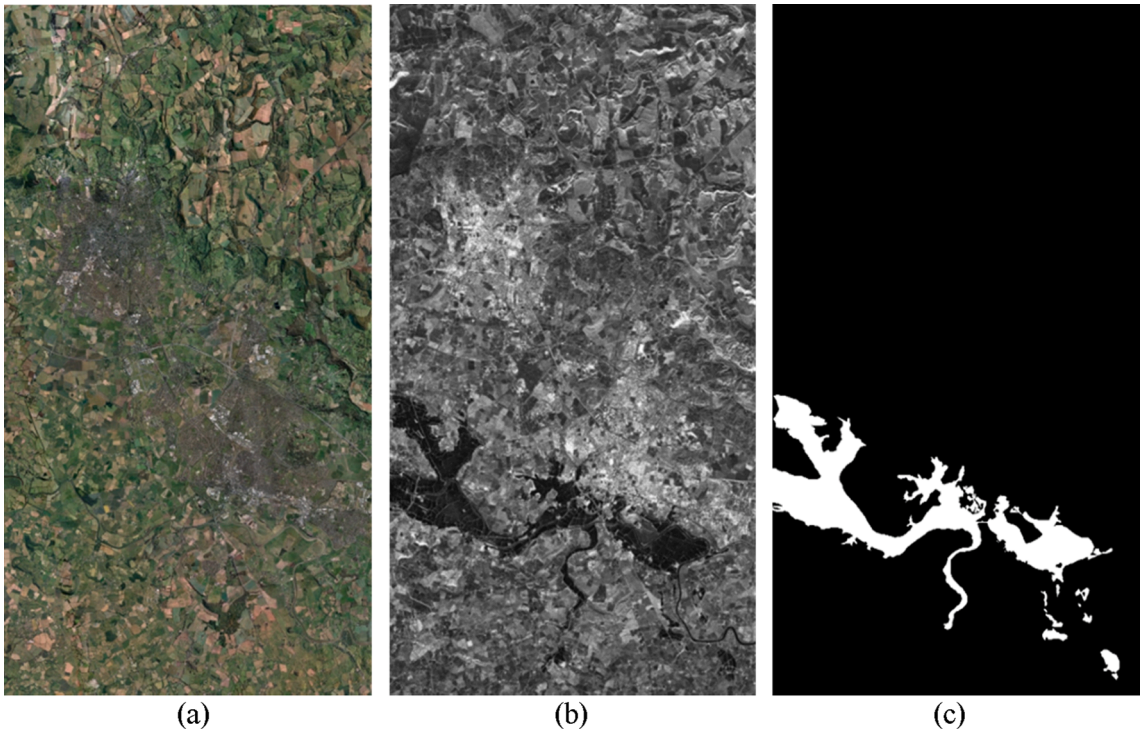

Fig. 1. Gloucester I data set in the UK. (a) Optical image. (b) SAR image. (c) Ground truth. Note that the change areas are filled with white color in (c).

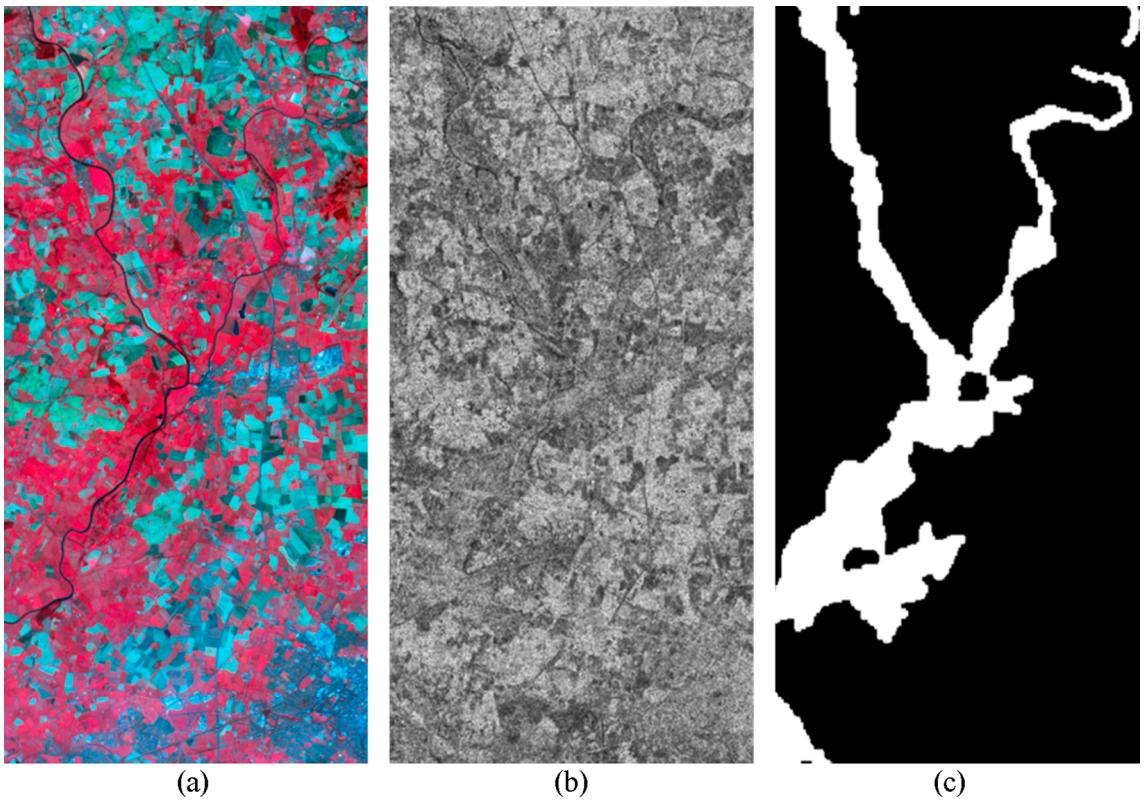

Fig. 2. Gloucester II data set in the UK. (a) Optical image. (b) SAR image. (c) Ground truth. Note that the change areas are filled with white color in (c).

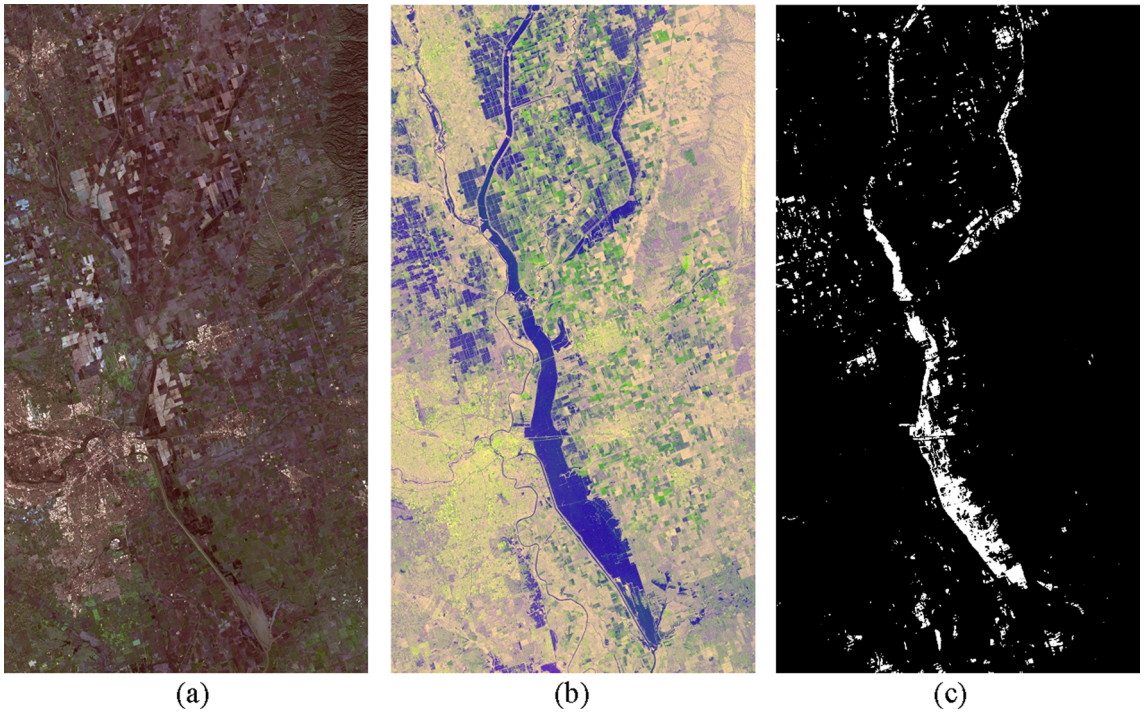

Fig. 3. California data set in the USA. (a) Optical image. (b) SAR image. (c) Ground truth. Note that the change areas are filled with white color in (c).

optical and SAR image pairs were fed into deep translation network. Besides, our CD network is supervised, which needs corresponding label.

3.1. Deep translation network

The deep translation network aims at reducing the differences between optical and SAR images by encoding and converting them to the same feature space, and also prepares for the CD network. The image-to-

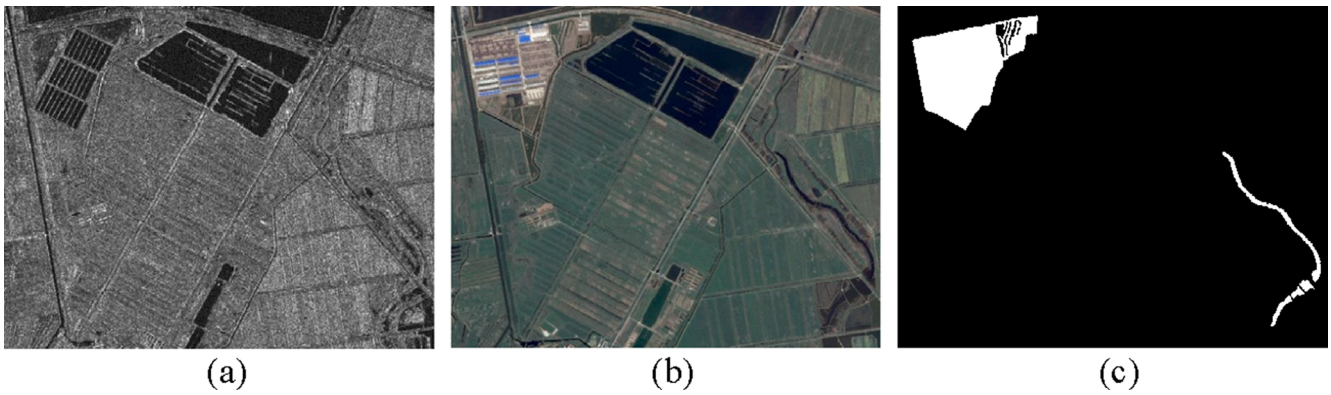

Fig. 4. Shuguang village data set in China. (a) SAR image. (b) Optical image. (c) Ground truth. Note that the change areas are filled with white color in (c).

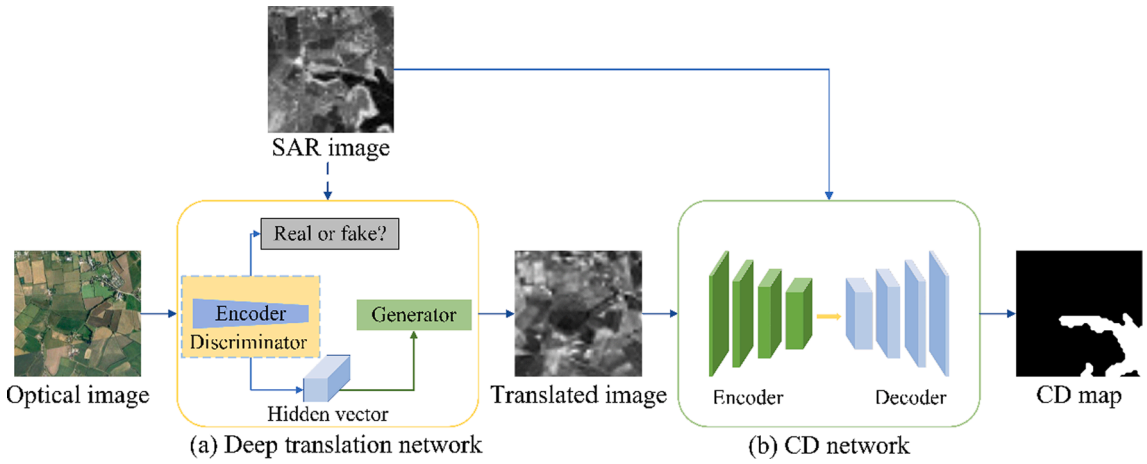

Fig. 5. The structure of proposed DTCDN with a two-step process. (a) Deep translation network and (b) CD network.

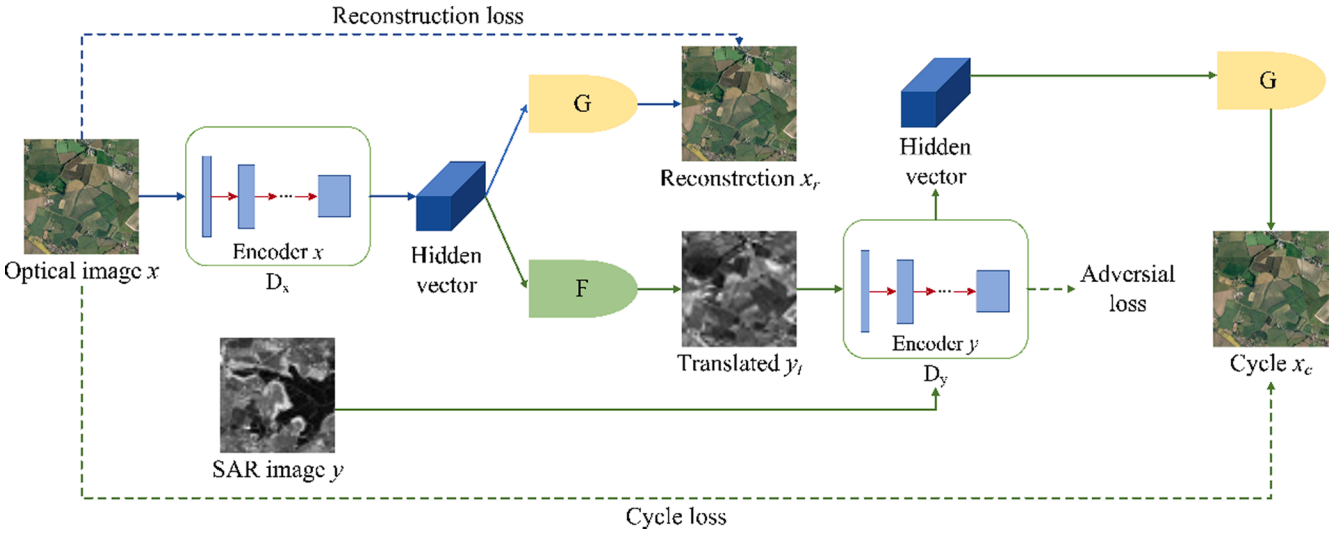

Fig. 6. The deep translation model, in which the translation from optical images to SAR images is taken as an example. The G and F represent the generators of domains X and Y, respectively. D_x and D_y are the discriminators of domains X and Y. With adversarial loss, reconstruction loss and cycle loss, the quality of generated SAR images can be improved during the cycle.

image translation is a common image processing method whose basic idea is to learn the mapping functions between an input image and an output image with a training set of aligned image pairs (Zhu et al. 2017). It originates from image analogies, which realizes a translation from a single image to a single image (Hertzmann et al. 2001). With the pix2pix appeared, image translation based GAN has attracted much more attention (Isola et al. 2017). The unsupervised cycle-consistent adversarial networks (Cycle-GAN) can transfer characteristics from a domain to another domain by cycle consistency loss (Zhu et al. 2017). The pix2pixHR is an improved version of pix2pix, which introduces a coarse

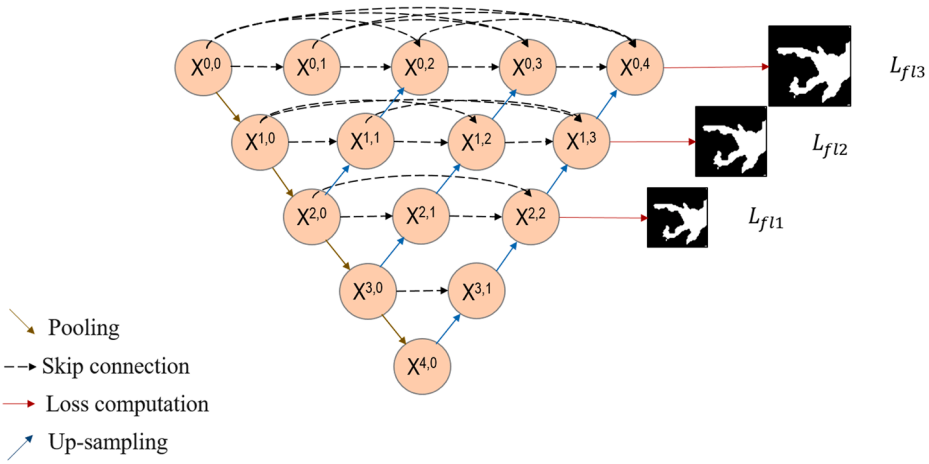

Fig. 7. The improved U-Net++. The black arrows represent the skip connections, and the red arrows mean the multi-scale loss in different levels of prediction structure and down-sampling labels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

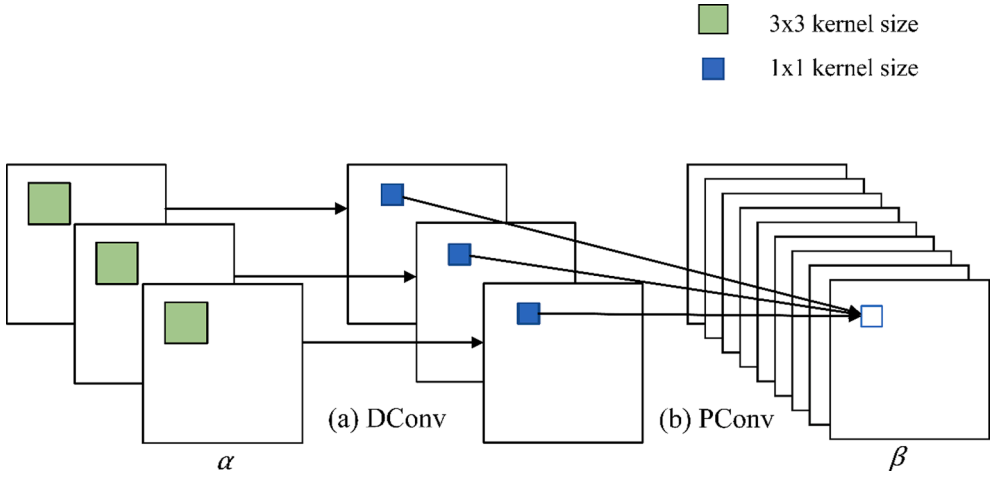

Fig. 8. The example of depthwise separable convolution, which is divided into two parts: (a) depthwise convolution (DConv), and (b) pointwise convolution (PConv).

Table 2
The CD results of Gloucester I data set.

Dataset	Method	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Gloucester I	RX	15.78	30.64	20.83	76.63
	INLPG	28.51	61.27	38.91	80.70
	SCCN	59.74	61.68	60.70	91.99
	cGAN	33.84	68.68	45.34	83.42
	X-Net	66.26	78.75	71.97	93.85
	ACE-Net	65.65	70.73	68.10	93.35
	HPT	31.75	65.18	42.68	82.46
	Without translation	83.21	80.92	82.05	96.45
	DTCDN	**89.96**	**89.93**	**89.95**	**97.98**

to a fine generator and a multi-scale discriminator architecture to produce high resolution generated images (Wang et al. 2018). The translation idea based excellent adversarial discriminative network is also applied in the field of remote sensing image (Shi et al. 2020). Additionally, the dialectical GAN converted the Sentinel 1 images to TerraSAR images (Ao et al. 2018). Fuentes Reyes et al. (2019) further discussed the limit of Cycle-GAN and optimized the structure for the translation from SAR to optical images. Turnes et al. (2020) proposed an Atrous-cGAN to enhance fine details in the generated optical image from SAR with the atrous spatial pyramid pooling (ASPP) for exploiting spatial context at multiple scales. These studies have demonstrated the image translation also has a good performance for remote sensing images.

The deep translation model used here is no-independent-component-for-encoding GAN (NICE-GAN) which is based on Cycle-GAN (Chen et al. 2020; Zhu et al. 2017). The original Cycle-GAN makes use of the encoder to transform the image to a vector containing hidden information, then the generator produces images in another domain, and finally the discriminator classifies whether the generated image is real or fake. Besides, the NICE-GAN, referred to the Introspective Networks (INN), reuses early layers of a certain number in the discriminator to improve the quality of generation and make the entire structure more compact and efficient (Jin et al. 2017; Lazarow et al. 2017; Lee et al. 2018). The original NICE-GAN with multi-scale formulation in discriminator and residual attention was used in our experiment. As shown in Fig. 6, the proposed NICE-GAN converts a domain X of heterogeneous remote sensing images to another domain Y, and even achieves the exchange of Y and X. Taking an optical image in the X domain as an example, the optical image is imported to the discriminator D_x, which has the functions of encoding into a hidden vector and classifies the image as true or false. The hidden vector enters into the generators G and F, and the reconstruction x_r and translated y_t are generated. Then, y_t is fed into the D_y to encode and compute the

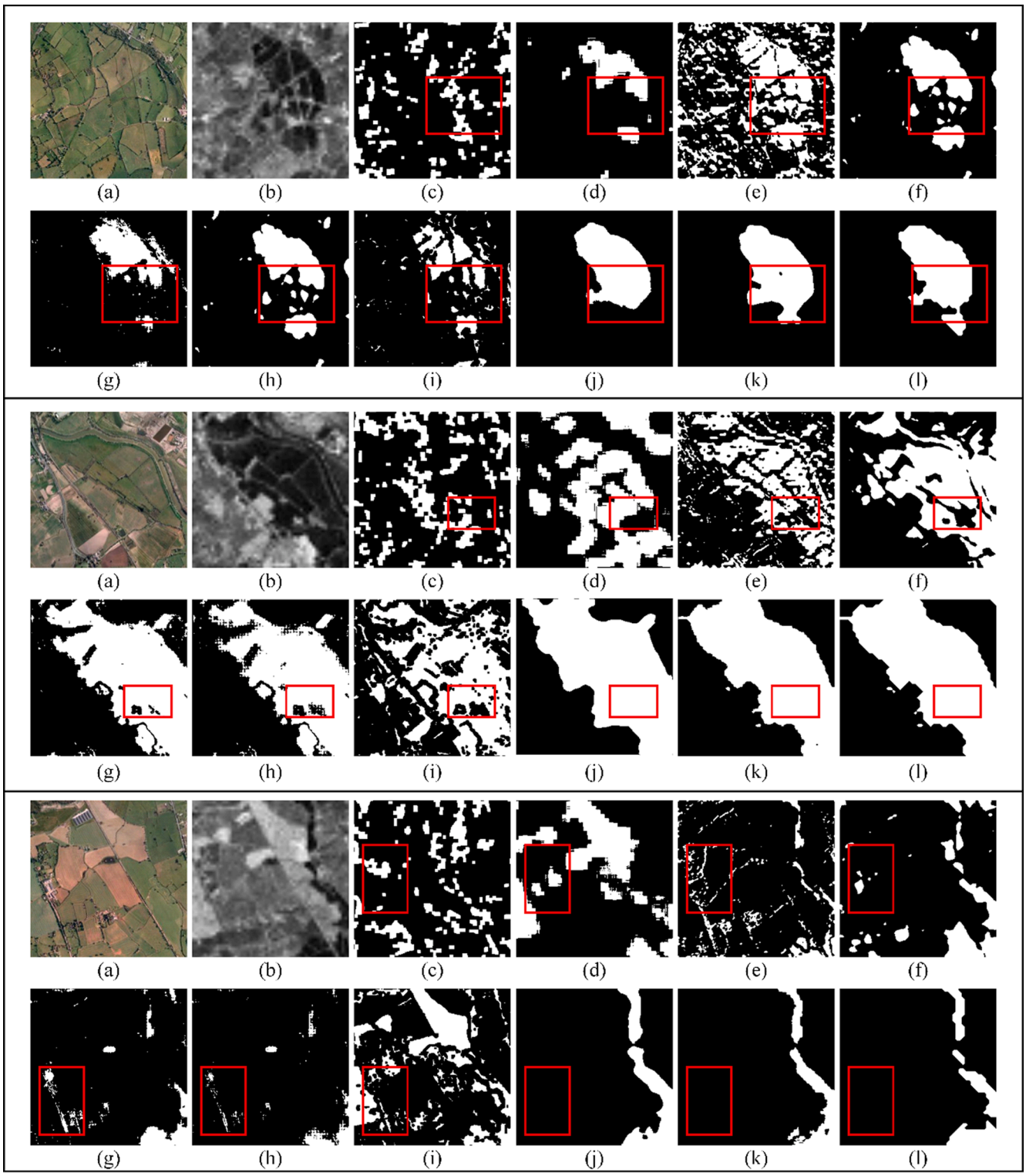

Fig. 9. Visual comparison of CD results using different approaches on Gloucester I data set. From left to right: (a) image T1, (b) image T2, (c) RX, (d) INLPG, (e) SCCN, (f) cGAN, (g) X-Net, (h) ACE-Net, (i) HPT, (j) Without translation, (k) DTCDN, (l) ground truth.

adversarial loss. In this process, the SAR image y will be provided as references and participate in the computation. After that, the hidden vector from the D_y is input to G to obtain the cycle x_c. This means one complete iteration. Besides, the SAR image y in domain Y also takes the same operations in every iteration. In network structure, three important losses are defined, adversarial loss L_{adver}, reconstruction loss L_{recon} and cycle loss L_{cyc}, which are defined as follows:

$$L_{adver} = E_{y\sim\ p_{data}(y)}\left[\left(D_y(\mathrm{y})\right)^2\right] + E_{\mathrm{x}\sim\ p_{data}(\mathrm{x})}\left[1 - D_y(F(x))^2\right] + E_{x\sim p_{data}(x)}\left[\left(D_x(\mathrm{x})\right)^2\right] + E_{\mathrm{y}\sim\ p_{data}(\mathrm{y})}\left[1 - D_x(G(y))^2\right] \quad (1)$$

$$L_{recon} = E_{\mathrm{x}\sim\ p_{data}(\mathrm{x})}[\|G(\mathrm{x}) - \mathrm{x}\|_1] + E_{y\sim\ p_{data}(y)}[\|F(y) - y\|_1] \quad (2)$$

$$L_{cyc} = E_{x\sim\ p_{data}(x)}[\|G(F(\mathrm{x})) - \mathrm{x}\|_1] + E_{y\sim\ p_{data}(y)}[\|F(G(y)) - y\|_1] \quad (3)$$

Table 3
The CD results of Gloucester II data set.

Dataset	Method	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Gloucester II	RX	27.36	40.68	32.71	70.47
	INLPG	42.48	42.47	42.48	75.36
	SCCN	34.12	35.36	34.73	72.35
	cGAN	36.53	52.13	42.96	77.01
	X-Net	53.26	37.23	43.83	82.14
	ACE-Net	57.12	39.84	46.94	85.69
	HPT	50.46	49.65	50.05	88.52
	Without translation	83.05	80.65	81.83	93.45
	DTCDN	**90.78**	**86.66**	**88.67**	**96.33**

where the image distribution probabilities are denoted as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$. $||\,||_1$ represents the l_1 norm. With Eqs. (1), (2) and (3) in the adversarial discriminative process, the more precise optical and SAR images will be generated. Two translation modes were realized in this study: 1) from optical images to SAR images and 2) from SAR images to optical images. In this case, the optical and SAR images will be unified into the same space.

3.2. Change detection network

After deep image translation, the homogeneous images are obtained from the heterogeneous images, which will be fed to the CD network. Owing to the rapid development of deep convolutional neural networks (CNN), it has been demonstrated the deep learning methods can obtain good performance in remote sensing images (Daudt et al. 2018; Guo et al. 2020; Kampffmeyer et al. 2016; Liu et al. 2020b; Touati et al. 2020). The CD methods base on deep leraning has been studied and applied rapidly (Liu et al. 2021; Wang et al. 2019). U-Net is a better version of CNN because of the skip connection (Ronneberger et al. 2015). Our proposed CD network is based on the improved U-Net, called U-Net++ (Zhou et al. 2018). Compare with U-Net, U-Net++ fills the hollow of middle layers and utilizes a series of nested and dense skip

Table 4
The CD results of California data set.

Dataset	Method	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
California	RX	18.54	42.08	25.74	88.50
	INLPG	30.92	**80.27**	44.65	90.58
	SCCN	31.11	54.01	39.47	84.55
	cGAN	32.43	59.20	41.91	92.82
	X-Net	35.34	59.11	44.24	90.83
	ACE-Net	30.59	76.49	43.71	91.38
	HPT	53.95	51.32	52.60	95.62
	Without translation	58.61	64.99	61.63	96.42
	DTCDN	**66.73**	78.24	**72.03**	**97.61**

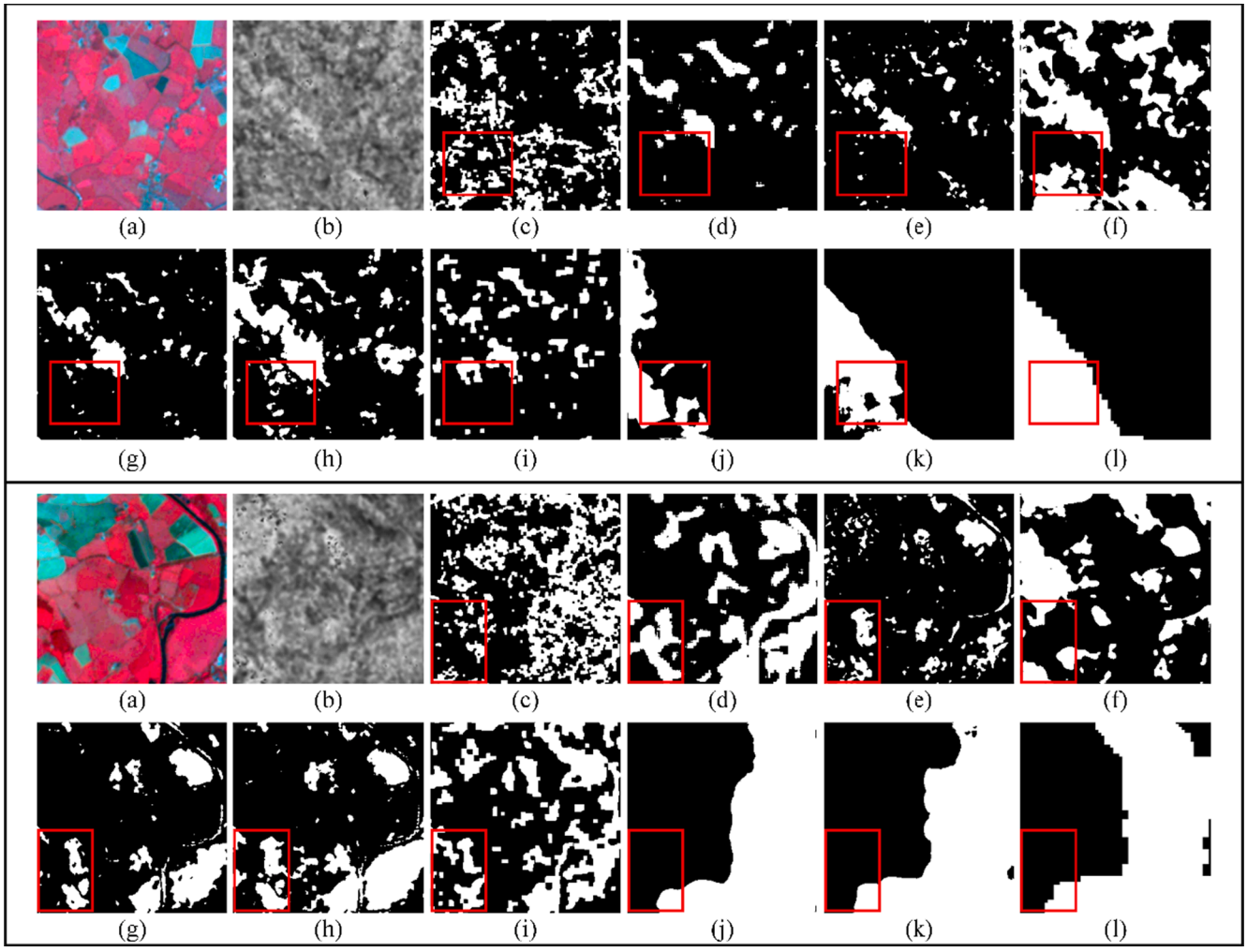

Fig. 10. Visual comparison of CD results using different approaches on Gloucester II data set. From left to right: (a) image T1, (b) image T2, (c) RX, (d) INLPG (e) SCCN, (f) cGAN, (g) X-Net, (h) ACE-Net, (i) HPT, (j) Without translation, (k) DTCDN, (l) ground truth.

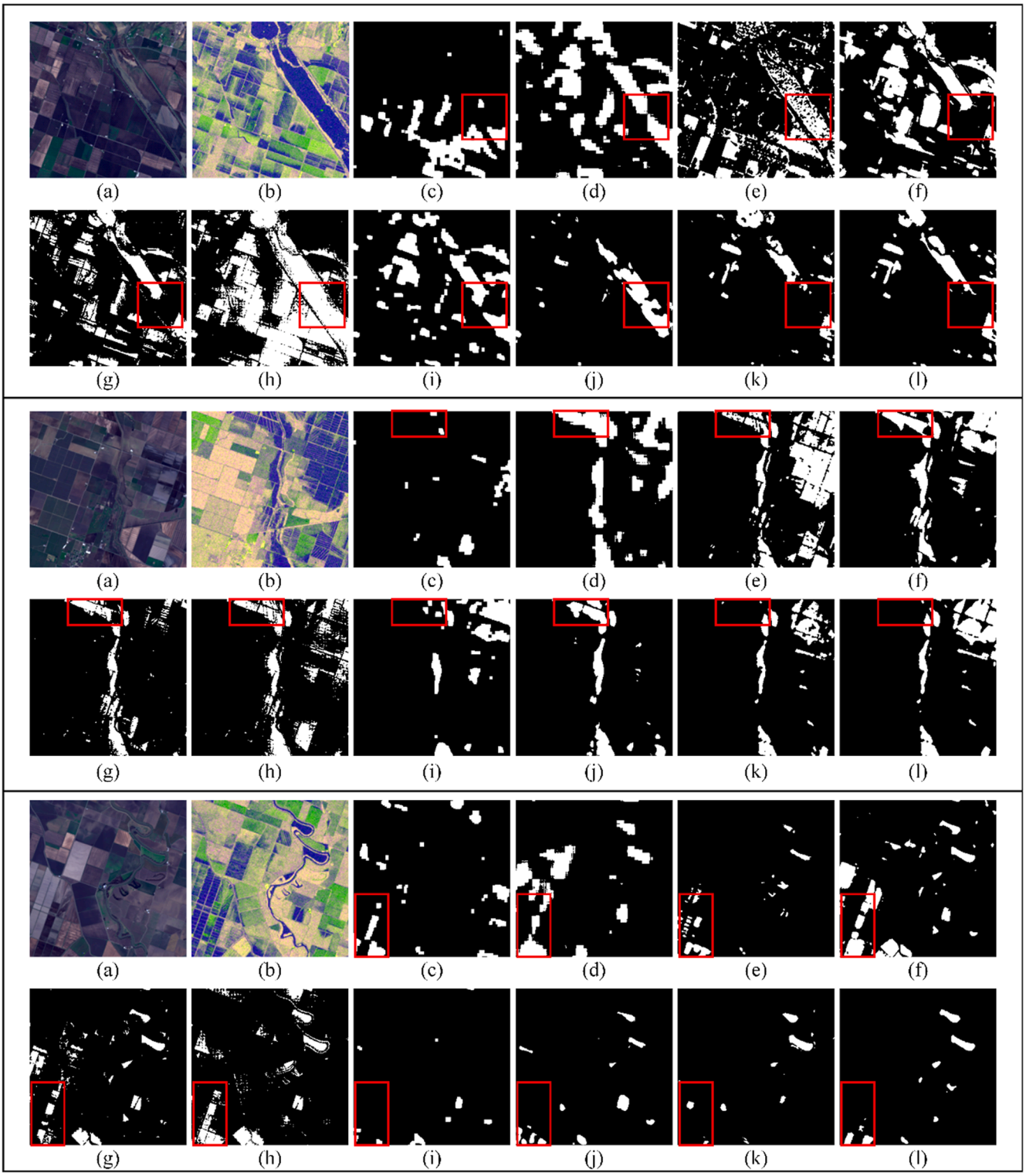

Fig. 11. Visual comparison of CD results using different approaches on California data set. From left to right: (a) image T1, (b) image T2, (c) RX, (d) INLPG (e) SCCN, (f) cGAN, (g) X-Net, (h) ACE-Net, (i) HPT, (j) Without translation, (k) DTCDN, (l) ground truth.

connections to combine features. The advantage is that it can capture and integrate features of different levels to strengthen robustness by feature concatenation. Although U-Net++ can achieve better results, its numerous dense connections cause a large increase in the amounts of parameters and inference time. Towards this end, depthwise separable convolution (Chen et al. 2018; Liu et al. 2020a; Shang et al. 2020) is introduced to replace the ordinary convolution. Simultaneously, the multi-scale loss is added to different-level predictions to extract more context information and speed up the training process. The improved UNet++ is shown in Fig. 7.

3.2.1. *Depthwise separable convolution for U-Net++*

Standard convolution *Conv* takes action to spatial and channel simultaneously, but this operation consumes a lot of computation. The

Table 5
The CD results of Shuguang data set.

Dataset	Method	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Shuguang village	RX	46.28	43.77	44.99	92.76
	INLPG	39.90	90.01	55.29	90.40
	SCCN	30.21	43.42	35.62	89.97
	cGAN	54.03	87.23	66.73	95.52
	X-Net	68.19	75.82	71.80	96.37
	ACE-Net	79.17	75.87	75.34	97.13
	HPT	51.81	89.99	65.76	96.25
	Without translation	80.34	79.07	79.70	97.49
	DTCDN	**92.92**	**90.25**	**91.56**	**99.75**

depthwise separable convolution *SConv* can decompose the standard convolution into two steps, the depthwise convolution *DConv* and the pointwise convolution *PConv* (Chollet 2017), which significantly reduces the number of mathematical operations and parameters. The formulation of the *Conv*, *DConv*, *PConv* and *SConv* can be expressed as follows:

$$Conv(W,a)_{(i,j)} = \sum_{m,n,k}^{M,N,K} W_{(m,n,k)} \bullet a_{(i+m,j+n,k)} \tag{4}$$

$$DConv(W_d,a)_{(i,j)} = \sum_{m,n,k}^{M,N,K} W_{d(m,n)} \bullet a_{(i+m,j+n)} \tag{5}$$

$$PConv(W_p,a)_{(i,j)} = \sum_{k}^{K} W_p \bullet a_{(i,j)} \tag{6}$$

$$SConv(W_p,W_d,a)_{(i,j)} = PConv(W_p,a)_{(i,j)}(W_p,DCov(W_d,a)_{(i,j)}) \tag{7}$$

where the W, W_d and W_p represent the weight matrix of *Conv*, *DCov* and *PCov*, respectively. i and j are the pixel's location in feature map a. m, n, and k are three dimensions of the kernel size.

Fig. 8 represents a specific example of *SConv*. Let $\alpha \in P^{H\times W\times 3}$ represents an input feature from image P with 3 channels. The first step *DConv* applies 3 × 3 convolution kernels to each channel for feature extraction separately. Certainly, the output of the 3 channels can also be obtained from *DConv* and as the input of *PConv*. The *PConv* is used to generate an output feature map $\beta \in R^{H\times W\times C_0}$ by feature fusion on the 1 × 1 convolution kernel, where C_0 is the number of output channels.

3.2.2. Loss function

A common problem in deep learning based classifiers is that some classes have a significantly higher number of training examples than other classes. This phenomenon is referred to as class imbalance in CD (Buda et al. 2018). The class of unchanged pixels usually takes the lead,

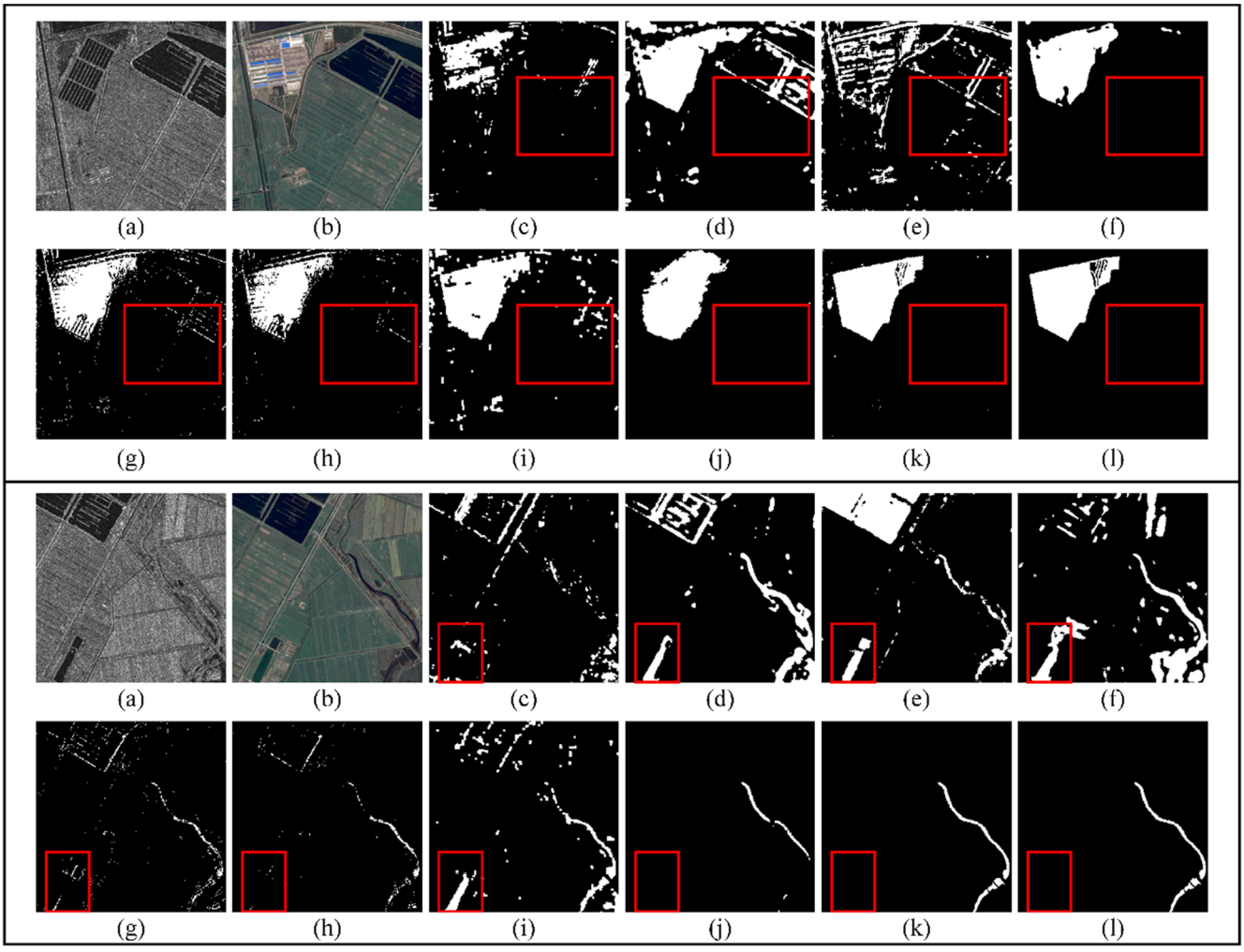

Fig. 12. Visual comparison of CD results using different approaches on Shuguang data set. From left to right: (a) image T1, (b) image T2, (c) RX, (d) INLPG (e) SCCN, (f) cGAN, (g) X-Net, (h) ACE-Net, (i) HPT, (j) Without translation, (k) DTCDN, (l) ground truth.

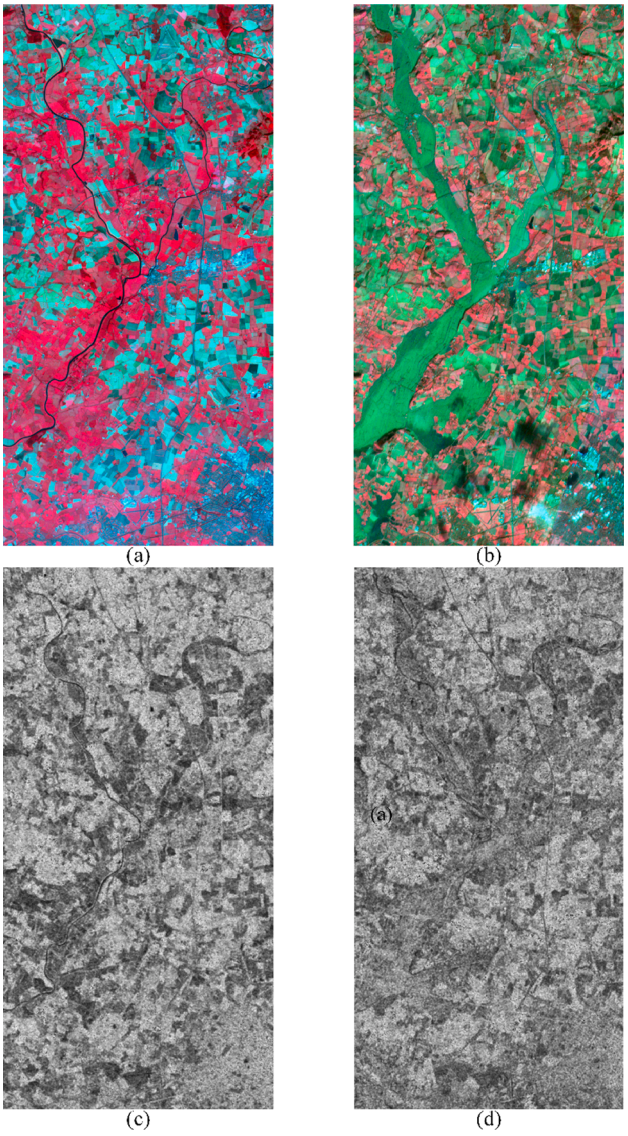

Fig. 13. The optical and SAR images from both timesteps in Gloucester II dataset. (a) Pre-event optical image. (b) Post-event optical image. (c) Pre-event SAR image. (d) Post-event SAR image.

Table 6
The assessment of CD results for three image pairs.

Image pairs	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Optical and optical	89.99	**90.84**	**90.41**	**96.81**
SAR and SAR	90.55	81.15	85.60	95.48
Optical and SAR (DTCDN)	**90.78**	86.66	88.67	96.33

while there is only a small part of changed pixels. To reduce the impact of class imbalance, the focal loss (Lin et al. 2017a) is resorted and modified from the standard cross-entropy loss (de Boer et al. 2005) in our model. This loss function can reduce the weight of easy-to-classify samples so that the focus is on difficult-to-classify samples during training. The focal loss is as follow:

$$L_f = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{8}$$

where t is the location of pixels, p_t means the probability of changed class, and α_t represents the weighting factor. The tunable focusing parameter $\gamma \geq 0$.

Inspired by feature pyramid networks, a multi-scale loss is used to optimize the model (Lin et al. 2017b). The multi-scale loss mainly acts on the upper three layers with long skip connections, because the upper layers contain more valuable features. The label is downsampled to calculate a focal loss from the prediction of each of the upper three layers. The total loss is calculated as follows:

$$Loss = w_1 * L_{f1} + w_2 * L_{f2} + w_3 * L_{f3} \tag{9}$$

where w_i means the weight of focal loss in ith layer, $i = 1, 2, 3$. Multi-scale loss can make use of more information in different scales, and make loss function converge quickly and become stable.

4. Experiments and results

In this section, with the above four datasets, the experiments were carried out to prove the effectiveness of our proposed model. The results of four datasets were analyzed in detail separately. To verify the validity of DTCDN, it was compared with other CD models of optical and SAR remote sensing images, including Reed-Xiaoli (RX) (Reed and Yu 1990), INLPG (Sun et al. 2021b), SCCN (Liu et al. 2018a), cGAN (Niu et al. 2019), X-Net and ACE-Net (Luppino et al. 2020) and HPT (Liu et al. 2018b). RX assumes that the background obeys the multivariate normal distribution and constructs the likelihood ratio detection operator. And it can be applied to CD of original and translated images (Guo et al. 2014; Zhou et al. 2016). INLPG constructs the K-NN graph and calculates the nonlocal patch similarity structure difference to measure the change level. SCCN obtains a differential image from the two encoded images in coupling layers. cGAN transfers optical image to SAR image through the translation network and SAR image gets a similar domain with transferred images through the approximation network. X-Net takes the affinity information as prior knowledge, and the difference image is obtained after the two images are transformed by the cyclic GAN network. Similarly, ACE-Net introduces a latent space, in which the image is translated into the architecture of four networks. HPT is a supervised heterogenous CD method through mapping pixel using K-nearest unchanged pixels to realize transformation. In our experiment, the label used in HPT is same as in DTCDN. There are not so many supervised methods for CD of heterogeneous remote sensing images, so the selected methods are only comparable with our method to some extent, because most of them are unsupervised. To this end, the experiments of without translation were carried in which the original images are not translated and directly imported to the CD network. Besides, the corresponding SAR and optical images from both timesteps in Gloucester II were directly inputted into the CD network to further explore the effectiveness of the proposed method.

In the image translation network of our proposed model, 6-layers generator with residual blocks and 7-layers encoder in the discriminator were selected. Adam with a learning rate of 0.0005 was chosen as an optimization algorithm. For the CD network, the 4-layers U-Net++ with the number of {32, 64, 128, 256} convolutional filters in the encoder part were used. By concatenating an optical image and a SAR image, the input tensor size is 256 × 256 × C, where C is the sum of two images' channels. The optimal weights of the multi-scale loss w_1, w_2, and w_3 were 0.2, 0.3 and 0.5 after repeated test experiments. For the focal loss parameters, they were set to $\gamma = 2$ and $\alpha = 0.25$. All experiments were done by Pytorch and running in the NVIDIA RTX 2080Ti with 11 GB RAM.

4.1. Evaluation metrics

In order to obtain a clear and quantitative assessment, the intermediate results of deep translation and the CD results were evaluated separately by different criteria. For the image translation, the Frchet inception distance (FID) (Heusel et al. 2017) and kernel inception distance (KID) (Bińkowski et al. 2018) can quantitatively describe the

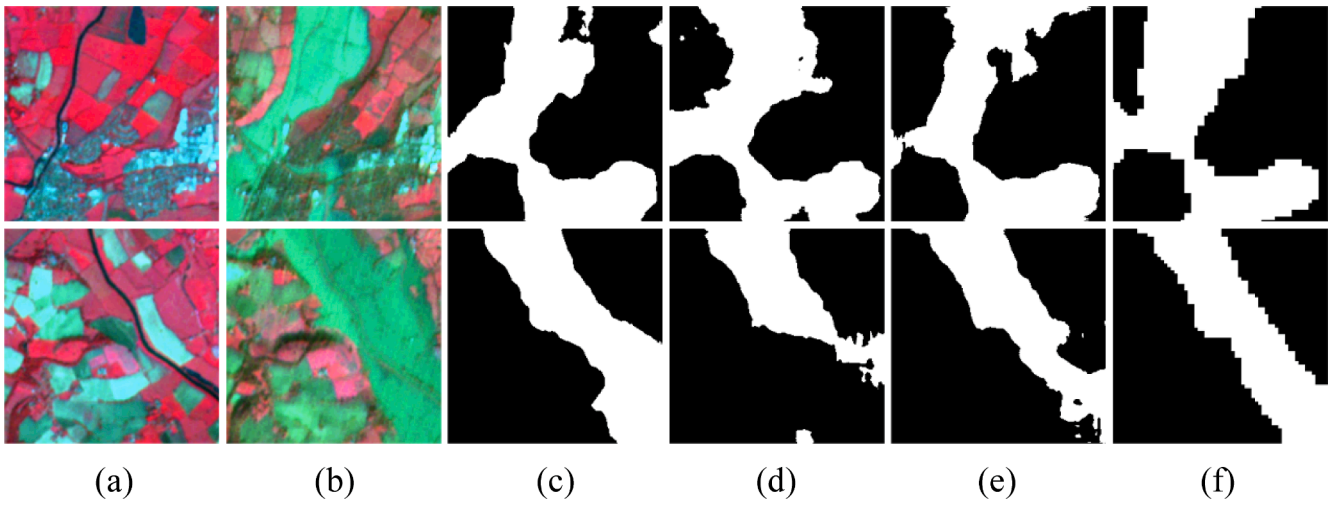

Fig. 14. The CD results from different image pairs. (a) T1 optical image, (b) T2 optical image, (c) optical and optical, (d) SAR and SAR, (e) optical and SAR (DTCDN), (f) ground truth.

Table 7
The assessments of deep translation through two different modes on four data sets.

Dataset	Mode	FID↓	KID × 100↓
Gloucester I	SAR → Op	182.47	7.37
	Op → SAR	130.16	5.76
Gloucester II	SAR → Op	121.92	5.23
	Op → SAR	**36.26**	**0.79**
California	SAR → Op	166.98	6.16
	Op → SAR	100.53	2.71
Shuguang	SAR → Op	86.05	1.76
	Op → SAR	117.13	3.62

distance of features between real images and generated images. The FID applies pre-trained Inception V3 to extract the 2048-dimensional vector before the fully connected layer as a feature of the picture.

$$FID = \|\mu_{r-}\mu_g\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \tag{10}$$

where r and g are the features of real image and generated image, respectively. μ_r and μ_g represent the mean value of r and g, respectively. Σ_r and Σ_g are the covariance matrix of r and g, respectively. Tr represents the trace of matrix. $||\ ||^2$ computes the square of difference between μ_r and μ_g.

The squared maximum mean discrepancy (MMD) between Inception representations with a polynomial kernel function k is used to calculate KID,

$$KID = E_{\substack{r,r^{'}\ pdata(r),\\ g,g^{'}\ pdata(g)}}[k(r,r^{'}) - 2k(r,g) + k(r,g^{'})] \tag{11}$$

$$k(\theta,\theta^{'}) = (\frac{1}{d}\theta^T\theta^{'} + 1)^3 \tag{12}$$

where $r, r^{'}$ are the samples from the real images, and $g, g^{'}$ are the samples from the generated images. d represents dimension of image θ. Compared with FID, KID has more advantages. KID is an unbiased estimate and does not use the parameter form of activation distribution. The low values of two indicators mean better translation results.

For the results of CD, the precision and recall were calculated by the samples of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) in the confusion matrix. F1 is a comprehensive result of precision and recall. Overall accuracy (OA) is the evaluation of the correct ratio of all samples. The calculation equations of the four in-

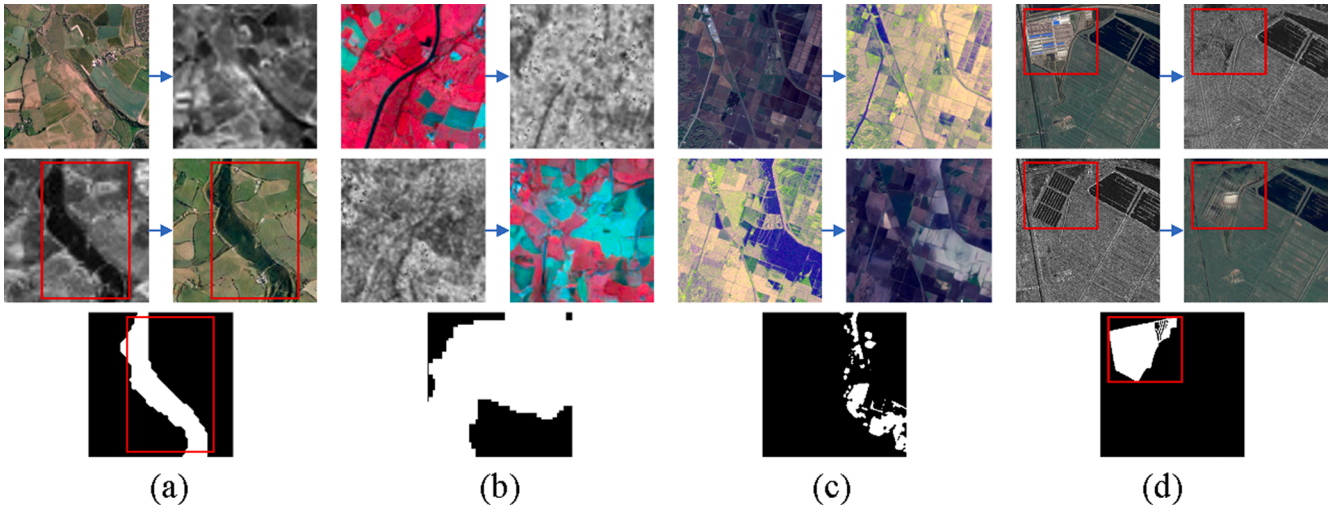

Fig. 15. The translated images on four data sets. (a) Gloucester I data set. (b) Gloucester II data set. (c) California data set. (d) Shuguang village data set. From top to bottom: optical images to SAR images, SAR images to optical images, ground truth.

Table 8
CD results of two different modes on four data sets.

Translation Method	Mode	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Gloucester I	SAR → Op	86.13	79.40	82.63	96.66
	Op → SAR	**89.96**	**89.93**	**89.95**	**97.98**
Gloucester II	SAR → Op	**90.78**	**86.66**	**88.67**	**96.33**
	Op → SAR	79.64	85.65	82.54	92.76
California	SAR → Op	**67.82**	60.53	63.97	96.70
	Op → SAR	66.73	**78.24**	**72.03**	**97.61**
Shuguang	SAR → Op	90.44	81.59	85.46	99.76
	Op → SAR	**92.92**	**90.25**	**91.56**	**99.75**

dicators are as follows:

$$Precision = TP/(TP + FP) \tag{13}$$

$$Recall = TP/(TP + FN) \tag{14}$$

$$F1 = 2{*}Precison{*}Recall/(Precison + Recall) \tag{15}$$

$$OA = (TP + TN)/(TP + TN + FP + FN) \tag{16}$$

Table 9
FID and KID of four different translation methods for Gloucester I data set.

Translation Method	Mode	FID↓	KID × 100↓
pix2pix	SAR → Op	286.67	10.03
	Op → SAR	309.40	10.76
pix2pixHR	SAR → Op	271.53	9.59
	Op → SAR	190.44	8.45
Cycle-GAN	SAR → Op	217.44	8.31
	Op → SAR	256.62	9.71
Nice-GAN	SAR → Op	182.47	7.37
	Op → SAR	**130.16**	**5.76**

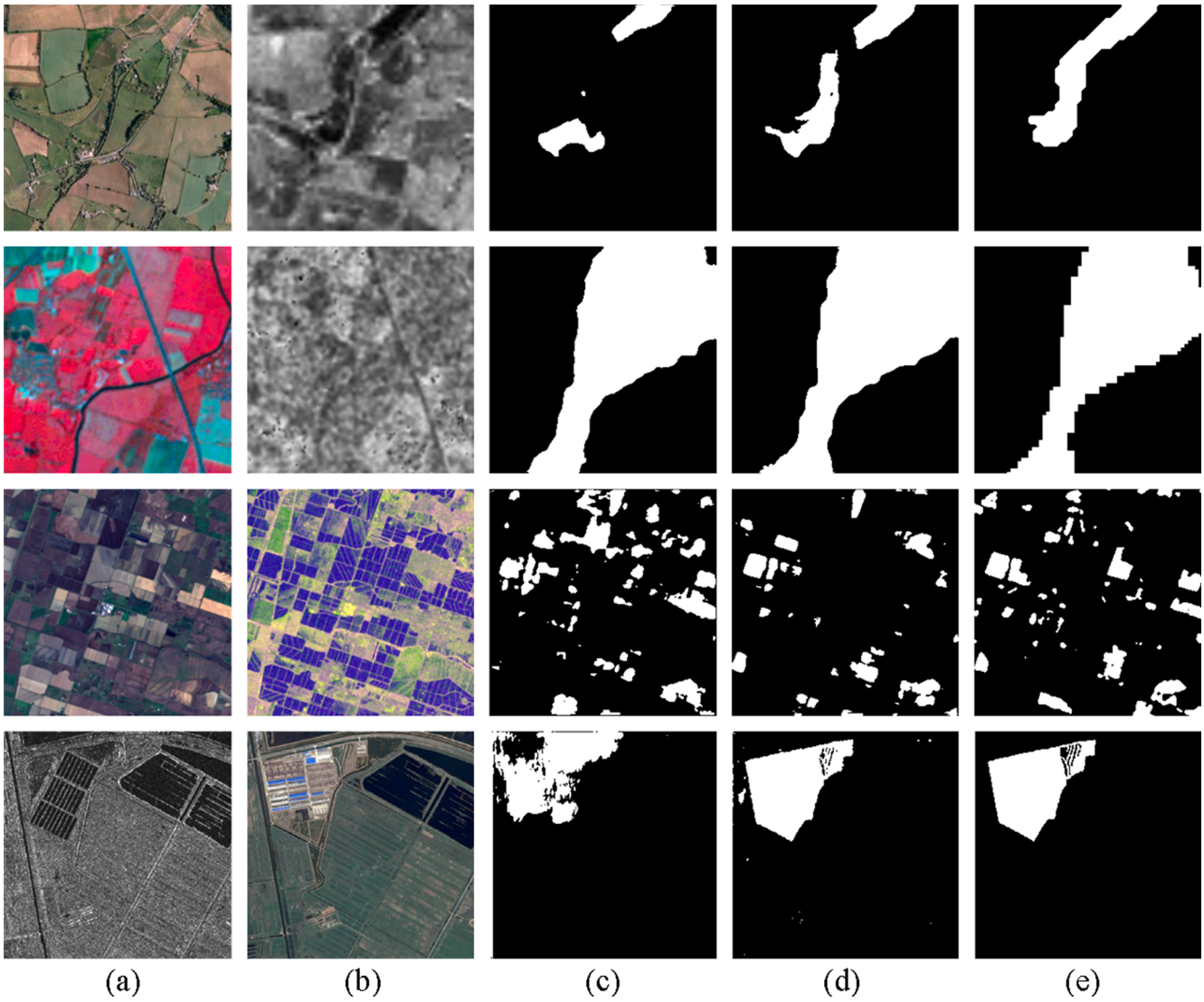

Fig. 16. Visual comparison of CD results using two translation modes for four data sets. From left to right: (a) image T1, (b) image T2, (c) SAR → Op, (d) Op → SAR, (e) ground truth. From top to bottom: Gloucester I, Gloucester II, California and Shuguang datasets.

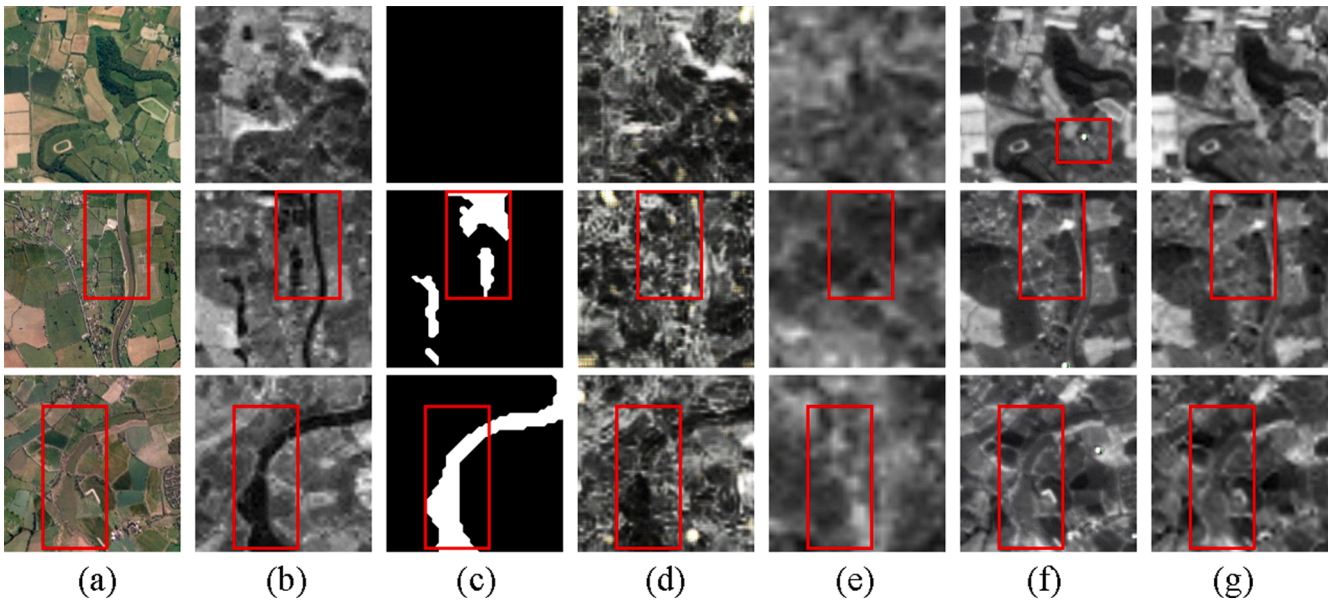

Fig. 17. Visual comparison of image translation results using different approaches on Gloucester I data set. From left to right: (a) image T1, (b) image T2, (c) ground truth, (d) pix2pix, (e) pix2pixHR, (f) Cycle-GAN, (g) NICE-GAN.

Table 10
The CD assessments of different translation methods and without translation on Gloucester I data set. Note that here results only use the mode of optical to SAR images.

Translation method	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Without translation	83.21	80.92	82.05	96.45
pix2pix	85.19	82.03	83.58	96.77
pix2pixHR	87.94	86.15	87.04	97.43
Cycle-GAN	83.81	**90.65**	87.10	97.31
NICE-GAN	**89.96**	89.93	**89.95**	**97.98**

4.2. Change detection results of Gloucester I data set

For the Gloucester I data set, the main changed pixels between the two images is that a flood disaster covers part of the ground, including farmland and buildings. The images were cropped to 256 × 256 patches of which 60% was used for training, 20% for validation and 20% for testing in the translation network. The CD network uses the same samples included original images, translated images and corresponding labels.

As shown in Table 2, compared to other methods, the CD results obtained from the Gloucester I data set of our method have the superiority on detecting changed and unchanged areas and less misclassification. Especially in F1, DTCDN has an obvious improvement which means the other methods have more error detection and missed

Table 11
Comparison of the parameter and time using standard convolution and depthwise separable convolution.

Convolution method	Trainable parameter	Inference Time(s)
Standard	9,182,466	1.28
Depthwise separable	5,266,636	1.13

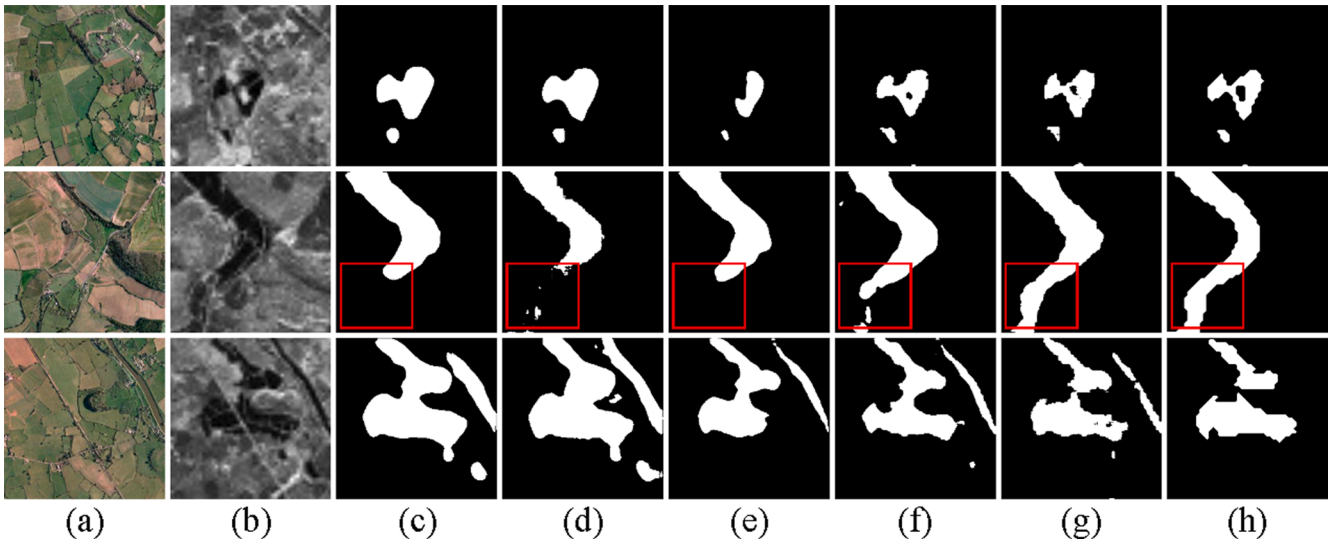

Fig. 18. Visual comparison of CD results using different approaches on Gloucester I data set. From left to right: (a) image T1, (b) image T2, (c) without translation, (d) pix2pix, (e) pix2pixHR, (f) Cycle-GAN, (g) NICE-GAN, (h) ground truth.

Table 12
Comparison of the CD results using standard convolution and depthwise separable convolution.

Dataset	Convolution method	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Gloucester I	Standard	75.51	**92.52**	83.16	96.24
	Depthwise separable	**89.96**	89.93	**89.95**	**97.98**
Gloucester II	Standard	86.88	**89.11**	87.98	95.97
	Depthwise separable	**90.78**	86.66	**88.67**	**96.33**
California	Standard	64.56	77.06	70.26	96.35
	Depthwise separable	**66.73**	**78.24**	**72.03**	**97.61**
Shuguang village	Standard	90.33	87.48	88.88	98.00
	Depthwise separable	**92.92**	**90.25**	**91.56**	**99.75**

detection. RX has a bad performance due to the lowest precision which indicates it mistakenly takes many unchanged pixels as changed pixels. Although INLPG has a good recall which over 60%, it gets the second worst performance. Most of the deep learning based methods achieve good performance, X-Net and ACE-Net with similar structures have the second and third high F1 and OA. And SCCN is the only non-GAN-based method, which ranks in the middle. Due to low precision, the overall results of cGAN are not good. It worth noting that although HPT uses the unchanged pixels as supervised in our experiment, it was worse than some unsupervised methods. Thanks to the outstanding ability of feature analysis of CD network, the method of without translation obtains the second high evaluation, while DTCDN has a much better discriminability and has the best performance in recognizing changed pixels.

To compare the results intuitively, the change binary maps of the Gloucester I data set are shown in Fig. 9. It is obvious that the DTCDN [Fig. 9 (j)] achieves the best performance. As shown in Fig. 9 (c), RX performs worst in detecting the flooding area. INLPG [Fig. 9 (d)] misses many changed pixels. The predictions of SCCN [Fig. 9 (e)] are broken and have a large amount of blur. The X-Net [Fig. 9 (g)] and ACE-Net [Fig. 9 (h)] based on adversarial encoders network achieve similar results. The result of cGAN [Fig. 9 (f)] looks like the third best result in general. There are many discontinuous small noise trips in the predictions of HPT [Fig. 9 (i)]. In detail, for Rows 1–4, the DTCDN [Fig. 9 (k)] model can detect more blank holes in the flooding areas. In addition, an unchanged building and a road are included in the red frame of Row 5 and 6. Since its characteristics are similar to floods in SAR remote sensing images, most other methods regard them as changed areas. Generally, DTCDN has the best performance for detecting the shape of the changed area completely and accurately.

4.3. Change detection results of Gloucester II data set

The event that took place in this dataset is the same as Gloucester I. The processing of images and labels can refer to Gloucester I. Besides, we observed that the SAR image of this dataset contains some small noise which may affect our final CD results.

The metrics of heterogeneous CD results are displayed in Table 3. As Table 3 is shown, similar to the first dataset, our DTCDN significantly outperforms other detection methods. Different from the Gloucester I data sets, the supervised HPT is better than other unsupervised methods. RX still achieves the worst performance than the baseline due to the drawback of the simple assumption of images relationship. And in unsupervised methods, X-Net, ACE-Net and cGAN based on deep learning obtain a higher score than INLPG. SCCN achieves the second worst results in recognizing changed pixels. Benefited from the closer image features, DTCDN achieves higher accuracy than without translation.

There are some binary CD results of all the baselines are displayed in Fig. 10. As we can observe, except for DTCDN [Fig. 10 (k)], other methods are prone to lose most change pixels (Row 1 and 2) and detect the extra areas belonging to the unchanged class (Row 3 and 4). DTCDN [Fig. 10 (k)] can cope with the two problems well, however, some relatively small regions that were taken as unchanged belonged to the changed regions in the ground truth.

4.4. Change detection results of California data set

As stated previously, a flood occurred in the California data set. Inconsistent with the other two data sets, the SAR images in this data set are synthesized from multi-polarized VV and VH data, which brings us challenges. The images and labels are also cropped to 256 × 256 patches, and the split method of dataset is same as the above two datasets.

The CD results of the California data set are shown in Table 4 and our DTCDN also achieves the highest value for most evaluations. RX still has a worst performance in this dataset. HPT achieves the third high score of F1 and OA. Besides, INLPG performs best in recall, however, it has low precision, which means it is prone to take unchanged pixels for changed pixels. Similarly, ACE-Net has only slightly worse recall than DTCDN, and obtains bad results in precision. There are very similar precision and recall of cGAN and X-Net. The result shows that the translation affects the final CD, and without translation is lower than DTCDN in all evaluations. It is worth noting that the proposed method significantly outperforms the eight comparative methods.

The change binary maps of our proposed methods and all comparison methods are shown in Fig. 11. From the first and second row of Fig. 11, it can be seen that INLPG [Fig. 11 (d)], SCCN [Fig. 11 (e)] and ACE-Net [Fig. 11 (h)] connect the white changed areas in the part that should be separated, which indicates they contain more false-positive pixels. As shown in Rows 3–6 of Fig. 11, the red frames contain unchanged pixels whose color and texture are similar to changed pixels on both optical and SAR images. DTCDN [Fig. 11 (k)] can avoid taking these pixels as changed pixels, while the rest methods all recognize them as changed pixels. Besides, HPT [Fig. 11 (i)] ignores many small changed objects in our results. In this data set, the DTCDN method still gets the best result and does not cause too many misclassifications.

4.5. Change detection results of Shuguang village data set

For the Shuguang village data set, it includes building reconstruction and a changed river. After cropped in 256 × 256, because its size is small, a cross-validation approach was adopted for the two main change areas.

The CD results of the Shuguang village data set are shown in Table 5. SCCN has the worst performance of all criteria. X-Net and ACE-Net are better than cGAN because of significantly higher precision. HPT and INLPG both achieve good results in recall almost 90%, however, they misdetect some changed pixels which leads to a decrease in F1. What is noteworthy is that the DTCDN model achieves over 90% of F1. In terms of OA, since the unchanged pixels dominate the images, all algorithms have high OA. Moreover, DTCDN performs the best.

To qualitatively assess our results, the prediction maps of all methods are shown in Fig. 12. Intuitively, the results of DTCDN [Fig. 12 (k)] are more consistent with the ground truth, and there are more correctly classified pixels that mean fewer false-positive samples. In the building area, although DTCDN [Fig. 12 (k)] has the best performance in detecting the boundary of the building, it exists some holes that cannot be recognized completely.

4.6. Heterogeneous vs. homogeneous

To comprehensively evaluate DTCDN, another two images collected by the same sensors from Gloucester II at the almost same time were introduced, as Fig. 13 shown.

The homogenous optical and SAR images are trained by our CD network part separately to compare the DTCDN more objectively. The

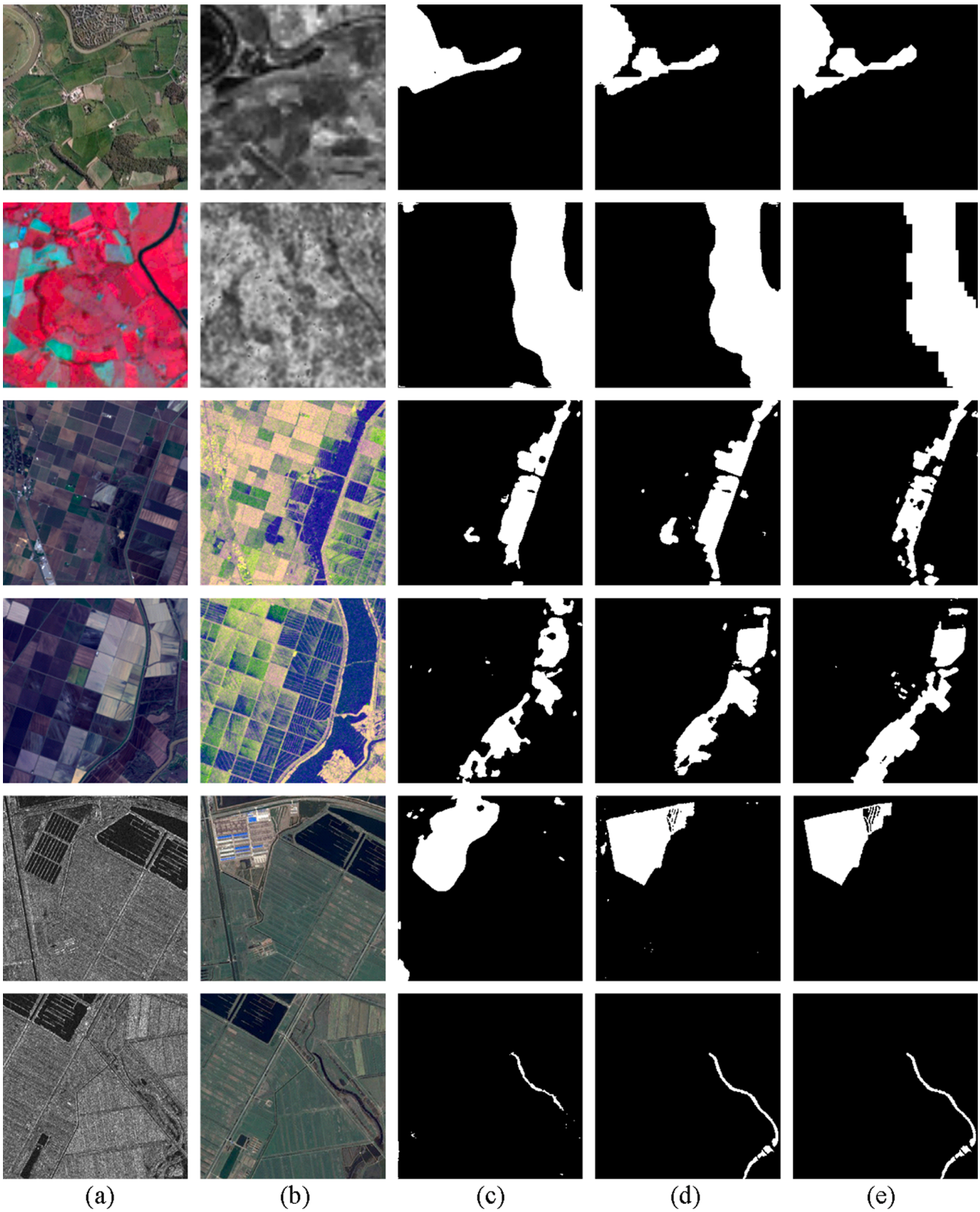

Fig. 19. Visual comparison of CD results using different convolutions methods for four data sets. From left to right: (a) image T1, (b) image T2, (c) standard convolution, (d) depthwise separable convolution, (e) ground truth.

Table 13
Comparison of the CD results using direct computation, MSOF and multi-scale loss.

Dataset	Loss computation	Precision (%)↑	Recall (%)↑	F1 (%)↑	OA (%)↑
Gloucester I	Direct computation	84.44	85.63	85.03	96.79
	MSOF	86.36	**91.40**	88.81	97.69
	Multi-scale loss	**89.96**	89.93	**89.95**	**97.98**
Gloucester II	Direct computation	80.17	91.13	85.30	94.80
	MSOF	85.87	**88.58**	87.21	95.69
	Multi-scale loss	**90.78**	86.66	**88.67**	**96.33**
California	Direct computation	63.85	74.33	68.69	95.19
	MSOF	**67.41**	76.32	71.59	96.23
	Multi-scale loss	66.73	**78.24**	**72.03**	**97.61**
Shuguang village	Direct computation	87.53	83.39	85.41	96.52
	MSOF	**94.31**	81.34	87.35	97.84
	Multi-scale loss	92.92	**90.25**	**91.56**	**99.75**

quantitative assessments of three image pairs CD are shown in Table 6. As we can see that the optical image pair achieves the best performance that is what we expected. However, to our surprise, most of the evaluation metrics of heterogeneous image pair are better than homogeneous SAR images. In particular, compared to SAR image pair, the heterogeneous image pair yields an improvement of 5.51%, 3.07%, 0.85% for recall, F1 and OA. We think that it may be caused by the quality of SAR images, which proves that our method can improve the image quality and reduce some errors, while translating SAR image.

The predictions of the final detection are presented in Fig. 14. Here we only show the T1 and T2 optical image pairs [Fig. 14 (a) and (b)] because they are enough to show the change areas. Obviously, Fig. 14 (c) has the best result because the optical image pairs provide efficient spectral and spatial information. Our DTCDN [Fig. 14 (e)] has some false positives and false negatives that need to be improved in the future. The experiment result proves that our method can be close to the results of either optical or SAR images, which helps us to realize change detection when homogeneous images cannot be available.

5. Discussion

In this section, the factors of our DTCDN model are further analyzed. The effects of the main structures of our deep translation and CD network are discussed. On this basis, the solutions to the CD of optical and SAR remote sensing images are also explored.

5.1. Effect of deep translation

The four data sets were translated by two modes from optical images to SAR images and SAR images to optical images. The quantitative values of the image translation are shown in Table 7. As can be seen, the FID of four data sets are all under 200, and the best translation effect is the Gloucester II data set from optical to SAR. However, its SAR image translated to optical image is the worst. This may be influenced by the original SAR image, which contains a certain amount of speckle noise. The conversion result of California data set better than Gloucester I data set. And both performances of two different ways on the Shuguang dataset are not bad. Additionally, there are some differences between the two translation modes for the four data sets. For California, Gloucester I and II data set, the generated SAR images are closer to the reference. This is because the optical images include more information and are more complicated than the SAR images. The generation of optical images is a more difficult process. Conversely, the Shuguang village data set is better for converting to optical images. The reason may be the small amount of this data set.

Note that the Op → SAR and SAR → Op mean the translation from optical images to SAR images and from SAR images to optical image, respectively. The examples of the visual translation results of the four data sets are shown in Fig. 15. The translated images can accurately match the texture information and some color information of the original images, but the style is closer to referenced images. For the changed pixels, the translated image retains a large difference, while in the non-change areas, it can almost show the same feature as the reference. In the red area that should have been flooded shown in Fig. 15 (a), although it cannot be converted into a water object directly, it was translated to the shape of a pit which is inconsistent with the surrounding features. Besides, in the red frame of Fig. 15 (d), the building in the original optical image varies largely in the translated optical image, which can be recognized more easily.

FID and KID represent the similarity between the translated images and the real images. However, for CD it is not to say that closer means better. So the two directions of the translation experiments were carried on these four datasets. As Table 8 is shown, Gloucester I, California and Shuguang achieve better performance on the translation of optical to SAR images, because they are easy to generate the better quality of fake SAR images. By contrast, the SAR to fake optical image has a better result on Gloucester II dataset. We speculate that CD will be affected by the noise of the original image, and the accuracy will be lower when the original image is fuzzy.

There are some CD examples of two translation modes shown in Fig. 16. For the Gloucester I dataset, the optical to fake SAR image can detect more flood areas, which causes a high recall. The third row shows a lot of buildings are covered by flood in the California dataset. It reflects that the fake SAR images do not lead to misdetect and the shape is more consistent with the ground features. For the Shuguang dataset of last row, the prediction of fake optical image is unable to extract the precise shape of the building. From Table 8 and Fig. 16, we can conclude that although the distance can measure the similarity of the generated image and reference image, it cannot absolutely explain the effect of the final change detection.

Image translation has been used in the field of unsupervised heterogeneous remote sensing images CD. However, it is the first time to explore the possibilities in which the translation result is used as the input of the CD network. Therefore, the current image translation methods were compared with the NICE-GAN and Gloucester I data set was used as the case study for brevity. FID and KID of image translation results of pix2pix, pix2pixHR, Cycle-GAN and NICE-GAN are shown in Table 9. It is found that the NICE-GAN, which combines the classifier and encoder in the discriminant phase, has lower FID and KID for all two modes. The pix2pix and pix2pixHR do not have a cyclic structure, which needs double training time and may affect the efficiency of translation. What's more, the pix2pix lacks a multiscale method, which we think is the reason for the lowest accuracy. Cycle-GAN and pix2pix have a higher FID and KID in the mode that SAR images translate to optical images than another conversion mode. On the contrary, pix2pixHR and Nice-GAN have a better performance in translating to SAR images.

In order to compare these four methods more intuitively, the visual translations from optical to SAR are shown in Fig. 17. Visually, Nice-GAN [Fig. 17 (g)] and Cycle-GAN [Fig. 17 (f)] both achieve a pretty good effect on transferring style, and the original characteristics are retained. pix2pix [Fig. 17 (d)] is the worst with noise and blurry effect. Besides, because pix2pixHR [Fig. 17 (e)] is more suitable for generating higher resolution images, the translated images are too smooth. Although the image generated by Cycle-GAN [Fig. 17 (f)] is similar to the NICE-GAN [Fig. 17 (g)], it has salt noise as shown in the red frame of the first row, which may influence the CD result. Moreover, the second and third rows in Fig. 17 show some image pairs containing changed areas between optical and SAR images. The changed flooded pixels in translated SAR image of NICE-GAN [Fig. 17 (g)] do not show the flood characteristics, which leads to discrimination between original SAR images and translated images. This phenomenon demonstrates image translation can not only reduce the difference of unchanged pixels but

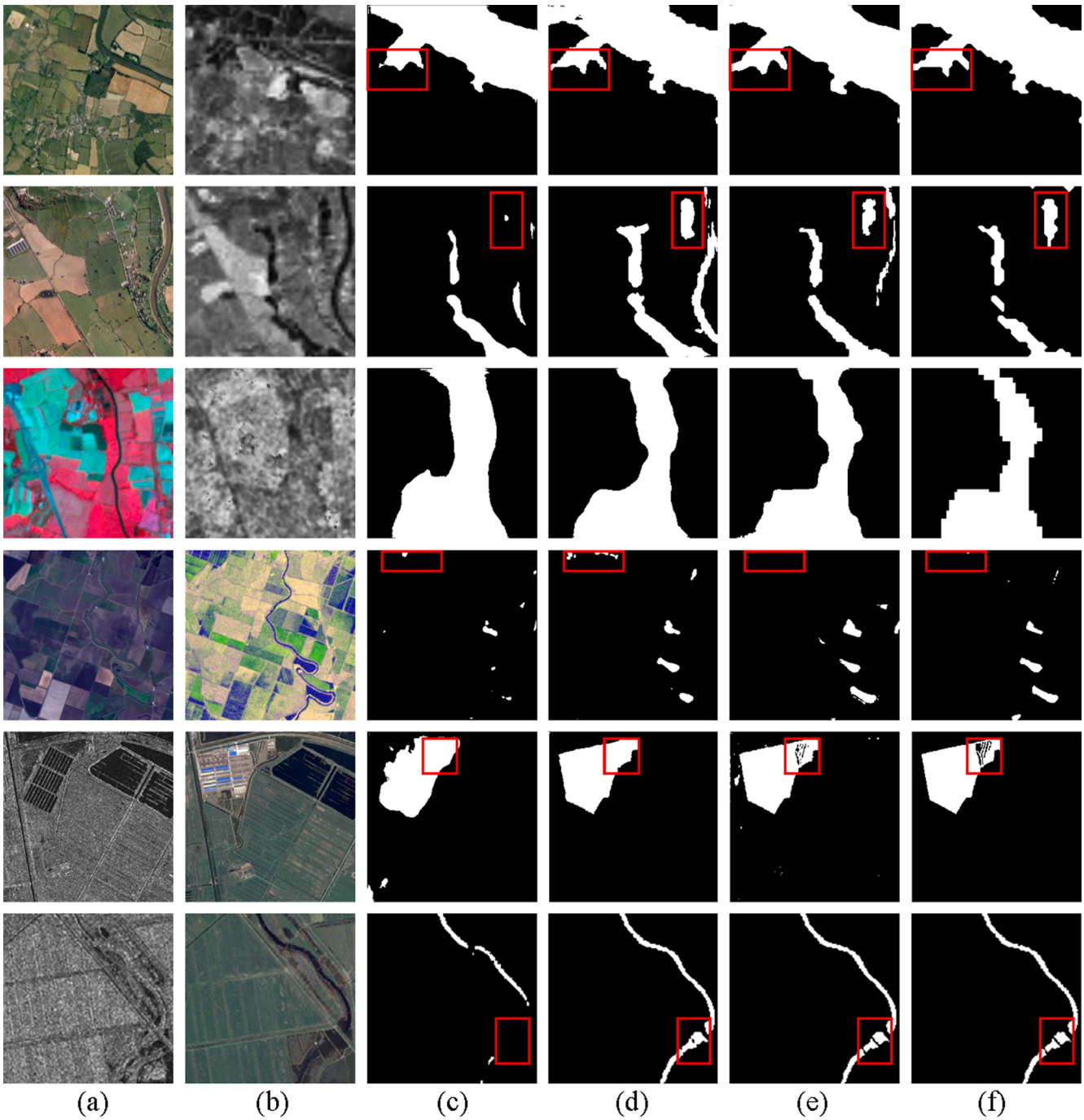

Fig. 20. Visual comparison of CD results using different loss computation methods for four data sets. From left to right: (a) image T1, (b) image T2, (c) direct computation, (d) MSOF, (e) multi-scale loss, (f) ground truth.

also maintain a certain distinction of changed pixels. The results of image translation inspire us that a robust translation model can generate more effective input of the CD network.

More experiments on the translation from optical images to SAR images were conducted to observe the image translation actual effect on the result of CD. The CD results of different deep image translation methods are shown in Table 10. It is obvious that the CD can be improved after translation. Through Nice-GAN conversion, it is increased by 0.06 in precision and 0.09 in recall which causes the F1 to improve 0.07. As expected, pix2pix has the worst effect of the four translation methods. The results of pix2pixHR are closer to the Cycle-GAN. And Cycle-GAN is slightly better than NICE-GAN in recall, while the NICE-GAN has the best performance in precision, F1 and OA.

The CD maps with different translation models are shown in Fig. 18. Obviously, NICE-GAN [Fig. 18 (g)] is better than other methods, which has small FID and KID in translation. Although Cycle-GAN [Fig. 18 (f)] has a pretty good result, it cannot detect some changed areas completely in the red frame of Row 2. Similarly, the pix2pixHR [Fig. 18 (e)] exists a lot of leak detection. Compare with no translation [Fig. 18 (c)], the pix2pix [Fig. 18 (d)] just improves a little because the generated images have a lot of noise. To some extent, the accurate prediction results prove that image translation is practicable in CD of optical and SAR images. Meanwhile, the results show that with the improvement of image translation methods, the effect of optical and SAR remote sensing images CD will be enhanced.

5.2. Effect of depthwise separable convolution for U-Net++

In this paper, the depthwise separable convolution is proposed to replace the standard convolution. Table 11 shows the results of standard convolution and depth separable convolution in the CD network. As mentioned above, the depth separable convolution can reduce the number of model parameters and improve performance. The result proves that the number of parameters of the depth separable convolution is only 52.7 M, which is 40% less than the standard convolution with the parameter amount of 91.8 M. Fewer parameters make the training dataset fit better, and the inference time is decreased by 0.15 s for every image.

The depth separable convolution defaults to an assumption that the standard convolution kernel has a decomposition characteristic similar to a linear combination in the channel dimension of the feature map. The standard convolution kernel needs to learn the spatial and channel correlation at the same time. The deep separable convolution separates the two correlations explicitly. The quantitative evaluations on the CD results of the two convolution methods are shown in Table 12. For the four data sets, the most evaluations obtain improvements through depthwise separable convolution. Especially on the Gloucester I data set, F1 almost increases by 6.5%. In this dataset, the standard convolution has a slightly better recall than the depthwise separable method, while the depthwise separable convolution outperforms largely in precision. However, for the Gloucester II data set, the gains are not as prominent, with the largest F1 gain smaller than 1%. And in the other two datasets, our model is higher than standard convolution in all criteria.

The binary change maps of two convolution methods are shown in Fig. 19. It is clear that the detection of binary maps using deep separable convolution [Fig. 19 (d)] achieve better visual performance than the standard convolution [Fig. 19 (c)]. In detail, our algorithm can detect the edge of the changed areas more accurately, and has a better extraction effect for small ground objects, such as the river in the last row.

5.3. Effect of multi-scale loss for U-Net++

In the initial U-Net++, the loss function of depth supervision is calculated by concatenating all the features of the top layer except the first one (Zhou et al. 2018). The MSOF based deep supervision method is also used in remote sensing image CD (Peng et al. 2019). However, these methods will cause the loss function to reach the minimum local prematurely, which may limit the optimal parameters. Therefore, such a loss function was not used in our model. Alternatively, the multi-scale loss is used to calculate the loss after the last feature maps obtained on different scales. Multi-scale loss aims to integrate different levels of information and improve CD results.

Table 13 shows the quantitative evaluations of direct computation, MSOF and multi-scale loss. Compared with directly computing the loss of the last feature, MSOF and multi-scale loss achieve improvement. Among all these data sets, the multi-scale loss is higher than MSOF in most of the evaluation metrics. Because of its ability to make use of contextual information and to map the features at different resolutions, it can obtain a better result in detailed and global detection.

The CD results with different loss functions are shown in Fig. 20. Visually, the overall effect of multi-scale loss has a higher similarity with ground-truth than the other two methods. From the red frames in Rows 1 and 2, some changed pixels of direct loss function are not detected, which represents a higher false-positive rate. In Row 4, although our method misses the small changed pixels, it does not lead to a more serious error that a lot of unchanged pixels were recognized as the changed pixels by the other two methods. Besides, as shown in Rows 5 and 6, the directly calculated loss has the lowest accuracy in detecting the change of building and river. Like multi-scale loss, MSOF can detect the complete boundary of the building, but it cannot detect the holes in it. In terms of results for detecting the unchanged pixels close to or in the changed areas, our algorithm has a better performance than MSOF.

6. Conclusion

Heterogeneous remote sensing images can overcome the limitations of a single data source and expand the potential of CD. In this paper, we proposed a deep translation (GAN) based CD network DTCDN for optical and SAR remote sensing images. The basic idea is to first make use of image translation to reduce the difference of heterogeneous remote sensing images, and then utilize an improved deep network to detect changes between two periods. The experimental CD results of the optical and SAR remote sensing in four different data sets demonstrate the validity of the proposed method. Compared with the previous methods, our method has the ability to significantly improve the performance of precision, recall, F1 and OA. Additionally, another two potentials were discovered in the improvements to U-Net++. (1) Depthwise separable convolution shows a satisfactory performance compared to ordinary convolution and can reduce the number of model parameters and time. However, the number of network parameters is a contradiction, too few parameters will lead to the instability of the network. Therefore, how to solve more complex CD problems while reducing the amounts of parameters is the future research work. (2) Multi-scale loss uses different levels of information to calculate the loss, which can fully combine the context of different receptive fields to enhance CD performance. Only the layers that contain long skip connections of a four-layer U-Net++ network were chosen. Because too low resolution will bring errors of small change areas after upsampling. Of course, whether there is a better way to calculate the loss function in the other CD networks remains to be studied.

The proposed method has some limitations. For example, the detection capabilities of more complex multi-source remote sensing images need to be further improved. Compared with the single-polarization data, the multi-polarization SAR image of the California dataset has more complicated characteristics and noise. The CD with more sourced and higher complexity images is a problem that deserved exploring. Moreover, our DTCDN still has some disadvantages compared to using two optical images for CD and needs to be further explored.

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.

Acknowledgement

The work was supported by the National Key R&D Program of China under Grant No. 2019YFB2102904, and the National Natural Science Foundation of China (NSFC) under Grant No. 41701394. The authors are very grateful to Prof. Max Migonotte from Montreal University for sharing the optical and SAR datasets of Gloucester I and Shuguang village. The images of the California dataset are supported by NASA's Land Processes Distributed Active Archive Center and ESA's Copernicus. The authors also would like to thank Luigi T. Luppino from UiT. The Arctic University of Norway for providing the ground truth of California datasets and the codes of comparison methods.

References

Alberga, V., 2009. Similarity measures of remotely sensed multi-sensor images for change detection applications. Remote Sensing 1 (3), 122–143. https://doi.org/10.3390/rs1030122.

Ao, D., Dumitru, C.O., Schwarz, G., Datcu, M., 2018. Dialectical gan for sar image translation: from sentinel-1 to terrasar-x. Remote Sensing 10, 1597. https://doi.org/10.3390/rs10101597.

Ashbindu, S., 1989. Review article digital change detection techniques using remotely-sensed data. Int. J. Remote Sens. 10 (6), 989–1003. https://doi.org/10.1080/01431168908903939.

Ayhan, B., Kwan, C., 2019. A new approach to change detection using heterogeneous images. In: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0192–0197. https://doi.org/10.1109/UEMCON47517.2019.8993038.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. In: International Conference on Learning Representations (ICLR), Vancouver, Canada, 1-8.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing 12 (10), 1662. https://doi.org/10.3390/rs12101662.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (ECCV), Munich, Germany, pp. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49.

Chen, R., Huang, W., Huang, B., Sun, F., Fang, B., 2020. Reusing discriminators for encoding: towards unsupervised image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, pp. 8168–8177. https://doi.org/10.1109/CVPR42600.2020.00819.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, pp. 1251–1258. https://doi.org/10.1109/CVPR.2017.195.

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

de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134 (1), 19–67. https://doi.org/10.1007/s10479-005-5724-z.

Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C., Schmitt, M., 2019. Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits. Remote Sensing 11, 2067. https://doi.org/10.3390/rs11172067.

Geng, J., Ma, X., Zhou, X., Wang, H., 2019. Saliency-guided deep neural networks for sar image change detection. IEEE Trans. Geosci. Remote Sens. 57 (10), 7365–7377. https://doi.org/10.1109/TGRS.2019.2913095.

Guo, H., Shi, Q., Du, B., Zhang, L., Wang, D., Ding, H., 2020. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 1–20. https://doi.org/10.1109/TGRS.2020.3014312.

Guo, Q., Zhang, B., Ran, Q., Gao, L., Li, J., Plaza, A., 2014. Weighted-rxd and linear filter-based rxd: Improving background statistics estimation for anomaly detection in hyperspectral imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (6), 2351–2366. https://doi.org/10.1109/JSTARS.460944310.1109/JSTARS.2014.2302446.

Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H., 2001. Image analogies. In: 28th annual conference on Computer graphics and interactive techniques, New York, USA, pp. 327–340. https://doi.org/10.1145/383259.383295.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Conference on Neural Information Processing Systems (NIPS), Long Beach, California, USA, pp. 6629–6640.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, pp. 1125–1134. https://doi.org/10.1109/CVPR.2017.632.

Jaturapitpornchai, R., Matsuoka, M., Kanemoto, N., Kuzuoka, S., Ito, R., Nakamura, R., 2019. Newly built construction detection in sar images using deep learning. Remote Sensing 11 (12), 1444. https://doi.org/10.3390/rs11121444.

Ji, M., Liu, L., Du, R., Buchroithner, M.F., 2019. A comparative study of texture and convolutional neural network features for detecting collapsed buildings after earthquakes using pre- and post-event satellite imagery. Remote Sensing 11, 1202. https://doi.org/10.3390/rs11101202.

Jin, L., Lazarow, J., Tu, Z., 2017. Introspective classification with convolutional nets. In: Conference on Neural Information Processing Systems (NIPS), Long Beach, California, USA, pp. 823–833. https://dl.acm.org/doi/abs/10.5555/3294771.3294850.

Kampffmeyer, M., Salberg, A.-B., & Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, Nevada, USA, pp. 680–688. https://doi.org/10.1109/CVPRW.2016.90.

Khan, S.H., He, X., Porikli, F., Bennamoun, M., 2017. Forest change detection in incomplete satellite images with deep neural networks. IEEE Trans. Geosci. Remote Sens. 55 (9), 5407–5423. https://doi.org/10.1109/TGRS.2017.2707528.

Kwan, C., Ayhan, B., Larkin, J., Kwan, L., Bernabé, S., Plaza, A., 2019. Performance of change detection algorithms using heterogeneous images and extended multi-attribute profiles (emaps). Remote Sensing 11 (20), 2377. https://doi.org/10.3390/rs11202377.

Lazarow, J., Jin, L., Tu, Z., 2017. Introspective neural networks for generative modeling. In: IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2774–2783. https://doi.org/10.1109/ICCV.2017.302.

Lee, K., Xu, W., Fan, F., Tu, Z., 2018. Wasserstein introspective neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, pp. 3702–3711. https://doi.org/10.1109/CVPR.2018.00390.

Lei, L., Sun, Y., Kuang, G., 2020. Adaptive local structure consistency-based heterogeneous remote sensing change detection. IEEE Geosci. Remote Sens. Lett. 1–5 https://doi.org/10.1109/LGRS.2020.3037930.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017a. Focal loss for dense object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, pp. 2980–2988. https://doi.org/10.1109/TPAMI.2018.2858826.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017b. Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, pp. 936–944. https://doi.org/10.1109/CVPR.2017.106.

Liu, J., Gong, M., Qin, K., Zhang, P., 2018a. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans. Neural Networks Learn. Syst. 29 (3), 545–559. https://doi.org/10.1109/TNNLS.2016.2636227.

Liu, R., Jiang, D., Zhang, L., Zhang, Z., 2020a. Deep depthwise separable convolutional network for change detection in optical aerial images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1109–1118. https://doi.org/10.1109/JSTARS.460944310.1109/JSTARS.2020.2974276.

Liu, S., Shi, Q., Zhang, L., 2020b. Few-shot hyperspectral image classification with unknown classes using multitask deep learning. IEEE Trans. Geosci. Remote Sens. 1–18. https://doi.org/10.1109/TGRS.2020.3018879.

Liu, T., Yang, L., Lunga, D., 2021. Change detection using deep learning approach with object-based image analysis. Remote Sens. Environ. 256, 112308. https://doi.org/10.1016/j.rse.2021.112308.

Liu, Z., Li, G., Mercier, G., He, Y., Pan, Q., 2018b. Change detection in heterogenous remote sensing images via homogeneous pixel transformation. IEEE Trans. Image Process. 27 (4), 1822–1834. https://doi.org/10.1109/TIP.2017.2784560.

Longbotham, N., Pacifici, F., Glenn, T., Zare, A., Volpi, M., Tuia, D., Christophe, E., Michel, J., Inglada, J., Chanussot, J., Du, Q., 2012. Multi-modal change detection, application to the detection of flooded areas: outcome of the 2009–2010 data fusion contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (1), 331–342. https://doi.org/10.1109/JSTARS.460944310.1109/JSTARS.2011.2179638.

Lunetta, R.S., Knight, J.F., Ediriwickrema, J., Lyon, J.G., Worthy, L.D., 2006. Land-cover change detection using multi-temporal modis ndvi data. Remote Sens. Environ. 105 (2), 142–154. https://doi.org/10.1016/j.rse.2006.06.018.

Luppino, L.T., Anfinsen, S.N., Moser, G., Jenssen, R., Bianchi, F.M., Serpico, S., Mercier, G., 2017. A clustering approach to heterogeneous change detection. In: Scandinavian Conference on Image Analysis (SCIA), Tromsø, Norway, pp. 181–192. https://doi.org/10.1007/978-3-319-59129-2_16.

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 (12), 9960–9975. https://doi.org/10.1109/TGRS.2019.2930348.

Luppino, L.T., Kampffmeyer, M., Bianchi, F.M., Moser, G., Serpico, S.B., Jenssen, R., & Anfinsen, S.N., 2020. Deep image translation with an affinity-based change prior for unsupervised multimodal change detection. arXiv:2001.04271 1-16.

Lyu, H., Lu, H., Mou, L., Li, W., Wright, J., Li, X., Li, X., Zhu, X.X., Wang, J., Yu, L., Gong, P., 2018. Long-term annual mapping of four cities on different continents by applying a deep information learning method to landsat data. Remote Sensing 10, 471. https://doi.org/10.3390/rs10030471.

Mercier, G., Moser, G., Serpico, S.B., 2008. Conditional copulas for change detection in heterogeneous remote sensing images. IEEE Trans. Geosci. Remote Sens. 46 (5), 1428–1441. https://doi.org/10.1109/TGRS.2008.916476.

Mignotte, M., 2020. A fractal projection and markovian segmentation-based approach for multimodal change detection. IEEE Trans. Geosci. Remote Sens. 58 (11), 8046–8058. https://doi.org/10.1109/TGRS.2020.2986239.

Mubea, K., Menz, G., 2012. Monitoring land-use change in Nakuru (Kenya) using multi-sensor satellite data. Adv. Remote Sensing 1, 74–84. https://doi.org/10.4236/ars.2012.13008.

Nguyen, T.L., Han, D., 2020. Detection of road surface changes from multi-temporal unmanned aerial vehicle images using a convolutional siamese network. Sustainability 12, 2482. https://doi.org/10.3390/su12062482.

Niu, X., Gong, M., Zhan, T., Yang, Y., 2019. A conditional adversarial network for change detection in heterogeneous images. IEEE Geosci. Remote Sens. Lett. 16 (1), 45–49. https://doi.org/10.1109/LGRS.2018.2868704.

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.

Planinsic, P., Gleich, D., 2018. Temporal change detection in sar images using log cumulants and stacked autoencoder. IEEE Geosci. Remote Sens. Lett. 15 (2), 297–301. https://doi.org/10.1109/LGRS.2017.2786344.

Prendes, J., Chabert, M., Pascal, F., Giros, A., Tourneret, J.-Y., 2015. A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors. IEEE Trans. Image Process. 24 (3), 799–812. https://doi.org/10.1109/TIP.2014.2387013.

Reed, I.S., Yu, X., 1990. Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. IEEE Trans. Acoust. Speech Signal Process. 38, 1760–1770. https://doi.org/10.1109/29.60107.

Ronneberger, O., Fischer, P., & Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (MICCAI), Munich, Germany, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Saha, S., Bovolo, F., Bruzzone, L., 2018. Destroyed-buildings detection from vhr sar images using deep features. In: the XXIVth Image and Signal Processing for Remote Sensing Berlin, Germany. https://doi.org/10.1117/12.2325149.

Shang, R., He, J., Wang, J., Xu, K., Jiao, L., Stolkin, R., 2020. Dense connection and depthwise separable convolution based cnn for polarimetric sar image classification. Knowl.-Based Syst. 194, 1–12. https://doi.org/10.1016/j.knosys.2020.105542.

Shi, Q., Liu, M., Liu, X., Liu, P., Zhang, P., Yang, J., Li, X., 2020. Domain adaption for fine-grained urban village extraction from satellite images. IEEE Geosci. Remote Sens. Lett. 17 (8), 1430–1434. https://doi.org/10.1109/LGRS.885910.1109/LGRS.2019.2947473.

Sublime, J., Kalinicheva, E., 2019. Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the tohoku tsunami. Remote Sensing 11, 1123. https://doi.org/10.3390/rs11091123.

Sun, Y., Lei, L., Li, X., Sun, H., Kuang, G., 2021a. Nonlocal patch similarity based heterogeneous remote sensing change detection. Pattern Recogn. 109, 107598. https://doi.org/10.1016/j.patcog.2020.107598.

Sun, Y., Lei, L., Li, X., Tan, X., Kuang, G., 2020. Patch similarity graph matrix-based unsupervised remote sensing change detection with homogeneous and heterogeneous sensors. IEEE Trans. Geosci. Remote Sens. 1–21. https://doi.org/10.1109/TGRS.2020.3013673.

Sun, Y., Lei, L., Li, X., Tan, X., Kuang, G., 2021b. Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images. IEEE Trans. Geosci. Remote Sens. 1–21 https://doi.org/10.1109/TGRS.2021.3053571.

Tian, J., Cui, S., Reinartz, P., 2014. Building change detection based on satellite stereo imagery and digital surface models. IEEE Trans. Geosci. Remote Sens. 52 (1), 406–417. https://doi.org/10.1109/TGRS.2013.2240692.

Tong, X., Zhang, X., Liu, M., 2010. Detection of urban sprawl using a genetic algorithm-evolved artificial neural network classification in remote sensing: A case study in jiading and putuo districts of shanghai, china. Int. J. Remote Sens. 31 (6), 1485–1504. https://doi.org/10.1080/01431160903475290.

Touati, R., Mignotte, M., 2017. An energy-based model encoding nonlocal pairwise pixel interactions for multisensor change detection. IEEE Trans. Geosci. Remote Sens. 56, 1046–1058. https://doi.org/10.1109/TGRS.2017.2758359.

Touati, R., Mignotte, M., Dahmane, M., 2017. A new change detector in heterogeneous remote sensing imagery. In: International Conference on Image Processing Theory, Tools and Applications (IPTA), Montreal, Canada, 1–6. https://doi.org/10.1109/IPTA.2017.8310138.

Touati, R., Mignotte, M., Dahmane, M., 2018. Change detection in heterogeneous remote sensing images based on an imaging modality-invariant mds representation. In: IEEE International Conference on Image Processing (ICIP), Athens, Greece, pp. 3998–4002. https://doi.org/10.1109/ICIP.2018.8451184.

Touati, R., Mignotte, M., Dahmane, M., 2019. Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based markov random field model. IEEE Trans. Image Process. 29, 757–767. https://doi.org/10.1109/TIP.2019.2933747.

Touati, R., Mignotte, M., Dahmane, M., 2020. Partly uncoupled siamese model for change detection from heterogeneous remote sensing imagery. J. Remote Sensing GIS 9, 1–8.

Turnes, J.N., Castro, J.D.B., Torres, D.L., Vega, P.J.S., Feitosa, R.Q., Happ, P.N., 2020. Atrous cgan for sar to optical image translation. IEEE Geosci. Remote Sens. Lett. 1–5 https://doi.org/10.1109/lgrs.2020.3031199.

Wan, L., Xiang, Y., You, H., 2019. A post-classification comparison method for sar and optical images change detection. IEEE Geosci. Remote Sens. Lett. 16 (7), 1026–1030. https://doi.org/10.1109/LGRS.885910.1109/LGRS.2019.2892432.

Wang, Q.i., Yuan, Z., Du, Q., Li, X., 2019. Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection. IEEE Trans. Geosci. Remote Sens. 57 (1), 3–13. https://doi.org/10.1109/TGRS.2018.2849692.

Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 8798–8807. https://doi.org/10.1109/CVPR.2018.00917.

Zhou, J., Kwan, C., Ayhan, B., Eismann, M.T., 2016. A novel cluster kernel rx algorithm for anomaly and change detection using hyperspectral images. IEEE Trans. Geosci. Remote Sens. 54 (11), 6497–6504. https://doi.org/10.1109/TGRS.2016.2585495.

Zhou, W., Troy, A., Grove, M., 2008. Object-based land cover classification and change analysis in the baltimore metropolitan area using multitemporal high resolution remote sensing data. Sensors 8, 1613–1636. https://doi.org/10.3390/s8031613.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2223–2232. https://doi.org/10.1109/ICCV.2017.244.