Super-Resolution-Based Change Detection Network With Stacked Attention Module for Images With Different Resolutions

Mengxi Liu, *Student Member, IEEE*, Qian Shi, *Senior Member, IEEE*, Andrea Marinoni, *Senior Member, IEEE*, Da He, *Member, IEEE*, Xiaoping Liu, *Member, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— Change detection (CD) aims to distinguish surface changes based on bitemporal images. Since high-resolution (HR) images cannot be typically acquired continuously over time, bitemporal images with different resolutions are often adopted for CD in practical applications. Traditional subpixel-based methods for CD using images with different resolutions may lead to substantial error accumulation when the HR images are employed, which is because of intraclass heterogeneity and interclass similarity. Therefore, it is necessary to develop a novel method for CD using images with different resolutions that are more suitable for the HR images. To this end, we propose a super-resolution-based change detection network (SRCDNet) with a stacked attention module (SAM). The SRCDNet employs a super-resolution (SR) module containing a generator and a discriminator to directly learn the SR images through adversarial learning and overcome the resolution difference between the bitemporal images. To enhance the useful information in multiscale features, a SAM consisting of five convolutional block attention modules (CBAMs) is integrated to the feature extractor. The final change map is obtained through a metric learning-based change decision module, wherein a distance map between bitemporal features is calculated. Ablation study and comparative experiments on two large datasets, building change detection dataset (BCDD) and season-varying change detection dataset (CDD), and a real-image experiment on the Google dataset fully demonstrate the superiority of the proposed method. The source code of SRCDNet is available at https://github.com/liumency/SRCDNet.**

***Index Terms*— Change detection (CD), fully convolutional networks (FCNs), metric learning, remote sensing images, super-resolution.**

I. Introduction

CHANGE detection (CD) aims to identify surface changes in bitemporal images of the same area and provides quantitative data for various important applications [1], such as land and resource investigation, ecological monitoring and protection, and urban planning [2]–[4]. Because of their ability to provide rich information on wide areas of earth surface with high temporal efficiency, remote sensing images have been widely used in these applications for a long time [5].

In the past few decades, multispectral satellite images have been extensively employed in remote sensing applications [6], [7], including CD [8]. Therefore, traditional CD methods, such as change vector analysis (CVA) [9] and principal component analysis (PCA) [10], mainly exploit the spectral information in bitemporal images. However, as the spectral and spatial resolutions of an image are mutually restricted, multispectral satellite images usually have a low spatial resolution, making it difficult to achieve precise CD. Recently, with the rapid development of remote sensing technology, high-resolution (HR) images with rich spatial information have become the main data source for CD, especially for fine-grained scenarios such as urban renewal [11]. In order to fully exploit the opportunities provided by the HR images, advanced techniques for remote sensing data analysis have been explored and proposed in technical literature. Deep learning-based methods that include a powerful feature learning structure, namely convolutional neural networks (CNNs), to extract spatial and semantic features from HR images hierarchically [12] have provided a remarkable solution for HR image CD. Examples of these methods include classification-based methods [13]–[15] and metric learning-based methods [16], [17].

Although HR images might help in improving the characterization of the phenomena occurring on earth surface, it is also true that several meteorological and technical effects—such as small observation range, low temporal resolution, and the effect of clouds and fog—might limit the actual capacity and impact of the investigation of the HR images. Hence, the use of traditional CD methods relying on the analysis of bitemporal images showing same resolution properties

Manuscript received February 13, 2021; revised May 17, 2021; accepted June 19, 2021. Date of publication July 2, 2021; date of current version January 14, 2022. This work was supported in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X355, in part by the Guangdong Natural Science Foundation under Grant 2019A1515011057, in part by the National Natural Science Foundation of China under Grant 61976234, in part by the Open research fund of National Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, in part by the Guangzhou Applied Basic Research Project, in part by the Center for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA), and in part by the Research Council of Norway (RCN) under Grant 237906. *(Corresponding author: Qian Shi.)*

Mengxi Liu, Qian Shi, Da He, and Xiaoping Liu are with the Guangdong Provincial Key Laboratory for Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: liumx23@mail2.sysu.edu.cn; shixi5@mail.sysu.edu.cn; heda@mail.sysu.edu.cn; liuxp3@mail.sysu.edu.cn).

Andrea Marinoni is with the Department of Physics and Technology, UiT the Arctic University of Norway, 9019 Tromsø, Norway, and also with the Department of Engineering, University of Cambridge, Cambridge CB2 1TN, U.K. (e-mail: andrea.marinoni@uit.no).

Liangpei Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3091758

might not be adequate to tackle the main issues of remote sensing-based CD, especially on a large-scale scenario. For example, let us suppose that we have obtained an HR image of a certain area at time T1, but only low-resolution (LR) images corresponding to time T2 are available; then, we need to detect the changes that occurred between T1 and T2 using the bitemporal images with different resolutions. Therefore, to realize large-scale and rapid CD, it is often necessary to use bitemporal images with different resolutions for real-life applications [18]–[20].

To tackle these issues, the most intuitive method is to simply down-sample the HR image to the resolution of the LR image [21], or to interpolate the LR image to the resolution of the HR image to obtain bitemporal images with the same resolution [22] and then employ common CD methods to detect changes. However, the down-sampling step at the core of the first approach induces a lack of detailed spatial information of the outcomes, which leads to a strong degradation of the precision of the obtained results. On the other hand, the second approach does not take into account semantic information in common interpolation operations (such as linear, bilinear, and bicubic interpolation) applied to remote sensing images, which leads to a scarce capacity to achieve detailed information on the region of interest.

In addition to the simplest interpolation methods mentioned above, other methods have been proposed to solve the problem of CD with remote sensing images of different resolutions: in this context, the subpixel-based method is the most prevalent. Considering the excellent ability of subpixel mapping (SPM) to obtain fine-resolution land-cover maps from coarse-resolution images [23]–[25], Ling *et al.* [26] first introduced SPM to CD using images with different resolutions (henceforth referred to as "different-resolution CD" for brevity) using the spatial dependence principle and a novel land-cover change (LCC) rule to obtain the spatial pattern of LCC maps at the subpixel scale. Later, Wang *et al.* [27] proposed a Hopfield neural network with SPM to overcome the resolution difference between Landsat and Moderate-Resolution Imaging Spectroradiometer (MODIS) images for subpixel-resolution LCCs. Li *et al.* [28] applied an iterative super-resolution CD method for Landsat-MODIS CD, which combines end-member estimation, spectral unmixing, and SPM. Wu *et al.* [29] proposed a back propagation neural network to obtain subpixel LCC maps from the soft-classification results of LR images.

These SPM-based methods obtain fine-resolution change maps by establishing a mapping between a former fine-resolution land-cover map and a coarse-resolution image and have been proved to be effective in dealing with large-scale differences for remote sensing image CD, especially on Landsat and MODIS images. However, in these cases, the accuracy of fine-resolution change maps is largely limited by the accuracy of the former fine-resolution land-cover map, leading to redundant error accumulation. Such a problem would be much more severe for HR images due to the intraclass heterogeneity and interclass similarity in HR images. Therefore, there is an urgent need to develop more precise CD methods for HR images with different spatial resolutions.

In this article, we propose an end-to-end super-resolution-based change detection network (SRCDNet) for different-resolution change detection. To deal with the unmatched spatial resolution of bitemporal images, the SRCDNet employs a super-resolution module (SRM) to learn a super-resolution (SR) image directly from the LR images to recover more semantic information and avoid redundant errors. The SR image is then input into a feature extractor together with the HR images corresponding to other timestamps. To fully extract the multilevel information in the HR images, so as to facilitate the subsequent prediction, a stacked attention module (SAM) consisting of five convolutional block attention modules (CBAMs) is also added to the feature extractor. Then, in order to learn precise change maps from the multiscale features of the bitemporal images, the distance map between the features is calculated and compared with the ground truth, where a contrastive loss, a common loss in metric learning, is adopted to help increase the distance of the changed area and decrease that of the unchanged area. Finally, the change map can be obtained from the distance map through simple thresholding. The main contributions of this study can be summarized as follows:

1) We provide an end-to-end super-resolution-based network for HR image CD; the proposed scheme learns the SR image through the mapping between the LR image and the initial HR image to avoid the error accumulation encountered in traditional subpixel-based methods.
2) We integrate a SAM consisting of five CBAM blocks into the feature extractor of the network to enhance valid information in the hierarchical features for more distinguishable feature pairs, which can greatly help subsequent change decision through metric learning.
3) Comparative experiments on two common change detection datasets (CDDs), building change detection dataset (BCDD) [30] and CDD [31], were conducted to verify the effectiveness of the SRCDNet, and a real-image dataset based on Google dataset [32] was also constructed to further test the SRCDNet on real images. The results showed that the proposed model could not only achieved the state-of-the-art (SOTA) performance on simulated different-resolution images in BCDD and CDD, but also obtained highest accuracy on real images.

The remainder of this article is organized as follows: Section II further presents some related works. Section III presents detailed information regarding the proposed network. Section IV elucidates the settings of all the experiments conducted in the study. In Section V, we present our results and detailed analysis. Section VI discusses the effects of different settings on the model. Finally, we conclude this article in Section VII.

II. Related Works

A. Deep Learning-Based CD

After the proposal of a fully convolutional network (FCN) [33] provided a more intuitive method for dense prediction, many methods based on FCNs and their variants, especially U-Net [34], have been proposed for pixel-wise CD.

Daudt *et al.* [14] explored three different methods of image inputs, including early fusion, Siamese difference, and Siamese concatenation, based on U-Net for bitemporal CD. To fully utilize both global and local information in bitemporal images, Peng *et al.* [35] proposed U-Net++ with a multiscale feature fusion strategy to generate final change maps. The encoder–decoder structure of U-Net is commonly used in semantic segmentation tasks, where the encoder is used to extract multilevel semantic features of bitemporal images, and the decoder is used to recover spatial information from the hierarchical features and generate CD maps by classification.

In recent years, some studies have introduced metric learning into CD to replace the decoder's upsampling process, which directly obtains change maps by calculating the distance between the features of the bitemporal images. During the training process, the distance between "unchanged" features is minimized, while that between "changed" features is maximized; the loss function plays an important role in this process. For example, Zhang *et al.* [36] used an improved triplet loss to learn the semantic relation between multiscale information in paired features. Wang *et al.* [37] employed contrastive loss to detect changes based on features extracted using a Siamese convolutional network. To mitigate the effects of class imbalance in CD, Chen and Shi [16] employed batched contrastive loss to train the proposed spatial–temporal attention-based network (STANet).

B. CD Strategies

While HR images critically lead to false alarms or missed alarms due to the influence of intraclass heterogeneity and interclass similarity, many attempts have been made to generate more discriminative features, including recurrent neural networks (RNNs) [38] and attention mechanisms. Papadomanolaki *et al.* [13] integrated long short-term memory blocks (LSTMs) [39] into an FCN to detect urban changes in Sentinel-2 images and proved that RNNs are effective for capturing spectral or temporal relationships between images. In addition, Song *et al.* [40] combined a convolutional LSTM with a 3-D FCN to capture joint spectral–spatial–temporal features in hyperspectral images. Although RNNs can work well on multispectral and hyperspectral inputs, they are still limited because of infrequent spectral information in HR images and deficient time information in bitemporal images.

Owing to their ability to enhance useful information in extracted features, attention mechanisms have been adopted to make better use of the abundant spatial information in HR images. Chen and Shi [16] integrated a self-attention module in the feature extractor of a CD network to strengthen the spatial–temporal relationships between bitemporal features. Chen *et al.* [17] used dual CBAMs [41] for each bitemporal feature to emphasize change information in images, facilitating subsequent metric learning-based change decisions. However, because of memory limitations, in the existing methods in which spatial information is effectively exploited for CD, the attention mechanism is usually only applied to high-level semantic features, whereas the enhancement of shallow features is ignored.

C. Super-Resolution

Since the quality of image plays a vital role in many visual applications, super-resolution aims to obtain higher quality image from the LR image. In recent years, the prosperity of deep learning has brought new solutions to image super-resolution, especially for single image super-resolution (SSIR). To better restore the detailed information of an image when reconstructing HR images from LR images, Dong *et al.* [42] first introduced CNNs into SR applications as super-resolution convolutional neural network (SRCNN). After that, a series of effective SR algorithms have been successively proposed by researchers. By combining an SRCNN and a Visual Geometry Group (VGG) backbone, Kim *et al.* [43] designed a deep neural network called very deep super-resolution (VDSR) with 20 layers. Kim *et al.* [44] also proposed a deep recursive convolutional network (DRCN) for SR. In view of the excellent performance of generative adversarial networks (GANs) [45] in various other fields, Ledig *et al.* [46] applied a GAN to SR and proposed super-resolution generative adversarial network (SRGAN), which achieved SOTA performance at the time.

Combined with the above researches, we propose a deep metric learning change detection network (CDNet), integrating super-resolution to learn the mapping from LR image to HR image, and attention mechanism to obtain more effective multiscale features, for different-resolution CD.

III. Methodology

In this section, a brief overview of the proposed method is provided, followed by the detailed description of each part and the optimization process of the model.

A. Overview

The SRCDNet, as shown in Fig. 1, consists of two parts: a super-resolution (SR) module and a CD module. Based on GAN architecture, the SR module, consisting of a generator and a discriminator, aims to reconstruct the LR images of the bitemporal images to the HR images. The CD module is responsible for feature extraction and CD. A SAM with five CBAMs, a lightweight attention mechanism that can enhance the features both channel-wise and spatial-wise, is integrated into the feature extractor to extract multiscale features. After this, deep metric learning is employed for subsequent change decision.

Let us assume that we need to detect changes between the HR images obtained at T1 and the LR images obtained at T2, whose resolution difference can be denoted as N ($N = 4, 8$). Given a set of HR bitemporal images at T1 and T2, referred as $I_{\mathrm{T1}}^{\mathrm{HR}}$ and $I_{\mathrm{T2}}^{\mathrm{HR}}$, respectively, and the corresponding change annotation Y. $I_{\mathrm{T2}}^{\mathrm{HR}}$ would be N times down-sampled to generate LR images at T2, which can be referred as $I_{\mathrm{T2}}^{\mathrm{LR}}$. Thus, we can obtain a training set consisting of a set of images $\left(I_{\mathrm{T1}}^{\mathrm{HR}}, I_{\mathrm{T2}}^{\mathrm{HR}}, I_{\mathrm{T2}}^{\mathrm{LR}}\right) = \left\{\left(i_{\mathrm{T1}}^{\mathrm{HR}}, i_{\mathrm{T2}}^{\mathrm{HR}}, i_{\mathrm{T2}}^{\mathrm{LR}}\right)^1, \ldots, \left(i_{\mathrm{T1}}^{\mathrm{HR}}, i_{\mathrm{T2}}^{\mathrm{HR}}, i_{\mathrm{T2}}^{\mathrm{LR}}\right)^n\right\}, n \in N$, and the ground truth $\mathrm{GT} = \{gt^1, \ldots, gt^n\}, n \in N$, then the flowchart of SRCDNet can be summarized as follows.

1) With the input of $I_{\mathrm{T2}}^{\mathrm{LR}}$ to the SR module, the generator G would produce the SR image at T2, $I_{\mathrm{T2}}^{\mathrm{SR}} = G(I_{\mathrm{T2}}^{\mathrm{LR}})$, with

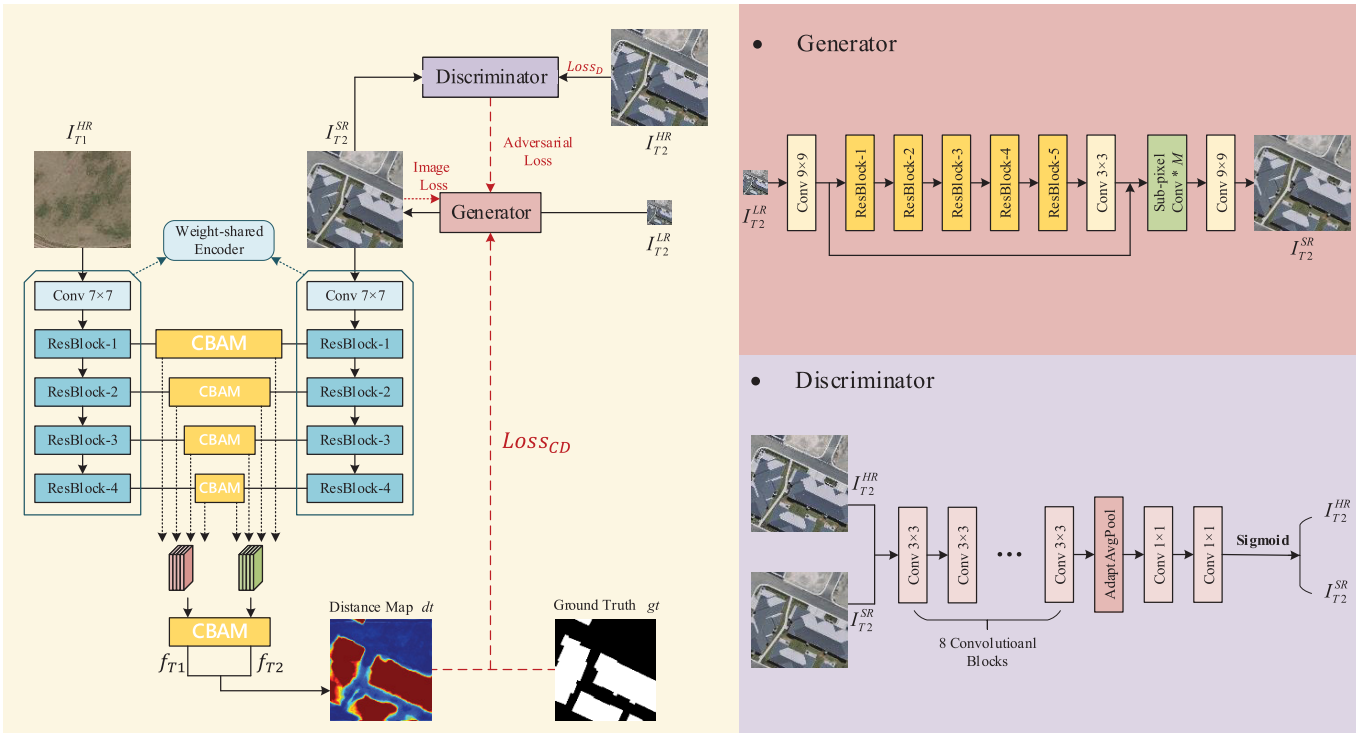

Fig. 1. Overview of the proposed SRCDNet.

the same size as I_{T2}^{HR}; after that, the discriminator D is responsible for learning to discriminate I_{T2}^{SR} from I_{T2}^{HR} through a loss Loss_D, which is constituted by the output of the discriminator, $D(I_{T2}^{HR})$ and $D(I_{T2}^{SR})$.

2) Then, I_{T1}^{HR} and I_{T2}^{SR} are both fed to the weight-shared feature extractor for hierarchical features. The first four CBAMs are applied to four intermediate features, before they are stacked into one, on which the fifth CBAM blocks would be applied.
3) Thereafter, a distance map dt is calculated based on the bitemporal features f_{T1} and f_{T2} to measure the distance between I_{T1}^{HR} and I_{T2}^{SR}, which would be compared with the ground truth gt and obtain a contrastive loss, Loss_{CD}. The metric Loss_{CD} would be then optimized to push away the distance between the changed area on the ground truth and pull in the distance between the unchanged area.
4) Finally, the generator G is optimized according to the difference between I_{T2}^{SR} and I_{T2}^{HR}, the discriminator's result, and the CD performance Loss_{CD}, in order to generate the SR images with rich semantic information.

It is worth noting that when there is no spatial resolution difference between bitemporal images, the SRCDNet can be easily degenerated into a simple CDNet by removing the SR module; this detail significantly improves the generality of the model.

B. Super-Resolution Module

The structure of the SR module is inspired by the SRGAN scheme [46], where a generator is responsible for generating the SR image from the LR image, whereas the discriminator distinguishes the SR image from the initial HR image until they are indistinguishable from each other and then the generator is able to output SR images that are sufficient for fine-grained CD.

The generator first employs a 9×9 convolutional layer to capture the shallow features of the input LR image and then five residual blocks to extract high-level features. Each residual block is composed of two 3×3 convolutional layers followed by a batch normalization (BN) layer [47], of which the first one is a parametric rectified linear unit (PReLU) layer that serves as the activation layer. The deep features further extracted by residual blocks are fused with the shallow features obtained in the first convolutional layer to obtain features with rich spatial and semantic information. Thereafter, M subpixel convolution layers are used to increase the feature size to that of the HR image. Since each subpixel convolution layer can magnify the size of the input feature by twice through a pixel shuffle operation, the value of M can be calculated by $M = \log_2 N$. Finally, the generator produces the SR image by means of a fully convolutional layer.

Inheriting the structure of VGG-19 [48], the discriminator contains eight convolutional layers, wherein BN layers and leaky ReLU functions are used. Two fully connected layers and a sigmoid function are used to output the distribution of the input image, which is a binary classification task. Because the objective of the discriminator is to distinguish the SR images from the HR images, the generator can be prompted to generate the SR samples that are more similar to the original HR image through adversarial training.

C. CD Module

While the SR module aims to produce the SR images similar to the HR images, the task of the CD module is to generate

precise change maps based on the SR image and the HR image at another timestamp. The CD module in the SRCDNet employs metric learning to obtain change maps based on features from the feature extractor.

We use a pretrained ResNet-18 [49] as the feature extractor after removing the last fully connected layers, which is extended to a Siamese structure to receive bitemporal inputs. A 7×7 convolutional layer with a stride of one is used to extract the shallow features with rich spatial information, followed by a BN layer and the ReLU function, after which a max-pooling layer with a stride of two is employed. Then, four residual blocks are employed to fully exploit the information in the images. The size of the output features of each residual block is 1/2, 1/4, 1/8, and 1/8 of the input image, and the channel of them are 64, 128, 256, and 512, respectively.

To fully capture the effective information in multiscale features, we integrated a SAM with five CBAMs in the feature extractor. More specifically, four CBAM blocks are applied to the output features of each residual block to emphasize useful information; these features are then uniformly adjusted to half the size of the original image and fused into ones with multiscale information. Thereafter, the fifth CBAM block is applied to make more distinguishable feature pairs for subsequent detection.

Each CBAM block contains two parts: a channel attention module to capture channel-wise relationship and a spatial attention module to explore spatial-wise contextual information. Given a feature F with size $C \times H \times W$, an average pooling layer and a max pooling layer are first applied on the input feature, respectively, to obtain two vectors with size $C \times 1 \times 1$ in the channel attention module, then a weight sharing multilayer perception (MLP) module with two 1×1 convolutional layers is used to learn and give weights to each channel. Finally, the two factors are summed up into one and a sigmoid function σ is applied to get the channel attention map factor, which can be expressed as

$$M_c(F) = \sigma(\mathrm{MLP}(\mathrm{Avg}(F)) + \mathrm{MLP}(\mathrm{Max}(F))). \tag{1}$$

The channel-refined feature F' is the result of multiplication of $M_c(F)$ and F, which can be denoted as

$$F' = M_c(F) \otimes F. \tag{2}$$

Thereafter, the spatial attention module would be applied on feature F' with size of $C \times H \times W$, which is the same with F. Here, an average pooling layer and a max pooling layer are utilized to squeezed the F' into two matrixes of size $1 \times H \times W$, which are then stacked into one and input into a 3×3 convolutional layer. At last, the spatial-refined matrix is obtained through a sigmoid function, which can be represented as

$$M_s(F') = \sigma\left(f^{3\times 3}\right)\left(\mathrm{Avg}(F'); \mathrm{Max}(F')\right). \tag{3}$$

Thus, the CBAM-refined feature can be obtained by the following formula:

$$F'' = M_s(F') \otimes F'. \tag{4}$$

Since CBAM does not change the size of the input features, the output feature pairs of the fifth CBAM keep the size of half of the original image, between which a Euclidean distance is calculated to measure their similarity. The distance map needs to be interpolated to the same size of the original images and then a contrastive loss is employed as the metric to weigh the disparity between the distance map and ground truth. Thereafter, through the optimization of the metric, the distance of the changed area on the ground truth was increased while that of the unchanged area was reduced. In other words, we can obtain the distance map with the value difference between the changed area and unchanged area as large as possible through metric learning. Consequently, we could obtain more precise change maps from the distance map by threshold segmentation.

D. Loss Function

There are three submodels that need to be optimized in our network: the generator, discriminator in the SR module, and change network. The objective of each submodel is provided in the following.

1) Discriminator: After receiving I_{T2}^{HR} and I_{T2}^{SR} as inputs, the discriminator outputs the probability of the input image being I_{T2}^{HR}. To improve the discriminator's ability to accurately distinguish I_{T2}^{SR} from I_{T2}^{HR}, the loss function of the discriminator is designed as follows:

$$\mathrm{Loss}_D = 1 - D\left(I_{T2}^{HR}\right) + D\left(G\left(I_{T2}^{LR}\right)\right). \tag{5}$$

According to the formula, the discriminator is bound to output a probability close to 1 for an HR image and a probability close to 0 for an SR image after adversarial training.

2) Discriminator: Then, I_{T2}^{SR} is fed to the CDNet together with I_{T1}^{HR}, where the multiscale features of the bitemporal inputs are extracted using the Siamese feature extractor. A Euclidean distance map between the feature pairs is calculated based on the feature pairs enhanced by the CBAM block, and based on this map, the final change map is generated using threshold segmentation. Therefore, the objective of the CDNet is to draw the corresponding values of "changed" areas and "unchanged" areas on the distance map as much as possible according to the ground truth. Thus, a batch contrastive loss is used to help minimize the distance between "unchanged" areas and maximize the distance between "changed" areas on the distance map, which can be denoted as follows:

$$\mathrm{Loss}_{CD} = \sum_{i,j=0}^{M} \frac{1}{2}\left[\left(1 - gt_{i,j}\right)dt_{i,j}^2 + gt_{i,j}\max\left(dt_{i,j} - m\right)^2\right] \tag{6}$$

where M denotes the size of dt; $dt_{(i,j)}$ and $gt_{(i,j)}$ represent the values of the distance map and ground truth map at point (i, j), respectively, where $i, j \in [0, M)$; and m is the margin to filter out pixels whose values exceed the threshold, which is set to be 2 in the experiment.

3) Generator: The generator loss consists of the following losses: the image loss, the content loss, the adversarial loss, and the change loss. The image loss measures the alignment of I_{T2}^{SR} and I_{T2}^{HR} in the pixel-wise space by calculating the mean

square errors (MSEs) between them. The image loss can be expressed as

$$l_{\mathrm{MSE}} = \sum_{i,j=0}^{M} \left(I_{\mathrm{T2}\,i,j}^{\mathrm{HR}} - G\left(I_{\mathrm{T2}}^{\mathrm{LR}}\right)_{i,j} \right)^2. \tag{7}$$

Since preserving detailed information in an image is difficult as a result of pixel alignment, content loss focuses more on perceptual similarity. Specifically, content loss computes the MSEs between certain feature maps of $I_{\mathrm{T2}}^{\mathrm{SR}}$ and $I_{\mathrm{T2}}^{\mathrm{HR}}$ obtained by the pretrained VGG-19 network to yield a more visually realistic SR image. The formula for content loss is

$$l_{\mathrm{MSE}}^{\mathrm{VGG}} = \sum_{i,j=0}^{M} \left(\phi_{\mathrm{VGG}}\left(I_{\mathrm{T2}}^{\mathrm{HR}}\right)_{i,j} - \phi_{\mathrm{VGG}}\left(G\left(I_{\mathrm{T2}}^{\mathrm{LR}}\right)\right)_{i,j} \right)^2. \tag{8}$$

While the discriminator is designed to discriminate $I_{\mathrm{T2}}^{\mathrm{SR}}$ from $I_{\mathrm{T2}}^{\mathrm{HR}}$, the generator aims to increase the probability of the discriminator's misjudgment through adversarial loss, which can be defined as follows:

$$l_D = 1 - D\left(G\left(I_{\mathrm{T2}}^{\mathrm{LR}}\right)\right). \tag{9}$$

To make the SR image pixel-wise and perceptually similar to the original HR image and to improve at the same time the CD results, the batch contrastive loss employed in the CDNet is also added to the loss of the generator. In summary, the optimization objective of the generator is

$$\mathrm{Loss}_G = l_{\mathrm{MSE}} + \alpha l_{\mathrm{MSE}}^{\mathrm{VGG}} + \beta l_D + \lambda \mathrm{Loss}_{\mathrm{CD}} \tag{10}$$

where α, β, and λ are factors that balance different losses.

The flowchart of SRCDNet is further described in Algorithm 1.

Algorithm 1 Flowchart of SRCDNet

Input:

A set of image pairs $(I_{\mathrm{T1}}^{\mathrm{HR}}, I_{\mathrm{T2}}^{\mathrm{HR}}, I_{\mathrm{T2}}^{\mathrm{LR}})$ and corresponding ground truth gt;

Parameters including resolution difference N, learning rate lr, batch size n, training epochs M

Output:

a distance map dt

for each $m \in [1, M]$ **do**

generate $I_{\mathrm{T1}}^{\mathrm{SR}}$ from $I_{\mathrm{T2}}^{\mathrm{LR}}$;

extract multiscale features from $I_{\mathrm{T1}}^{\mathrm{HR}}$ and $I_{\mathrm{T1}}^{\mathrm{SR}}$;

obtain bitemporal feature maps, f_{T1} and f_{T2};

calculate the distance map dt based on f_{T1} and f_{T2};

calculate Loss_D by (5);

optimized the discriminator D according to Loss_D;

calculate the contrastive loss $\mathrm{Loss}_{\mathrm{CD}}$ by (6);

optimized the CD module according to $\mathrm{Loss}_{\mathrm{CD}}$;

calculate Loss_G by (7)–(10);

optimized the generator G according to Loss_G;

end for

IV. Experimental Settings

A. Datasets

Three CDDs are used in our experiments, which can be summarized as follows.

1) Building Change Detection Dataset (BCDD): The BCDD [29] provides pairs of 0.2-m images with a size of 32 507 × 15 354 and a ground truth for the building changes between them. Because the bitemporal images were selected before and after earthquake occurrence, the areas contain various building changes, including building reconstruction and renewal. For the convenience of model training and testing, we cropped the images into 7434 patches of size 256 × 256 without overlapping, which were then randomly divided into training, validation, and testing sets in a ratio of 8:1:1. To avoid overfitting, we rotated the images in the training set for data augmentation.

2) Change Detection Dataset (CDD): The CDD [30] contains 16 000 real season-varying Google Earth image pairs with a size of 256 × 256, including 10 000 training samples, 3000 validation samples, and 3000 testing samples. With a very high spatial resolution of 3–100 cm, the CDD not only provides change information of common objects, including buildings, roads, and forests, but also that of many detailed objects, such as cars and tanks. While both the BCDD and CDD are change detection datasets with the same-resolution bitemporal images, further processing is needed for different-resolution CD experiments. Therefore, the T2 images in BCDD and CDD were $N(N = 4, 8)$ times bicubically down-sampled to obtain corresponding LR images for subsequent experiments. Examples of different-resolution samples in BCDD and CDD are shown in Fig. 2.

3) Google Dataset: The Google dataset consists of 19 pairs of Google Earth images in Guangzhou, China, with a resolution of 0.55-m and the size ranging from 1006 × 1168 pixels to 4936 × 5224. All of the images are collected between 2006 and 2019, when Guangzhou was experiencing a period of rapid development. Therefore, the dataset provides building changes of various shapes and sizes. In our experiment, for all of the 19 images at T2 in the dataset, we obtained corresponding images of the same date with a resolution of 2.2-m from Google Earth, so as to construct a real-image dataset with 4 times resolution difference. Thereafter, the real-image dataset was split into 1118 sample pairs without overlapping and divided into three parts according to a ratio of 6:2:2 for training, validating, and testing. The training set was further augmented by random rotation and flip. Examples of different-resolution samples in the reconstructed Google dataset are shown in Fig. 3.

B. Experimental Design

To fully verify the effectiveness of the proposed SRCDNet, three groups of experiments were designed.

1) Ablation Study: An ablation study was first conducted on BCDD and CDD to test the validity of different modules in SRCDNet. Since the proposed SRCDNet aims to solve the resolution difference between bitemporal images in CD,

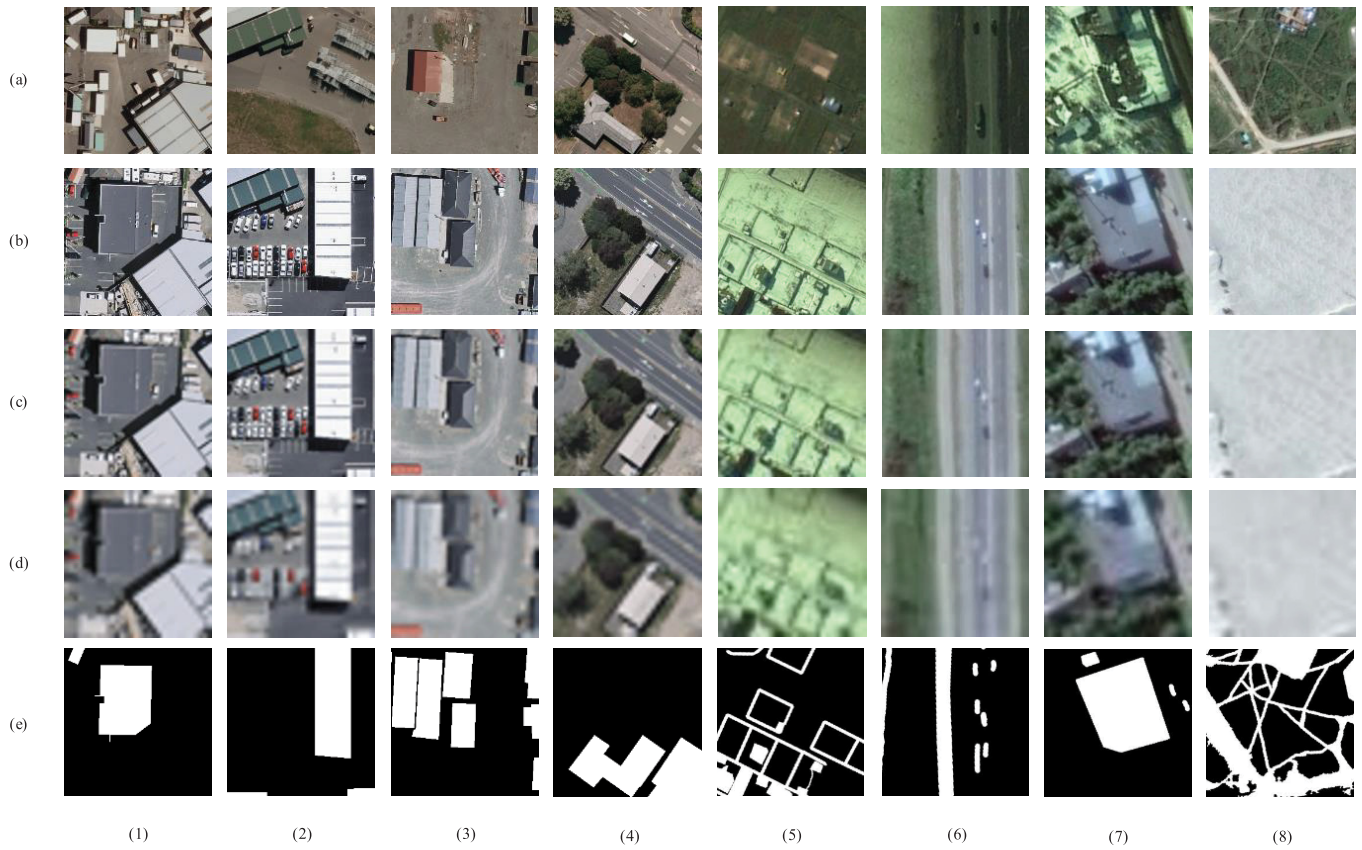

Fig. 2. Examples of different-resolution samples in BCDD (1)–(4) and CDD (5)–(8). (a) HR image at time T1. (b) HR image at time T2. (c) 4× LR image at time T2. (d) 8× LR image at time T2. (e) Ground truth.

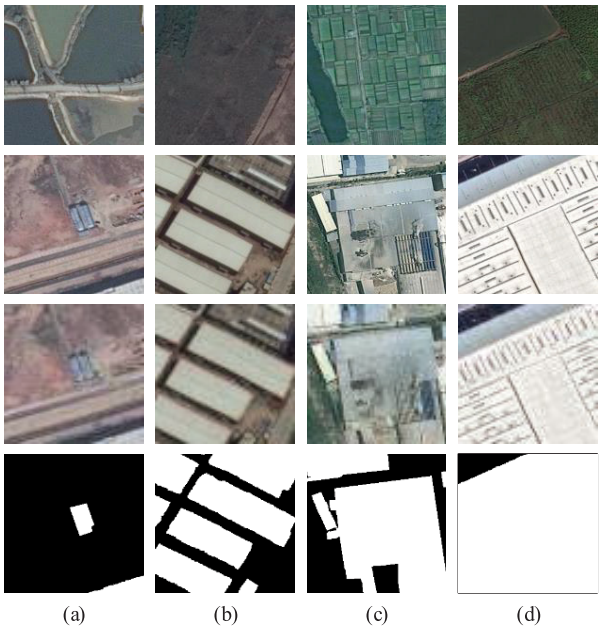

Fig. 3. Examples of different-resolution samples in the Google dataset. (a) HR image at time T1. (b) HR image at time T2. (c) 4× LR image at time T2. (e) Ground truth.

experiments on four times resolution difference (X4) and eight times resolution difference (X8) were conducted in this part.

2) Comparative Experiments: Comparative experiments with other CD methods were then carried out on the BCDD and CDD to examine the advance of SRCDNet. While the SRCDNet can also be transformed into a common CD model, except for experiments on X4 and X8, experiments on the initial images with no-resolution-difference (X1) were also involved to inspect the flexibility of the model.

3) Real-Image Experiments: In order to further investigate the utility of SRCDNet on real images, a real-image experiment was made on the Google dataset with four times resolution difference. In addition, we also transferred the model trained on the BCDD and CDD to the Google dataset, so as to explore the application potential of existing large CD datasets on real images.

C. Baselines

To validate the proposed SRCDNet, five SOTA CD methods were introduced into the experiments for comparison. The brief description of each method is provided below.

1) Fully Convolutional–Early Fusion (FC-EF) [30]: Based on the U-Net architecture, the FC-EF network detects changes by fusing the bitemporal images as a multispectral input. Skip connections are adopted to transfer features from the encoder to the decoder to recover spatial information at each level.

2) Fully Convolutional–Siamese Difference (FC-Siam-Diff) [30]: A variant of FC-EF, the FC-Siam-diff network extends the encoder to a Siamese structure to receive bitemporal inputs and extract their features separately. Features of the same layer in the Siamese encoder are transmitted

to the decoder through skip connections after the difference operation.

3) Fully Convolutional–Siamese Concatenation (FC-Siam-Conc) [14]: FC-Siam-conc also adopts the same Siamese encoder as FC-Siam-diff and features of the same level in the Siamese encoder are concatenated before being transferred to the decoder, rather than employing the difference operation.

4) BiDateNet [14]: BiDateNet integrates LSTM blocks into the skip connections of an FCN with U-Net architecture to learn the temporal dependence between bitemporal images to help detect changes.

5) Spatial–Temporal Attention-Based Network (STANet) [16]: STANet is a metric learning-based CDNet that provides a spatial–temporal attention module to further exploit spatial information and temporal relationships in the features.

Noted that all of the above baselines require the same-resolution bitemporal inputs, thus the LR images in the X4 and X8 experiments would be bicubically interpolated to the original image size, so as to be used as the input for the comparative methods.

D. Implementations

We implemented the SRCDNet with PyTorch Libraries. For a total of 100 training epochs, an Adam optimizer with an initial rate of 0.0001 was utilized to facilitate model convergence. A batch size of eight was adopted during training. BN and Dropout layer are both adopted to avoid over-fitting. The α, β and λ in the loss function of the generator are set as 0.006, 0.001, and 0.001, respectively. During the training process, the accuracy of the latest model on the validation set will be calculated at each epoch, so as to save the best model in time. Besides, early stopping is adopted when the accuracy on the validation set does not increase for 50 epochs. All of the baselines were run on a GeForce RTX 2080ti graphics card to achieve higher training efficiency.

Four commonly used factors are employed to measure the CD performance of different methods: precision, recall, F1-score, and IoU. Given that TP, FP, TN, and FN refer to true positives, false positives, true negatives, and false negatives, respectively, precision and recall can be defined as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

According to the formulas, precision represents the false alarm rate, whereas recall represents the missed alarm rate, both of which entail a tradeoff. Thus, to obtain more comprehensive evaluations, the F1 score combines both precision and recall and can be expressed as follows:

$$F1 = \frac{2\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{13}$$

The IoU refers to the intersection and union rate between the detection result and the ground truth, which can be intuitively represented as

$$\text{IoU} = \frac{\text{TP}}{\text{FP} + \text{TP} + \text{FN}}. \tag{14}$$

Moreover, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are employed as the index to measure the quality of restored images through bicubic interpolation and super-resolution. PSNR is the most widely used index to assess the image quality, which is calculated based on MSEs. Given a gray image X and the reference image Y, the MSE between X and Y can be denoted as

$$\text{MSE} = \frac{1}{hw} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [X(i, j) - Y(i, j)]^2 \tag{15}$$

where h and w are the height and width of X and Y, respectively.

Then the PSNR of image X can be calculated by the following formula:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}(X)^2}{\text{MSE}} \right) \tag{16}$$

where $\text{MAX}(X)$ denotes the maximum value in X. The larger the value of PSNR, the closer X is to Y. In our experiment, the RGB images need to be converted into the gray ones for the calculation of PSNR.

As can be seen from the above formula, PSNR only focuses on the proximity of X and Y in pixel values, ignoring the visual effect of the image. To make up for this deficiency, SSIM is often used as an auxiliary, which takes both luminance (l), contrast (c), and structure (s) into account. The three components can be calculated as

$$l(X, Y) = \frac{2\mu_X \mu_Y + c_1}{{\mu_X}^2 + {\mu_Y}^2 + c_1} \tag{17}$$

$$c(X, Y) = \frac{2\sigma_X \sigma_Y + c_2}{{\sigma_X}^2 + {\sigma_Y}^2 + c_2} \tag{18}$$

$$s(X, Y) = \frac{2\sigma_{XY} + c_2}{2\sigma_X \sigma_Y + c_2} \tag{19}$$

where μ and σ are the operation of mean and variance, respectively, and c_1 and c_2 are constants to avoid division by zero.

Then the SSIM between X and Y can be denoted as

$$\begin{aligned} \text{SSIM}(X, Y) &= l(X, Y) \cdot c(X, Y) \cdot s(X, Y) \\ &= \frac{(2\mu_X \mu_Y + c_1)(2\sigma_{XY} + c_2)}{\left(\mu_X^2 + \mu_Y^2 + c_1\right)\left(\sigma_X^2 + \sigma_Y^2 + c_2\right)}. \end{aligned} \tag{20}$$

The range of SSIM is $[0, 1]$, and the larger the value of SSIM, the closer X is to Y.

V. Results and Analysis

A. Ablation Study

With the aim to verify the effectiveness of the SAM and SRM, we first conducted the ablation study on SRCDNet on two resolution differences (X4 and X8) on BCDD and CDD. The SRCDNet without SAM and SRM was set as the Base model. Then, the SAM was added to the Base model as the second baseline, and the SRM was added to the Base model as the third baseline. The image produced by the bicubic interpolation of the LR image was used as the input because the first two baselines do not contain an SRM.

TABLE I
ABLATION STUDY ON SRCDNET

Dataset	Baseline	X4				X8			
		Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)
BCDD	Base	67.20	**91.63**	77.53	63.31	62.63	85.40	72.26	56.57
	Base+SAM	80.43	88.06	84.07	72.52	70.85	84.84	77.22	62.89
	Base+SRM	72.72	88.42	79.81	66.40	65.73	**87.77**	75.17	60.22
	SRCDNet	**84.44**	86.90	**85.66**	**74.91**	**81.61**	81.78	**81.69**	**69.05**
CDD	Base	90.03	83.35	86.56	76.31	87.08	70.62	77.99	63.92
	Base+SAM	**93.31**	84.66	88.77	79.82	91.41	70.97	79.91	66.54
	Base+SRM	93.03	82.65	87.53	77.83	88.85	73.74	80.59	67.49
	SRCDNet	92.07	**88.07**	**90.02**	**81.86**	**91.95**	**76.03**	**83.24**	**71.29**

1) Ablation Study on X4 Images: As shown in Table I, the Base model without any additions performs the worst on both datasets, with the lowest F1 values of 77.53% on the BCDD and 86.56% on the CDD. With the integration of the SAM, the F1 on the BCDD is significantly increased to 84.07% and that on the CDD also is increased to 88.77%. Notably, the precision rates of the second baseline are greatly improved compared with those of the Base model, which shows that the SAM can help extract changes more accurately. Compared with the Base model, the third baseline that includes the SRM also obtained better detection results, with an F1 of 79.81% on the BCDD and an F1 of 87.53% on the CDD. Therefore, SRCDNet, which integrates both the SAM and SRM into the Base model, outperforms all baselines, with the highest F1 values of 85.66% on the BCDD and 90.02% on the CDD.

As shown in Fig. 4, the performance of the Base model is poor on both datasets. More specifically, the building changes in the BCDD have an obvious spillover effect, whereas the changes corresponding to the CDD are substantially missed. The second baseline can solve the above problems and greatly improve the detection accuracy owing to the SAM. With the inclusion of the SRM to generate fine-grained SR images for CD, the results of the third baseline have more precise change boundaries. However, this is not accurate enough because of the poor detection performance of the Base model. Therefore, by combining the advantages of the SAM and SRM, SRCDNet can obtain the most precise change maps among all the baselines on the two datasets, based on the output SR images.

2) Ablation Study on X8 Images: The ablation study on 8× LR images on the BCDD showed the same tendency as that on 4× LR images. While the Base model could only obtain an F1 of 72.26%, the SAM-integrated model is able to boost the F1 to 77.22%, and the SRM-integrated model can improve it to 75.17%. SRCDNet far surpasses the above baselines, with an F1 of 81.69% and an IoU of 69.05%, which fully demonstrates the feasibility of combining SAM and SRM.

The performance of the Base model on the CDD is the worst among all baselines, with an F1 of 77.99%. Whereas SAM integration is more effective than SRM integration in the previous experiments, the opposite effect is observed on the CDD for 8× LR images. More specifically, the F1 of the second baseline with the SAM increased to 79.91%, whereas that of the third baseline increased to 80.59% owing to the addition of the SRM. This may be because the SRM can enhance the extraction of small changes in the images in the CDD by effectively recovering image information. Nevertheless, SRCDNet has the highest F1 of 83.24%, which is much higher than that of the other baselines.

According to Fig. 5, compared with the Base and Base+SAM models based on bicubic images, the Base+SRM model and SRCDNet can better capture the building boundaries on the BCDD owing to the integration of the SRM. Moreover, they can significantly alleviate missed alarms owing to SR images with more detailed information. Because the information of small objects in the CDD is difficult to recover from bicubic images, the change results obtained on the CDD by the Base and Base+SAM models also entail several missed alarms. Although it is difficult to restore the 8× LR images, with the help of the SRM, the Base+SRM and SRCDNet can yield more comprehensive change results.

3) Comparisons on Different Restored Images: According to the ablation study presented above, two different methods, bicubic interpolation and SR, can be utilized to restore detailed information from the LR images. While both the Base+SRM model and SRCDNet have adopted the SRM to generate the SR image from the LR image, there are three sets of restored images for CD for each group of experiments. Hence, we employ two common indices, namely PSNR and SSIM, to compare the effects of different restored images quantitatively and further understand the effect of the different modules in SRCDNet.

As can be seen from Table II, the bicubic image obtains the lowest PSNR and SSIM in each comparative experiment. Furthermore, the indices of the SR image obtained using the Base+SRM model are improved to a certain extent, which illustrates that the SRM can produce restored images with higher quality compared with bicubic interpolation. The SR image obtained using SRCDNet achieves the highest PSNR and SSIM in all experiments, with slightly higher metrics than those of the SR image obtained using the Base+SRM model. This further proves that the combination of SAM and SRM has a positive effect on CD, as mentioned before.

Another noteworthy phenomenon is that in experiments on the X4 images, the integration of the SRM leads to a greater increase in the PSNR and SSIM. Hence, more image information can be recovered through the SRM compared with the case of the X8 experiment. This shows that the greater the resolution difference, the greater the loss of image information, which causes substantial difficulties in image restoration.

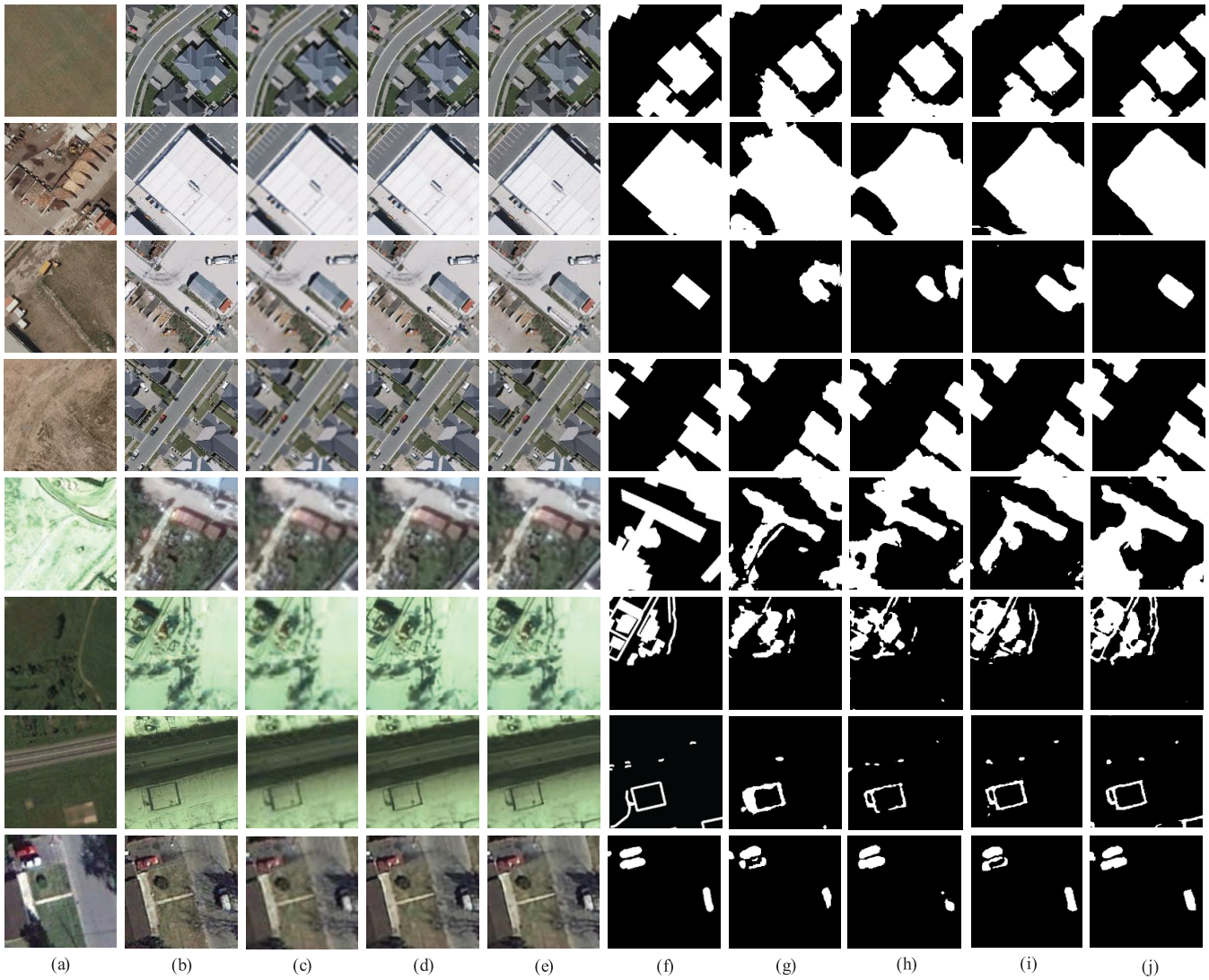

Fig. 4. Ablation study on X4 images (BCDD: Rows 1–4; CDD: Rows 5–8). (a) HR image at time T1. (b) HR image at time T2. (c) 4× Bicubic image at time T2. (d) 4× SR image at time T2 by Base+SRM. (e) 4× SR image at time T2 by SRCDNet. (f) Ground truth. (g) Base. (h) Base+SAM. (i) Base+SRM. (j) SRCDNet.

TABLE II
COMPARISONS ON DIFFERENT RESTORED IMAGES

Restored Images	Index	BCDD		CDD	
		X4	X8	X4	X8
Bicubic Image	PSNR SSIM	21.87, 0.5051	20.25, 0.3774	28.95, 0.7571	25.23, 0.6408
SR Image (Base+SRM)	PSNR SSIM	23.04, 0.5920	21.03, 0.4322	29.63, 0.7811	25.83, 0.6580
SR Image (SRCDNet)	PSNR SSIM	**23.09**, **0.5929**	**21.04**, **0.4335**	**29.88**, **0.7818**	**25.89**, **0.6582**

B. Comparative Experiments

1) Performance on X1 Images: As shown in Table III, our proposed method outperforms all the baseline methods on the BCDD with the highest recall, F1, and IoU of 90.13%, 87.40%, and 77.63%, respectively, and a very high precision of 84.84%. STANet is ranked second with an F1 and IoU of 84.96% and 73.86%, which are 2.44% and 3.77% lower than those of our method, respectively. BiDateNet obtains the best results among the U-Net-based methods, with an F1 and IoU of 83.55% and 71.75%, respectively, thus proving the feasibility of RNNs for enhancing the time relationship in bitemporal images. FC-Siam-diff performs the best among the three FCN variants, followed by FC-Siam-conc and FC-EF.

On the CDD, our proposed method and STANet again obtain the best performance, which further proves the superiority of metric learning-based methods compared with traditional classification-based methods. While our proposed method achieves the highest F1 and IoU of 92.94% and 86.81%, STANet obtains a relatively lower F1 and IoU of 91.44% and 84.23%, respectively, which shows the effectiveness of the SAM. The third-ranked BiDataNet has the highest precision of 95.98%. In contrast to the results on the BCDD, FC-Siam-conc performs better than FC-Siam-diff, with increases in F1 and IoU of 2.97% and 4.44%, respectively. This may be attributed to the fact that because the CDD contains various fine-grained objects, much useful

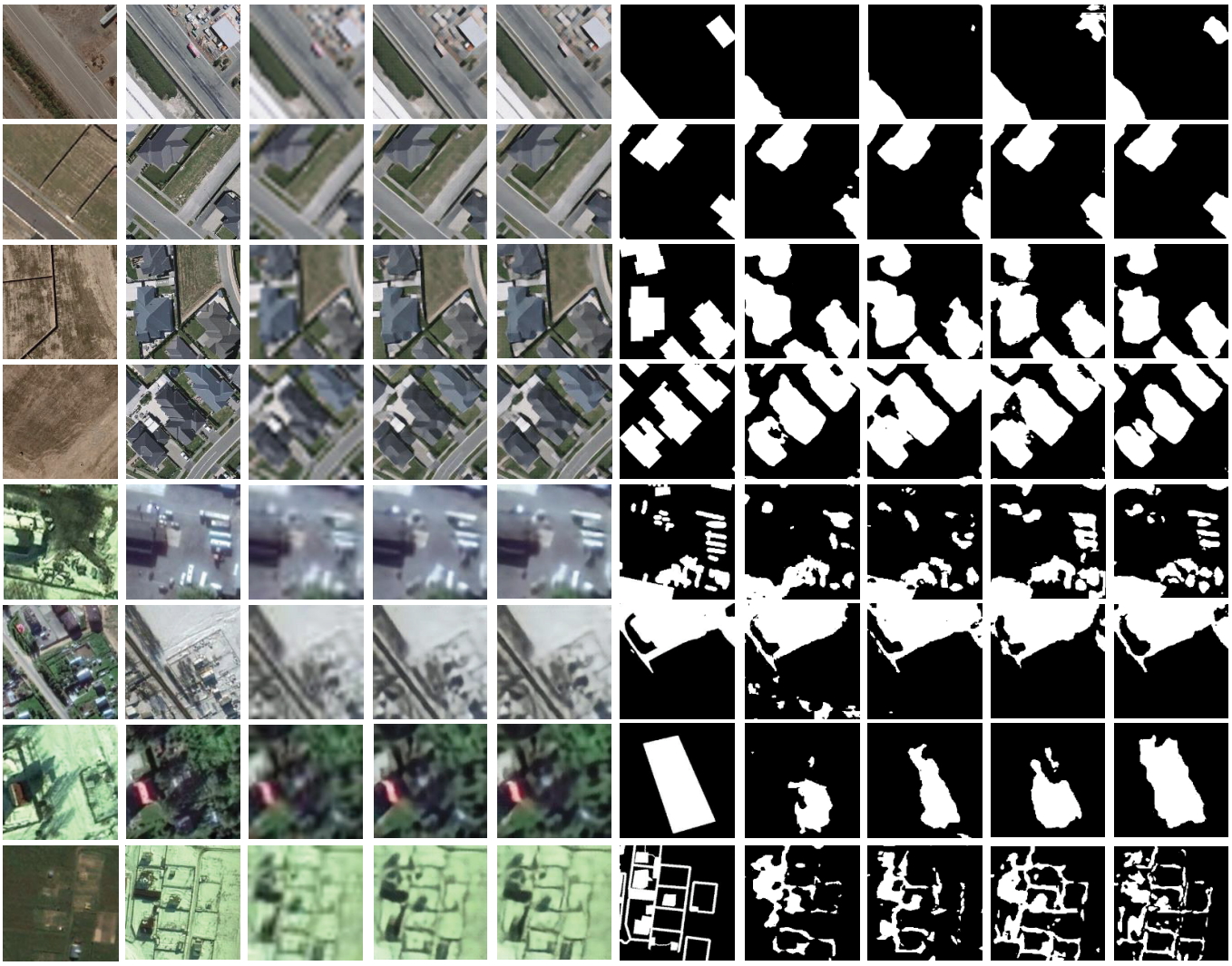

Fig. 5. Ablation study on X8 images (BCDD: Rows 1–4; CDD: Rows 5–8). (a) HR image at time T1. (b) HR image at time T2. (c) 8× Bicubic image at time T2. (d) 8× SR image at time T2 by Base+SRM. (e) 8× SR image at time T2 by SRCDNet. (f) Ground truth. (g) Base. (h) Base+SAM. (i) Base+SRM. (j) SRCDNet.

TABLE III
COMPARATIVE EXPERIMENTS ON BCDD AND CDD

Dataset	Bselines	X1				X4				X8			
		Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)
BCDD	FC-EF	85.47	72.59	78.51	64.62	**87.92**	68.01	76.69	62.20	**84.03**	65.30	73.49	58.09
	FC-Siam-diff	86.08	79.72	82.78	70.62	78.70	84.48	81.48	68.75	73.39	81.45	77.21	62.88
	FC-Siam-conc	77.12	85.89	81.27	68.44	73.58	**87.86**	80.09	66.79	69.13	68.89	69.01	52.69
	BiDateNet	85.73	81.48	83.55	71.75	80.87	83.11	81.97	69.45	76.74	64.36	70.01	53.86
	STANet	**90.07**	80.40	84.96	73.86	81.93	82.00	81.96	69.44	80.78	74.12	77.31	63.01
	Ours	84.84	**90.13**	**87.40**	**77.63**	84.44	86.90	**85.66**	**74.91**	81.61	**81.78**	**81.69**	**69.05**
CDD	FC-EF	90.41	69.60	78.65	64.81	91.16	66.03	76.58	62.05	86.81	63.75	73.51	58.12
	FC-Siam-diff	94.76	73.71	82.93	70.84	96.21	63.02	76.15	61.49	**95.43**	57.83	72.02	56.27
	FC-Siam-conc	92.15	80.44	85.90	75.28	**97.80**	59.09	73.67	58.31	94.71	57.87	71.84	56.05
	BiDateNet	**95.98**	84.74	90.01	81.83	94.27	79.54	86.28	75.87	92.80	67.71	78.29	64.33
	STANet	89.28	**93.71**	91.44	84.23	89.26	83.89	86.49	76.20	86.72	69.71	77.29	62.98
	Ours	92.55	93.34	**92.94**	**86.81**	92.07	**88.07**	**90.02**	**81.86**	92.15	**76.03**	**83.32**	**71.40**

information is over-filtered through the difference operation of bitemporal features in skip connections, whereas concatenation can better save such information in features.

The above results are further verified by visualization comparisons, as shown in Fig. 6, where the proposed method obtains the best visualization performance. All methods can

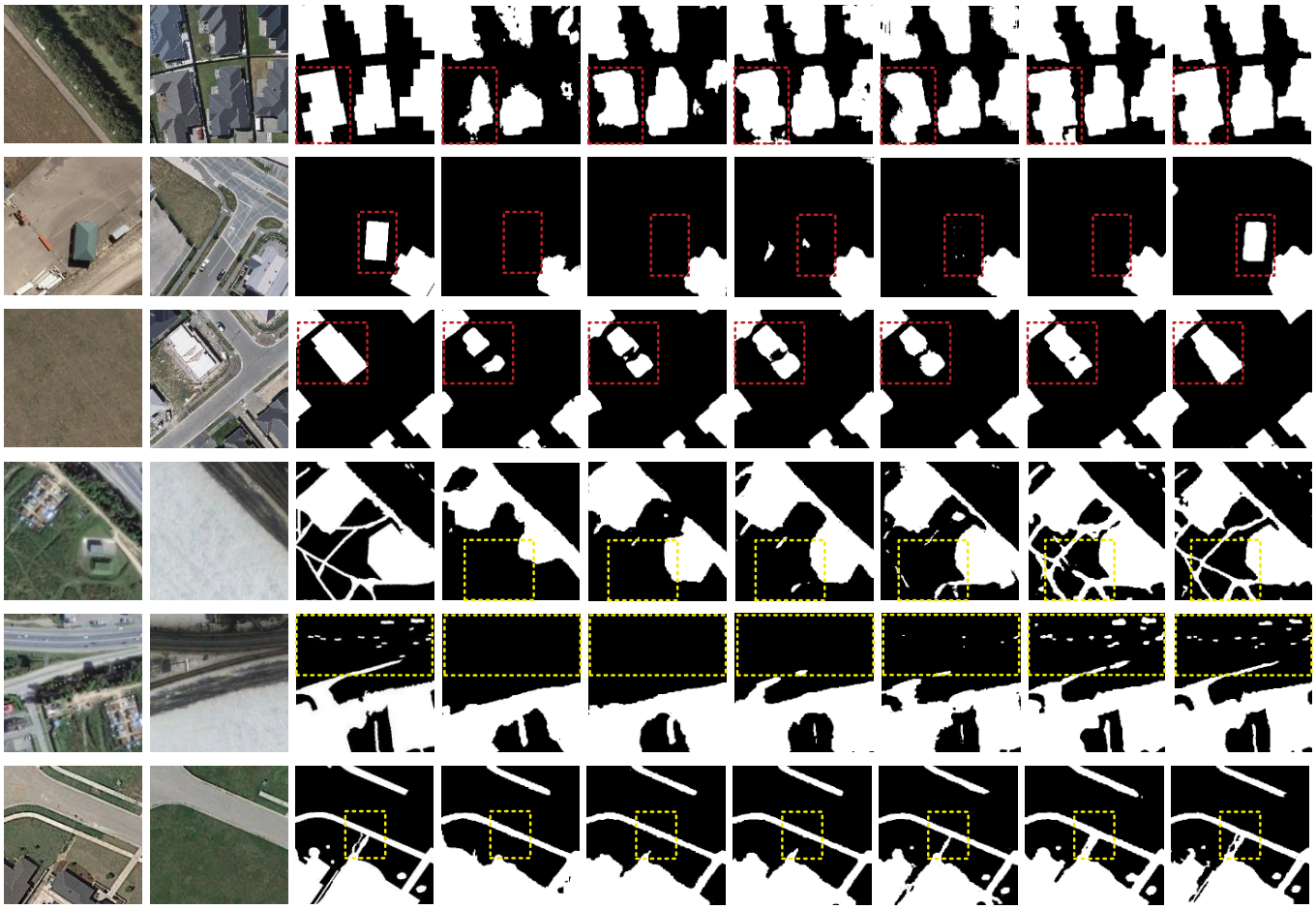

Fig. 6. Visualization of comparative experiments on X1 images (BCDD: Rows 1–3; CDD: Rows 4–6). (a) HR image at T1. (b) HR image at T2. (c) Ground truth. (d) FC-EF. (e) FC-Siam-diff. (f) FC-Siam-conc. (g) BiDateNet. (h) STANet. (i) SRCDNet.

accurately extract newly built buildings in the BCDD. In addition, the FC-EF and FC-Siam-diff entail many missed alarms; this corresponds to the low recall of the two methods, as shown in Table I. Compared with other methods, our proposed method is able to capture more precise building footprints, which is of great significance for practical applications. It is also important to note that our method is the only one to detect a building reduction, as shown in the second row of Fig. 6.

As for the CDD, only changes corresponding to large areas can be extracted by the three FCN variants, explaining their high precision and low recall, as shown in Table I. Benefitting from the integration of LSTM blocks, BiDateNet is better at extracting small changes compared with the other U-Net-based methods. The two metric learning-based methods are good at capturing small changes, including cars and roads. The STANet can extract the greatest number of small changes; however, the results show a spillover effect, which leads to its high recall but low precision, as shown in Table I. In terms of visualization results, our proposed method can not only extract small changes precisely, but also maintain their boundaries and shapes in a better way.

2) Performance on X4 Images: As shown in Table III, when there is a 4× resolution difference between the bitemporal images, both the F1 and IoU of all methods decline slightly on the BCDD dataset compared with experiments on X1 images. More specifically, SRCDNet obtains the highest F1 and IoU of 85.66% and 74.91% on the BCDD, which are 1.74% and 2.72% lower than those obtained in the X1 experiments, respectively. Although STANet outperforms BiDateNet in the X1 experiment, they achieve similar results in the X4 experiment, with F1 scores of 81.96% and 81.97%, respectively, suggesting that STANet is more affected by the resolution difference. Among the three FCN variants, FC-Siam-diff achieves the best performance, followed by FC-Siam-conc and finally FC-EF, which is consistent with the results of the X1 experiments.

The accuracies of all baselines decrease more significantly on the CDD. FC-Siam-conc has the worst detection results, with an F1 of 73.67%, which is 12.23% lower than that of the X1 experiment. The F1 of FC-Siam-diff is slightly higher, at 76.15%. Notably, FC-EF obtains the highest F1 of 76.58% among the three FCN variants, which is only 2.07% lower than that of the X1 experiment. BiDateNet and STANet have very close F1 values of 86.28% and 86.49%, which are 3.73% and 4.95% lower than those of the X1 experiment, respectively. In this case, SRCDNET still obtains a very high F1 and

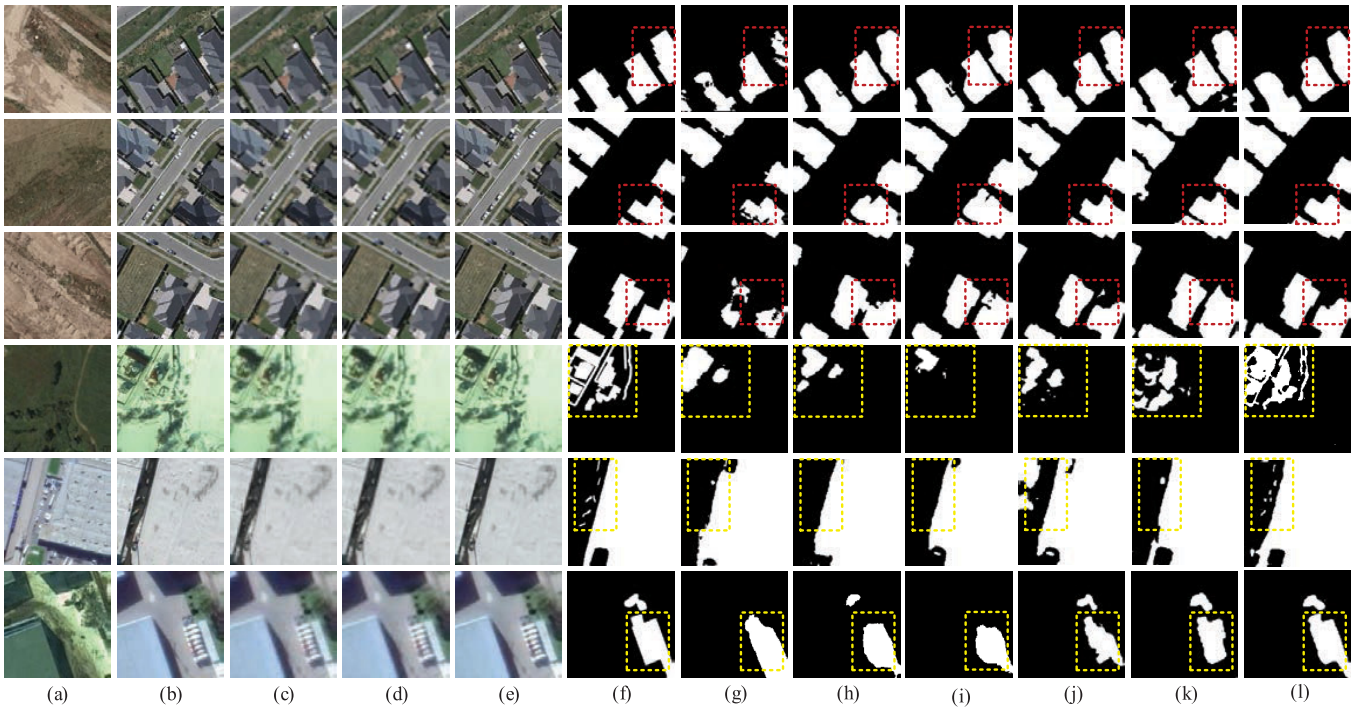

Fig. 7. Visualization of comparative experiments on X4 images (BCDD: Rows 1–3; CDD: Rows 4–6). (a) HR image at time T1. (b) HR image at time T2. (c) 4× LR image at time T2. (d) 4× Bicubic image at time T2. (e) 4× SR image at time T2. (f) Ground truth. (g) FC-EF. (h) FC-Siam-diff. (i) FC-Siam-conc. (j) BiDateNet. (k) STANet. (l) SRCDNet.

IoU of 90.02% and 81.86%, respectively, with only a 2.92% reduction in F1.

The visualization comparisons of these methods are shown in Fig. 7. Most of the building changes in the BCDD can be detected relatively completely from the 4× bicubic images, which indicates that owing to the large size of the buildings, the 4× resolution difference can be mitigated to some extent by bicubic interpolation. This also explains that the accuracy of the experiment on X4 images only exhibits a small decline compared with that of experiment on X1 images. However, the boundaries of the building changes obtained based on the comparative methods are not sufficiently regular and smooth, in contrast with those obtained via SRCDNet. This can be attributed to the fact that the SR image generated by the SRCDNet can better restore the boundary information of buildings compared with the bicubic image. In addition, while some bare lands adjacent to the building are easily extracted as part of the building changes, the SRCDNet can effectively reduce these pseudo-changes, as shown in Row 3 in Fig. 7.

Compared to the BCDD, the CDD has a higher spatial resolution and thus contains more detailed changes, which are difficult to reconstruct in 4× bicubic images; thus, the accuracy of the comparative methods declines significantly compared with that of the experiments on X1 images. While some detailed changes, such as alleys and cars, are rarely seen in the results of comparative methods, the SRCDNet is the only method to detect such changes, which further proves the validity of the SR module with regard to recovering spatial and semantic information. In addition to extracting smaller changes, SRCDNet also exhibits better performance in extracting large-area changes such as land changes and building changes. Thus, the quantitative and visualization comparisons fully demonstrate the effectiveness of SRCDNet on 4× LR images of both datasets.

3) Performance on X8 Images: The X8 experiment compares the performance of all the baselines on 8× LR images, as shown in Table III. On the BCDD, the accuracies of all methods are further reduced compared with the X4 experiment. FC-Siam-conc has the highest accuracy reduction, with the lowest F1 of 69.01%, followed by BiDateNet with an F1 of 70.01%. The F1 of FC-EF is decreased from 76.69% to 73.49%, and it is least affected by the resolution difference. FC-Siam-diff achieves the highest F1 accuracy among the U-Net-based models, with an F1 of 77.21%. STANet obtains an F1 of 77.31% and an IoU of 69.05%. Despite the large resolution differences, SRCDNet outperforms all the comparative methods and obtains the highest F1 of 81.69% and IoU of 69.05%.

It is difficult to recover many small objects from the LR images, and thus, the accuracies obtained on the CDD are lower than those obtained on the BCDD. More specifically, SRCDNet achieves the highest detection rate, with an F1 of 83.32% and an IoU of 71.40%, which are 6.70% and 10.46% lower than those obtained in the X4 experiments, respectively. Additionally, the F1 of BiDateNet is decreased from 86.28% to 78.29%, whereas that of STANet is decreased from 86.49% to 77.29%. Owing to the poor ability of the three FCN variants to extract small changes, the accuracies of these three methods are not greatly reduced in experiments on X8 images.

The visualization results of experiments on X8 images are shown in Fig. 8. Taking into account the results on the BCDD in Rows 1–3 of Fig. 8, it is possible to appreciate that it is

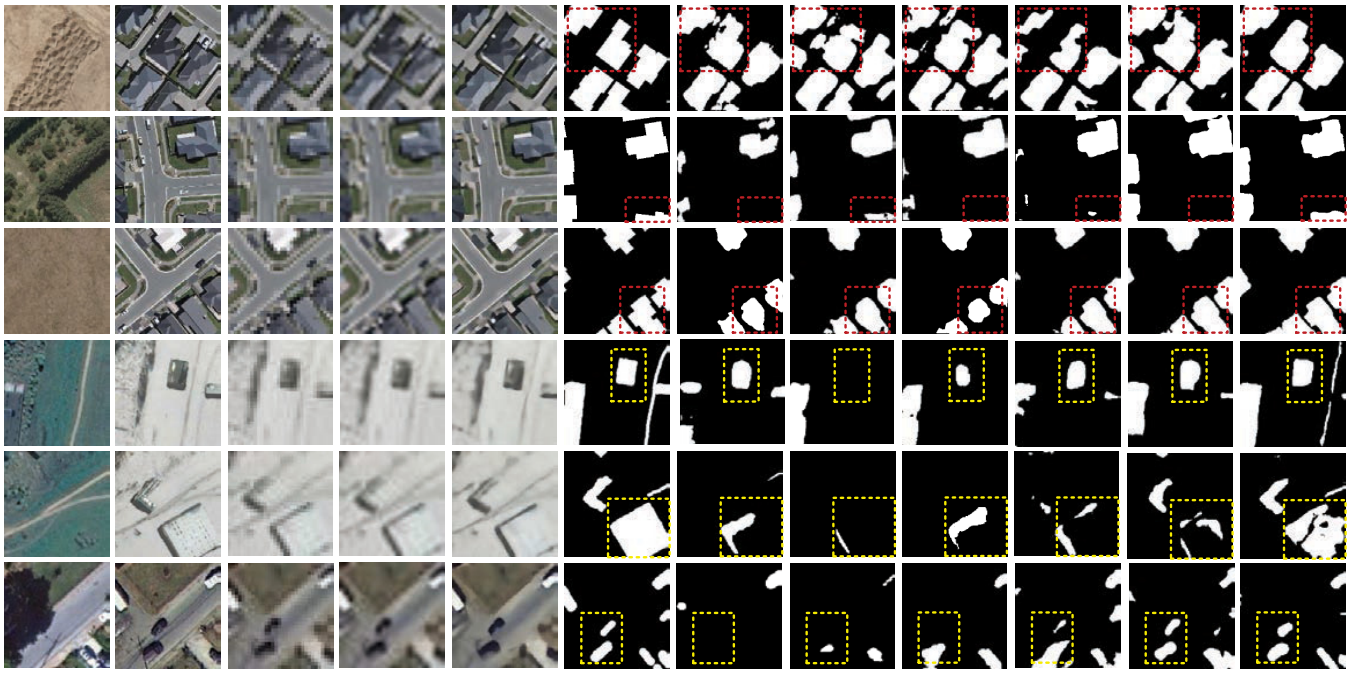

Fig. 8. Visualization of comparative experiments on X8 images (BCDD: Rows 1–3; CDD: Rows 4–6). (a) HR image at time T1. (b) HR image at time T2. (c) 8× LR image at time T2. (d) 8× Bicubic image at time T2. (e) 8× SR image at time T2. (f) Ground truth. (g) FC-EF. (h) FC-Siam-diff. (i) FC-Siam-conc. (j) BiDateNet. (k) STANet. (l) SRCDNet.

difficult to obtain regular building boundaries based on the bicubic images, as a result of the large resolution difference. Moreover, there are also a large number of missed alarms; that is, many houses with small volumes are missed. However, SRCDNet can better solve the aforementioned problems and generate more precise change results. It can be seen that despite the 8× resolution difference, the SR image can still recover the information of the building well, thanks to sufficient prior knowledge.

Numerous missed alarms also occur on the CDD, which is consistent with the low recall rates of all methods, as shown in Table I. This is because it is difficult to fully recover the information in the initial HR images, which has been greatly reduced after 8× down-sampling, by using bicubic interpolation. Therefore, not only small changes, including those of alleys and cars, but also some building changes are missed or incompletely detected. Notably, compared with the bicubic image, the SR image output by SRCDNet better regains information from the 8× LR image, which significantly helps in subsequent CD, leading to more complete and accurate change results.

C. Real-Image Experiments on Google Dataset

In the previous two sections, we have carried out sufficient ablation study and comparative experiments on the two simulated datasets of BCDD and CDD to verify the effectiveness of SRCDNet in solving the problem of resolution difference in CD. Next, comparative experiments would be conducted on the Google dataset to further test the effect of the model on real images. Besides, in order to explore the application potential of large datasets, transfer learning experiments were also conducted by adopting pretrained models on BCDD and CDD to be applied to the Google dataset.

1) Performance Analysis: All baselines are first trained from scratch on the Google dataset, where the SRCDNet still achieves the best accuracies, with an F1 of 77.13% and an IoU of 62.77%, indicating that the model also has a good effect on real images with 4 times resolution difference. The STANet and BiDateNet, which obtain 75.27% and 74.79% F1, respectively, are significantly superior to the other three UNet-based models, reflecting the gain effect of the attention mechanism model.

Fig. 9 demonstrates the CD results of buildings with different sizes. It can be seen that many pseudo-changes exist in the results of UNet-based models, which can be attributed to that some changes of impervious surfaces including roads are easily misclassified as buildings. With better discriminant ability, the STANet and SRCDNet can obtain more accurate results. In addition, compared with images restored by bicubic interpolation, images restored by SRCDNet can better recover the contour of buildings, which further guarantees the accuracy of building CD.

2) Transfer Learning Analysis: In recent years, more and more CDDs based on HR remote sensing images have been proposed for the test of CD algorithms, and many advanced algorithms have achieved good results on these datasets. However, it is sometimes difficult for a model trained on one dataset to be directly applied to real images due to domain shifts between different data. At present, one of the mainstream methods to solve this problem is transfer learning. Therefore,

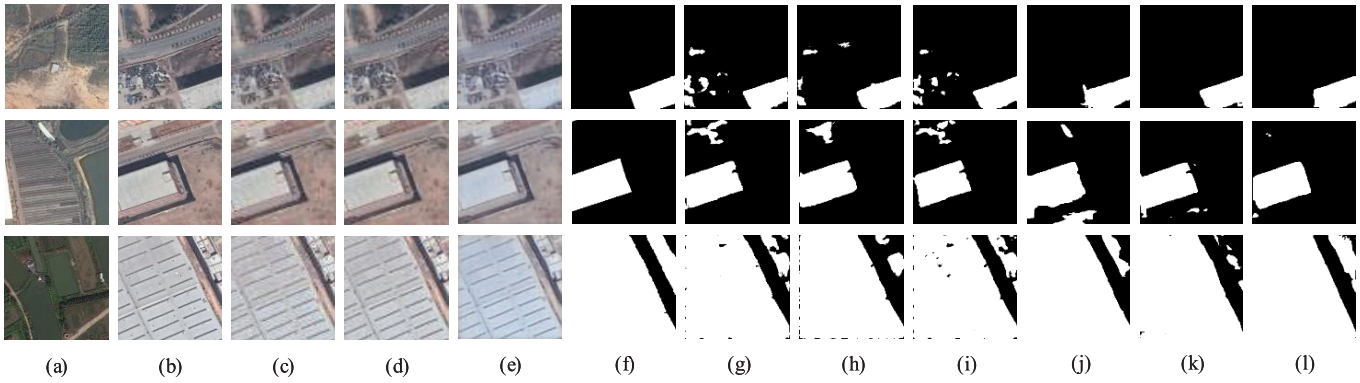

Fig. 9. Visualization comparisons on the Google dataset. (a) HR image at time T1. (b) HR image at time T2. (c) 4× LR image at time T2. (d) 4× Bicubic image at time T2. (e) 4× SR image at time T2. (f) Ground truth. (g) FC-EF. (h) FC-Siam-diff. (i) FC-Siam-conc. (j) BiDateNet. (k) STANet. (l) SRCDNet.

TABLE IV
REAL-IMAGE EXPERIMENT ON GOOGLE DATASET

Baseline	Google				BCDD - Google				CDD - Google			
	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)
FC-EF	80.81	64.39	71.67	55.85	85.85	63.48	72.99	57.47	**86.95**	61.72	72.20	56.49
FC-Siam-diff	85.44	63.28	72.71	57.12	**87.27**	63.72	73.66	58.30	85.80	64.80	73.83	58.52
FC-Siam-conc	82.07	64.73	72.38	56.71	84.56	66.66	74.55	59.43	84.94	65.42	73.92	58.62
BiDateNet	78.28	**71.59**	74.79	59.73	81.92	70.08	75.54	60.69	81.30	73.68	77.30	63.00
STANet	**89.37**	65.02	75.27	60.35	82.58	70.19	75.89	61.14	86.45	67.96	76.10	61.42
Ours	83.74	71.49	**77.13**	**62.77**	83.18	**76.34**	**79.62**	**66.14**	84.46	**76.34**	**80.20**	**66.94**

transfer learning experiments were also designed in this part, with the aim to explore the potential of applying large datasets to real images. Each baseline would be fine-tuned on the reconstructed Google dataset with 4× resolution difference by loading the best-trained model on BCDD and CDD as starting point.

According to Table IV, the accuracy of each baseline is improved by fine-tuning compared to that of training from scratch. As for FC-EF, FC-Siam-diff, and FC-Siam-conc, the pretrained models on BCDD have a slightly better gain effect than those on the CDD. A possible reason may lie in that BCDD and Google dataset both focus on building changes and share similar high-level characteristics, thus providing more immediate enhancements to the above model with relatively simple structure. Whereas, with BiDateNet, STANet, and SRCDNet, it is the reverse. This may be due to the larger amount of CDD, which can supply more diverse features conducive to the training of attention mechanism modules in these complex models. In a word, the transfer learning experiment indicates that the pretrained model on large datasets can improve the CD results on real images. In addition, compared with other baselines, the pretrained models on BCDD and CDD demonstrate largest gain effect for SRCDNet, with an increase by 2.49% and 3.07% on F1, respectively.

VI. DISCUSSION

In this section, further experiments and discussions on SRCDNet were made: at first, LR images obtained by different down-sampling strategies are used as the input of the model to test its robustness to different input images; then, comparative experiments were set up to examine the sensitivity of the loss function.

A. Comparisons on Different Down-Sampling Strategies

As mentioned above, in the comparative experiments, the second-phase images in BCDD and CDD were bicubically down-sampled to obtain the simulated LR images. However, the available images may be more complex in practical. Therefore, with the aim to verify the robustness of the model for different inputs, different down-sampling strategies, including nearest and bilinear, were applied to obtain two more sets of X4 and X8 images of BCDD for comparison.

Results in Table V showed that difference on the accuracies of using different down-sampling strategies were not significant, which indicates that SRCDNet has good robustness for different inputs. As for experiments on X4 images, the "Bicubic" inputs achieved the highest accuracies, while for experiments on X8 images, accuracies of "Bilinear" inputs show best performance. The "Nearest" inputs obtain lowest accuracies in both experiments on X4 and X8 images. The reason may be that the LR image obtained by nearest down-sampling consists more noises, as can be seen in Fig. 10. Moreover, the SRCDNet can obtain restored images with good visual effect for all of the three different inputs, which further verifies the good robustness of the model.

B. Sensitivity Experiments on the Loss Function

In the process of training SRCDNet, a value λ plays a vital role in the loss function to balance the benefit of CD results

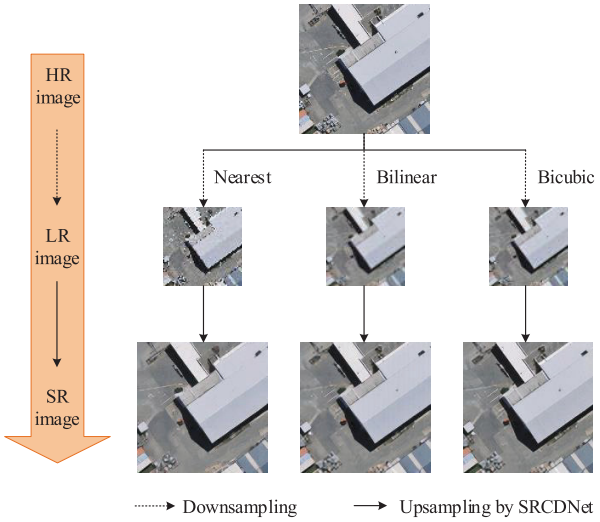

Fig. 10. Examples of the LR and SR images by different down-sampling strategies.

TABLE V
DIFFERENT DOWN-SAMPLING STRATEGIES ON BCDD

Down-sampling strategies	X4		X8	
	F1	IoU	F1	IoU
Nearest	85.36	74.46	80.47	67.32
Bilinear	85.41	74.54	**82.37**	**70.03**
Bicubic	**85.66**	**74.91**	81.69	69.05

TABLE VI
LOSS SENSITIVITY EXPERIMENTS ON BCDD

$\lambda(\times 10^{-3})$	X4		X8	
	F1	IoU	F1	IoU
0.25	83.78	72.09	79.69	66.24
0.5	84.57	73.27	81.10	68.21
0.75	84.86	73.70	81.62	68.95
1	**85.66**	**74.91**	**81.69**	**69.05**
1.25	83.82	72.14	81.02	68.10
1.5	84.38	72.98	80.56	67.46
1.75	84.62	73.35	80.96	68.05

on the optimization of generator. Therefore, we conducted sensitivity experiments on the BCDD by setting up a group of numbers with the step size of 0.25×10^{-3} on λ to explore the sensitivity of the loss function. The results are shown in Table VI.

The results on the BCDD show that different λ values had a great influence on change results; in other words, the loss function is sensitive to the λ value. More specifically, when λ increases from 0.25×10^{-3} to 1×10^{-3}, the accuracy of CD results increases gradually, while the best accuracy is obtained when $\lambda = 1 \times 10^{-3}$. When $\lambda > 1 \times 10^{-3}$, the accuracy begins to fluctuate. Based on the above results, it is suggested to choose a λ near 1×10^{-3} when apply.

VII. CONCLUSION

We propose an end-to-end SRCDNet for bitemporal images with different resolutions. With the aim of overcoming the resolution difference between bitemporal images, an SRM consisting of a generator and a discriminator is adopted to restore the LR images to the size of the HR images, which has been proved to be effective in generating realistic SR images from the LR images. A Siamese feature extractor extracts multiscale features from two input images: an SR image and an HR image corresponding to a different timestamp, on which a SAM is applied to capture more useful channel-wise and spatial information. SRCDNet employs deep metric learning to learn the final change map. The ablation study verifies the effectiveness of the SRM and SAM in SRCDNet. Then the comparative experiments on the BCDD and CDD, in which the SRCDNet not only obtains the best results for images with the same resolution, but also outperforms other comparative methods on 4 and 8 times different-resolution images, which fully proves the capacity of SRCDNet to be a general and useful solution for multiresolution CD. The real-image experiment on the Google dataset with 4 times resolution difference further verifies the validity of SRCDNet on real images. In the future, we seek to explore other methods for different-resolution CD and promote the development of deep learning in the application of CD.

REFERENCES

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.
[2] C. Marin, F. Bovolo, and L. Bruzzone, "Building change detection in multitemporal very high resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2664–2682, May 2015.
[3] C.-F. Chen *et al.*, "Multi-decadal mangrove forest change detection and prediction in honduras, central America, with landsat imagery and a Markov chain model," *Remote Sens.*, vol. 5, no. 12, pp. 6408–6426, Nov. 2013.
[4] B. Demir, F. Bovolo, and L. Bruzzone, "Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 300–312, Jan. 2013.
[5] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Informat.*, vol. 12, no. 2, pp. 143–160, Jun. 2019.
[6] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in landsat imagery," *Remote Sens. Environ.*, vol. 118, pp. 83–94, Mar. 2012.
[7] F. Yuan and M. E. Bauer, "Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in landsat imagery," *Remote Sens. Environ.*, vol. 106, no. 3, pp. 375–386, Feb. 2007.
[8] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.
[9] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.*, vol. 19, no. 3, pp. 411–426, Jan. 1998.
[10] G. F. Byrne, P. F. Crapper, and K. K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal landsat data," *Remote Sens. Environ.*, vol. 10, no. 3, pp. 175–184, Nov. 1980.
[11] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.
[12] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.
[13] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 214–217.
[14] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[15] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2115–2118.

[16] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.

[17] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[18] M. Gong, P. Zhang, L. Su, and J. Liu, "Coupled dictionary learning for change detection from multisource data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7077–7091, Dec. 2016.

[19] C. C. Petit and E. F. Lambin, "Integration of multi-source remote sensing data for land cover change detection," *Int. J. Geograph. Inf. Sci.*, vol. 15, no. 8, pp. 785–803, Dec. 2001.

[20] S. Dayanik, H. V. Poor, and S. O. Sezer, "Multisource Bayesian sequential change detection," *Ann. Appl. Probab.*, vol. 18, no. 2, pp. 552–590, Apr. 2008.

[21] J. Tu, D. Li, W. Feng, Q. Han, and H. Sui, "Detecting damaged building regions based on semantic scene change from multi-temporal high-resolution remote sensing images," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 5, p. 131, Apr. 2017.

[22] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[23] J. Verhoeye, "Land cover mapping at sub-pixel scales using linear optimization techniques," *Remote Sens. Environ.*, vol. 79, no. 1, pp. 96–104, Jan. 2002.

[24] P. Aplin and P. M. Atkinson, "Sub-pixel land cover mapping for per-field classification," *Int. J. Remote Sens.*, vol. 22, no. 14, pp. 2853–2858, Jan. 2001.

[25] G. M. Foody and D. P. Cox, "Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions," *Int. J. Remote Sens.*, vol. 15, no. 3, pp. 619–631, Feb. 1994.

[26] F. Ling, W. Li, Y. Du, and X. Li, "Land cover change mapping at the subpixel scale with different spatial-resolution remotely sensed imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 182–186, Jan. 2011.

[27] Q. Wang, W. Shi, P. M. Atkinson, and Z. Li, "Land cover change detection at subpixel resolution with a hopfield neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 3, pp. 1339–1352, Mar. 2015.

[28] X. Li, F. Ling, G. M. Foody, and Y. Du, "A superresolution land-cover change detection method using remotely sensed images with different spatial resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3822–3841, Jul. 2016.

[29] K. Wu, Q. Du, Y. Wang, and Y. Yang, "Supervised sub-pixel mapping for change detection from remotely sensed images with different resolutions," *Remote Sens.*, vol. 9, no. 3, p. 284, Mar. 2017.

[30] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[31] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 1–7, 2018.

[32] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 6, 2021, doi: 10.1109/TGRS.2020.3011913.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Göttingen, Germany: Copernicus Publications, 2015, pp. 234–241.

[35] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, Jun. 2019.

[36] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[37] M. Wang, K. Tan, X. Jia, X. Wang, and Y. Chen, "A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images," *Remote Sens.*, vol. 12, no. 2, p. 205, Jan. 2020.

[38] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*. [Online]. Available: http://arxiv.org/abs/1409.2329

[39] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Jan. 2014, pp. 338–342.

[40] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3D fully convolutional networks," *Remote Sens.*, vol. 10, no. 11, p. 1827, Nov. 2018.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[42] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[43] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[44] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[45] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014, *arXiv:1406.2661*. [Online]. Available: http://arxiv.org/abs/1406.2661

[46] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4681–4690.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

Mengxi Liu (Student Member, IEEE) received the B.S. degree in geographic information science from Sun Yat-sen University, Guangzhou, China, in 2019, where she is pursuing the Ph.D. degree in cartography and geographic information system with the School of Geography and Planning.

Her research interests include intelligent understanding of remote sensing images, change detection, and domain adaptation.

Qian Shi (Senior Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2015.

She is an Associate Professor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. Her research interests include remote sensing image classification, including deep learning, active learning, transfer learning.

Andrea Marinoni (Senior Member, IEEE) received the B.S., M.Sc. *(cum laude)*, and Ph.D. degrees in electronic engineering from the University of Pavia, Pavia, Italy, in 2005, 2007, and 2011, respectively.

From 2013 to 2018, he was a Research Fellow with the Telecommunications and Remote Sensing Laboratory, Department of Electrical, Computer and Biomedical Engineering, University of Pavia. In 2009, he was a Visiting Researcher with the Communications Systems Laboratory, Department of Electrical Engineering, University of California, Los Angeles (UCLA), Los Angeles, CA, USA. From 2015 to 2017, he was a Visiting Researcher with Earth and Planetary Image Facility, Ben-Gurion University of the Negev, Be'er Sheva, Israel; School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China; School of Computer Science, Fudan University, Shanghai, China; Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China; Instituto de Telecomunicações, Instituto Superior Tecnico, Universidade de Lisbon, Lisbon, Portugal. In 2020, he was a Visiting Professor with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia. He is an Associate Professor with the Department of Physics and Technology, Earth Observation Group, Center for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA), UiT the Arctic University of Norway, Tromsø, Norway. His main research interests are focused on efficient information extraction from multimodal remote sensing, nonlinear signal processing applied to large-scale heterogeneous records, earth observation interpretation and big data mining, analysis, and management for human–environment interaction assessment.

Da He (Member, IEEE) received the B.S. degree in remote sensing science and technology and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2015 and 2020, respectively.

He is a Postdoctor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. He has won the National Scholarship for doctoral students and awards for Survey Science and Technology Progression (ranked 15th). His research interests include multi- and hyperspectral remote sensing image processing, deep learning, space–time analysis, and change detection.

Xiaoping Liu (Member, IEEE) received the B.S. degree in geography and the Ph.D. degree in remote sensing and geographical information sciences from Sun Yat-sen University, Guangzhou, China, in 2002 and 2008, respectively.

He is a Professor with the School of Geography and Planning, Sun Yat-sen University. He has authored two books and over 100 articles. His research interests include image processing, artificial intelligence, and geographical simulation.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China in State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He was the Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has published more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He is the holder of 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE GRSS 2014 Data Fusion Contest, and his students have been selected as the Winners or the Finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He also serves as an Associate Editor or an Editor of more than ten international journals. He is serving as an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is the Founding Chair of IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter.