A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection

Qian Shi, *Senior Member, IEEE*, Mengxi Liu, *Student Member, IEEE*, Shengchen Li, Xiaoping Liu, *Member, IEEE*, Fei Wang, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— Change detection (CD) aims to identify surface changes from bitemporal images. In recent years, deep learning (DL)-based methods have made substantial breakthroughs in the field of CD. However, CD results can be easily affected by external factors, including illumination, noise, and scale, which leads to pseudo-changes and noise in the detection map. To deal with these problems and achieve more accurate results, a deeply supervised (DS) attention metric-based network (DSAMNet) is proposed in this article. A metric module is employed in DSAMNet to learn change maps by means of deep metric learning, in which convolutional block attention modules (CBAM) are integrated to provide more discriminative features. As an auxiliary, a DS module is introduced to enhance the feature extractor's learning ability and generate more useful features. Moreover, another challenge encountered by data-driven DL algorithms is posed by the limitations in change detection datasets (CDDs). Therefore, we create a CD dataset, Sun Yat-Sen University (SYSU)-CD, for bitemporal image CD, which contains a total of 20 000 aerial image pairs of size 256 × 256. Experiments are conducted on both the CDD and the SYSU-CD dataset. Compared to other state-of-the-art methods, our network achieves the highest accuracy on both datasets, with an F1 of 93.69% on the CDD dataset and 78.18% on the SYSU-CD dataset.**

***Index Terms*— Change detection dataset (CDD), convolutional block attention module (CBAM), deeply supervised (DS) layers, metric learning, remote sensing change detection (CD).**

I. Introduction

CHANGE detection (CD) is the process of quantitatively analyzing surface changes between different phases in the same area [1]. This process is of great significance to many fields, including environmental investigation [2], geological disaster monitoring [3], land cover surveys, and urban planning [4], [5]. In recent decades, regular monitoring and analysis of changes in land cover has become increasingly crucial because of the deterioration of the ecological environment. Meanwhile, high-resolution multisource and multitemporal remote sensing images can be obtained over different areas, which have been proven to be a key source of primary data for change detection (CD) because of their wide coverage, high temporal resolution, and diverse data types [6].

Traditional CD methods mainly detect changes by exploiting spectral information in the remote sensing images, such as change vector analysis (CVA) [7], principal component analysis (PCA) [8], multivariate alteration detection (MAD) [9], and so on. However, methods of this kind often require optimal threshold selection in the decision phase, which makes them scene dependent and time consuming. Accordingly, machine-learning algorithms, which can learn from a part of labeled samples to get an automated decision model, have been widely used for remote sensing CD [10], including support vector machine [11], decision tree [12], and random forest [13]. However, such methods rely heavily on hand-crafted features, which is difficult to effectively capture high-level features representations, resulting in lower accuracy.

In recent years, the rapid rise of big data and the popularization of high computational power have promoted myriad developments in deep learning (DL), which has made remarkable achievements in many fields [14], [15], including remote sensing image interpretation [16]–[19]. As a powerful DL structure, convolutional neural networks (CNNs) are able to automatically extract hierarchical multilevel features with rich spectral and spatial features from satellite images [20]. Although DL-based models have shown strong feature extraction ability and achieved great process in CD applications, some remarkable CD approaches based on the DL framework have been developed [6], [21]–[23].

The framework for CD methods can be summarized in two steps: 1) extract features with distinctive change information and 2) design a decision function to generate a change map based on the extracted features [24]. The improvements to CD methods in recent literature are mainly with reference to these two aspects. On the one hand, more discriminative features are of great significance to relieve pseudo-changes, which refers

Manuscript received November 5, 2020; revised April 7, 2021; accepted May 25, 2021. Date of publication June 29, 2021; date of current version January 12, 2022. This work was supported 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 the National Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, Guangzhou Applied Basic Research Project, and in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X355. *(Corresponding author: Mengxi Liu.)*

Qian Shi, Mengxi Liu, Shengchen Li, 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: shixi5@mail.sysu.edu.cn; liumx23@mail2.sysu.edu.cn; lishch8@mail2.sysu.edu.cn; liuxp3@mail.sysu.edu.cn).

Fei Wang is with the Xinjiang Common University Key Laboratory of Smart City and Environmental Stimulation, College of Resource and Environmental Sciences, Xinjiang University, Urumqi 830046, China (e-mail: wangfei1986@xju.edu.cn).

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.2021.3085870

to false alarms caused by the differences of some external factors, such as illumination and scale differences between bitemporal inputs. Therefore, many attempts have been made to produce more discriminative features to overcome this problem. Peng *et al.* [25] developed a UNet++ with dense skip connections, which aims to learn more effective features with multiscale semantic information. Because of their excellent ability to capture temporal dependencies between bitemporal images, recurrent neural networks (RNNs) are employed to obtain features with spatial-temporal information. Song *et al.* [26] presented a recurrent 3-D fully convolutional network (FCN) for hyperspectral images, which combine the advantages of a 3-D FCN and Long Short-Term Memory networks (LSTMs). Papadomanolaki *et al.* [27] proposed BiDateNet to enhance temporal information between bitemporal images by integrating LSTMs into a U-Net structure.

On the other hand, the strategy used to decide the category of each pixel in order to generate the change map plays a key role for accurate CD. Of late, deep metric learning [28] has also made significant progress in the area of CD [29], [30]. Wang *et al.* [30] proposed a Siamese convolutional network to detect changes by determining the difference between extracted features. Compared to common FCN-based methods, which obtain a change map via classification, metric-based methods obtain a change map by measuring the distance between the features. Because the distance between unchanged features would reduce while that of change features would move further apart, metric-based methods can further distinguish between the "changed" and "unchanged" features in the decision-making stage, thereby relieving the influence of pseudo-changes and improving the CD accuracy.

Nevertheless, there are still some problems with the existing CD methods. First, previous studies mainly focus on using RNNs to exploit the temporal dependencies between bitemporal images for more distinguishable features. However, for some pseudo-changes caused by objects with very confusing appearance, it is still difficult to distinguish through the time information. Thus, more effective techniques should be exploited to extract more discriminative information from the features and alleviate the influence of pseudo-changes. Second, the features generated by hidden layers are less semantically meaningful [31], which significantly influences the subsequent prediction and can lead to insufficiency in terms of the boundary and shape of the change area [32]. Therefore, to help the restoration of the morphology of the changed region, attempts to capture more semantic information still need to be explored.

In addition, large numbers of samples are required for model training and testing; this is an unavoidable challenge brought about by the data-driven nature of DL algorithms. In this case, large open-source CD datasets are highly important to the development of CD technology [21]. For one thing, it is highly labor- and time-consuming for researchers to collect satellite image pairs with well-labeled change information; for another thing, a standard dataset can provide a unified benchmark for different algorithms and models, which is helpful for improving CD methods. Recently, several CD datasets of different image types have been proposed, including multispectral images, hyperspectral images and optical images with RGB bands, and so on [23], [29], [33]–[39]. However, because of the limited number of datasets available, it is still difficult to satisfy diversified CD scenarios.

In light of the above-mentioned problems, a deeply supervised attention metric-based network (DSAMNet) for bitemporal image CD is proposed in this article. First, multiscale features from different levels of bitemporal images are extracted by means of a Siamese feature extractor. Subsequently, the convolutional block attention module (CBAM) [40] is exploited with the aim of making the features from different phases more distinguishable in both a channel-wise and spatial-wise sense. In the next step, a metric module learns a change map from the distance between pairwise bitemporal feature maps in a low-dimensional embedding space. Furthermore, a deeply supervised (DS) module is introduced into the feature extractor, with the goal of enhancing the feature extractor's learning ability and learning more effective information. Finally, a hybrid loss is used to combine the results of the two modules for network-training purposes.

Moreover, with the goal of alleviating the dependence of DL model on large number of samples and meeting the demands of CD under different scenarios, we constructed a new, large-scale, open-source change detection dataset (CDD), named "Sun Yat-Sen University (SYSU)-CD" (named after our university). The dataset consists of 20 000 pairs of image patches with a resolution of 0.5 m taken from 800 pairs of 1024 × 1024 orthographic aerial images in Hong Kong, along with corresponding binary change maps for each pixel. The SYSU-CD dataset provides not only common change information in urban and suburban areas, but also supplements the annotation of high-density building changes and sea construction; although these are important for urban decision makers, they have rarely been seen in previous datasets.

The contributions made by this article to the literature can be summarized as follows:

1) We provide a new CD network, DSAMNet, which integrates CBAM blocks for more discriminative features on both spatial-wise and channel-wise and DS layers for better feature extraction, to achieve fine-grained bitemporal CD.
2) We propose a new, large-scale, open-source CD dataset, SYSU-CD, which contains 20 000 pairs of 0.5 m aerial image patches for remote sensing CD. This dataset supplements multiple change samples to the existing datasets for more diversified CD.
3) The proposed DSAMNet achieves state-of-the-art performance, not only on the widely used CDD benchmark dataset but also on the proposed SYSU-CD dataset, with a highest F1 of 93.69% and 78.18%, respectively.

The remainder of this article is structured as follows: Section II provides an overall review of related works, whereas Section III describes the proposed SYSU-CD dataset in detail. The proposed network is introduced in Section IV. The settings and results of all experiments are presented in Section V. Section VI discusses the sensitivity of the loss in the DSAMNet and the necessity of the proposed dataset. Finally, we conclude this article in Section VII.

II. Related Works

A. Traditional CD

Simple algebra was applied in early CD methods to obtain the difference image as features, including image ratioing [41], image differencing [42], and image regression [43]. Because algebra-based methods heavily depend on the selection of the threshold to decide changed pixels and unchanged pixels, some transform techniques have been introduced into CD. These transform-based methods emphasize different information on the derived difference map by means of multiscale decomposition, such as PCA and Tasseled Cap Transformation [44]. As these linear transformation techniques rely heavily on the statistical property of images, MAD has been proposed to learn correlations between the two components based on canonical correlation analysis. The iteratively reweighted MAD (IR-MAD) method [45] applied different weights to each pixel on the basis of MAD. Although easy to implement, it remains difficult for these methods to satisfy large-scale, fine-grained CD in the context of shallowly extracted features and repetitive threshold adjustment for change map generation.

B. DL-Based CD

Nowadays, because of their strong feature extraction ability, the newly developed DL-based CD methods have obtained excellent results. Although CNNs only output a patch-wise category rather than pixel-wise prediction in order to gain a finer-grained change map, attempts have been made to assign a patch's output to its central pixel according to the principle of space proximity [39], [46], [47]. A novel recurrent convolutional neural network (ReCNN) was provided by Mou *et al.* [46], which is able to learn better feature representations with joint spectral-spatial-temporal information from multispectral images for CD using a very small input size of 5 × 5. general end-to-end 2-D convolutional neural network (GETNET) was proposed for hyperspectral image CD; this approach employs an effective mixed affinity matrix to mine the change patterns between two corresponding spectral vectors [39]. Nevertheless, applying such methods may result in very low computational efficiency and very high memory consumption because of the large overlap between adjacent patches. Moreover, the patch size determines the receptive field of CNNs, whereas a small receptive field with insufficient contextual information may lead to limited CD performance.

In light of these problems, a FCN [48] has been proposed to achieve pixel-wise prediction by replacing the fully connected layer of CNN with a fully convolutional (FC) layer. Since this time, FCN and its variants have provided another approach to fine-grained CD [22], [25]–[27], [49]. These methods can be summarized as classification-based methods [29], because they obtain the change map by classifying the extracted features. Daudt *et al.* [22] discussed three different U-Net-based variants, namely, FC-early fusion (EF), FC-Siam-conc, and FC-Siam-diff, by exploring two image input methods (EF and Siamese) and two skip connection methods (concatenation and difference). Although a simple connection can help the spatial information recovery in the upsampling stage, it remains difficult to fulfill the needs of multiscale change objects because of single feature extraction in each skip connection. Therefore, UNet++ employs dense skip connections to achieve multiscale feature extraction and reduce pseudo-changes caused by scale variance [25]. With the aim of fully exploiting temporal dependence between bitemporal images, the BiDateNet integrated LSTMs into the skip connection in order to obtain more temporally distinguishable features [27]. Unlike the above classification-based method, the spatial-temporal attention-based network (STANet) obtains the change map by means of metric learning, which uses a spatial and temporal attention mechanism to obtain more discriminative spatial and temporal features [29]. However, because of the limitation of temporal dependence to distinguish pseudo-changes that are very confusing on appearance, the improvement that these methods can achieve is limited. Moreover, the changed map exhibits poor morphology because of a lack of semantic information.

Consequently, our proposed method introduces CBAM into the network. This approach learns more discriminative features in a channel-wise and spatial-wise manner in order to alleviate the influence of pseudo-changes. In addition, a DS module is also integrated to supervise the learning of the feature extractor and supplement more useful information to generate the change map.

III. SYSU-CD Dataset

Over the last few decades, great efforts have been witnessed in relation to developing open CDDs. We collected the number of image pairs, size, resolution, and band number of the images of these datasets for our proposed dataset in Table I.

Evidently, the majority of these CD datasets are based on high-resolution images (HRIs); these contain abundant spatial information and are more favorable for visual interpretation compared to low- and medium-resolution images. The SZTAKI Air Change Benchmark set [33], which has a spatial resolution of 1.5 m with 13 pairs of 952 × 640 optical aerial images, was the earliest and mostly commonly used CD dataset in early studies. The aerial imagery change detection (AICD) dataset [34] comprises 1000 pairs of aerial images with synthetic changes, each with a size of 800 × 600 and resolution of 0.5 m. By contrast, synthetic images may be inadequate to reflect real changes. To make full use of the rich change information contained in high-resolution images, a CDD [35] was released containing 16 000 image pairs. The images are 256 × 256 pixels in size with a spatial resolution of 0.03–1 m, and were collected from seven pairs of 4725 × 2700 real season-varying remote sensing images. Recently, a few datasets have been specifically proposed to monitor changes in buildings, which is one of the most common and concerned types in CD, including the Wuhan University (WHU) Building CD [36], AIST Building Change Detection (ABCD) dataset [37], and the Learning, Vision and Remote Sensing Laboratory (LEVIR)-CD [29].

Multispectral images, because they have both relatively high temporal and spatial resolution, rich spectral information, and good data accessibility, have always been the main data source in all kinds of remote sensing applications. The Onera Satellite Change Detection (OSCD) benchmark [23] contains 24 pairs of Sentinel-2 images taken between 2015 and 2018. The images in OSCD are taken from urban areas

TABLE I
INFORMATION OF DIFFERENT CDDS

Dataset	Number of Image Pairs	Image Size	Resolution(m)	Number of Bands
SZTAKI	13	952×640	1.5	3
AICD	1000	800×600	0.5	3
		984×740	20	224
HCCD	3	600×500	20	224
		390×200	30	242
OSCD	24	600×600	10	13
WHU Building CD	1	32207×15354	0.2	3
CDD	16000	256×256	0.03-1	3
ABCD	4253	160×160	0.4	3
River	1	463×241	30	242
LEVIR-CD	637	1024×1024	0.3	3
SYSU-CD	20000	256×256	0.5	3

around the world (including Asia, Brazil, Europe, the Middle East, and the USA). Hence, the change information is mainly reflected in urban expansion and renewal but lacks natural changes. Moreover, multispectral images suffer from mixed pixel problem. This means that subtle changes, such as small buildings and alleys, are not noticed and can be easily ignored at a 10-m resolution. In addition to multispectral images, some hyperspectral image-based CDDs have also been proposed in recent years, including the hyperspectral change detection dataset (HCCD) [38] and "River" dataset [39].

Although these datasets have addressed the issue of detecting changes between multitemporal remote sensing images, there are still some improvements that could be made to achieve large-scale and fine-grained CD. First, considering that HRIs are able to provide more precise information for CD, the resolution of some datasets remains insufficient; this is particularly true with regard to the datasets of hyperspectral images. Second, the change types of the existing datasets are not diversified enough to cope with diverse needs in practical applications. Last but not least, the volumes of some datasets are slightly too small for data-driven DL methods, which may result in model over-fitting and poor performance.

The SYSU-CD dataset largely complements existing CD datasets in terms of image resolution, change types, and dataset volume and further provides a new benchmark for CD. The dataset contains 20 000 pairs of 0.5-m aerial images taken between the years 2007 and 2014 in Hong Kong, which has long been a prosperous and populous metropolis situated in the south of China. With a total land area of 1,106.66 square kilometers and a total population of about 7.2 million by the end of 2014, Hong Kong was ranked third in the world for population density, resulting in very high density of high-rise buildings in urban areas. Moreover, the amount of construction and maintenance in ports, sea routes, and oceanic and coastal projects in Hong Kong, as well as the major shipping hubs in international and Asia-Pacific areas, have increased rapidly from 2007 to 2014 under the rapid development of the littoral economy and nautical transport. Accordingly, our dataset greatly complements change instances of high-rise buildings, which are very difficult to mark in HRIs because of the influence of deviation and shadow, as well as change information related to the port compared to previous datasets.

According to the ratio of 6:2:2, we first divided the original 800 image pairs into training set, verification set, and test set. Thereafter, to generate a dataset for DL application, 25 sample pairs of 256 × 256 size are randomly collected from each image pair, where random flip and rotation are applied for data augmentation. The preprocessing results in a total of 20 000 pairs of aerial image patches of size 256 × 256. As shown in Fig. 1, the main types of changes in the dataset include (a) newly built urban buildings; (b) suburban dilation; (c) groundwork before construction; (d) change of vegetation; (e) road expansion; and (f) sea construction. The SYSU-CD dataset will be made openly available for all research needs.

IV. METHODOLOGY

In this section, an overview of our proposed network is first provided, after which each network module will be described in detail. Finally, the model optimization strategy is introduced.

A. Overview

The architecture of the proposed network is presented in Fig. 2. It consists of three parts: a feature extractor, a metric module, and a DS module. To learn representative features for CD, the feature extractor automatically extracts multiscale features from bitemporal inputs. The metric module then learns a distance map according to the bitemporal feature pairs; before this occurs, the CBAM blocks are used to make the features more discriminative from both channel-wise and spatial-wise perspective. Moreover, the deep supervision with two DS layers are applied to assist the hidden layers in capturing more useful features.

Let I_{T1} and I_{T2} represent a pair of images in the same area at different time points $T1$ and $T2$, respectively, while y represents the label with change annotations. The flowchart of DSAMNet can be summed up as follows:

1) First, the images I_{T1} and I_{T2} are input into the weight-sharing feature extractor, with each obtaining a group of multiscale feature vectors $F_m = \{\text{feat}_m^1, \ldots, \text{feat}_m^4\}, m = T1, T2$.
2) Next, the vectors from the same timestamp are all merged into one to obtain feature pairs with the same

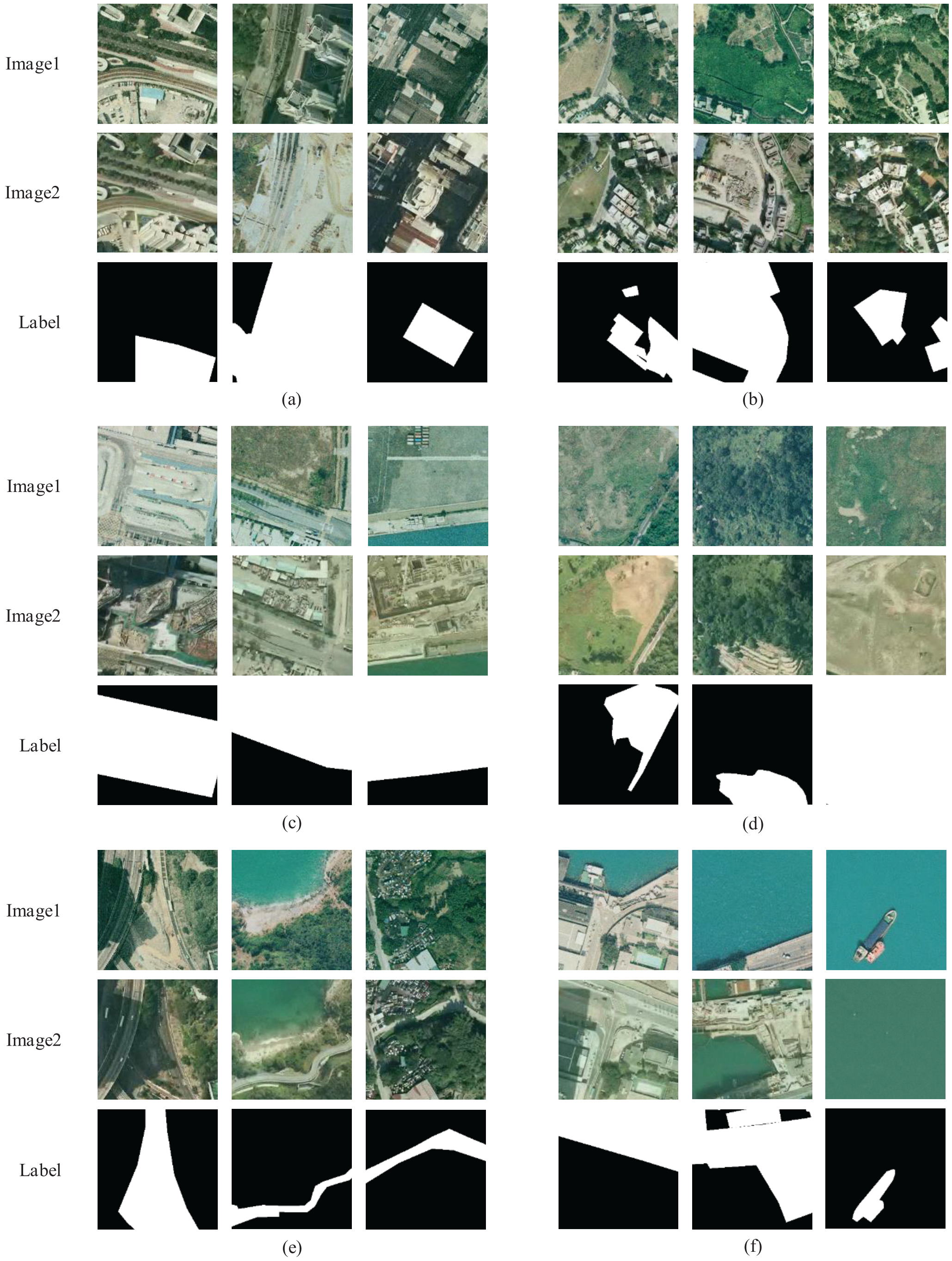


Fig. 1. Example samples (256 × 256) from the SYSU-CD dataset. Each column represents different change types: (a) Newly built urban buildings; (b) Suburban dilation; (c) Groundwork before construction; (d) Change of vegetation; (e) Road expansion; and (f) Sea construction.

dimension, Feat_{T1} and Feat_{T2}, where a CBAM block is applied to make them more discriminative. The metric module calculates a distance map D between Feat_{T1} and Feat_{T2}, whereas a batch contrastive loss (BCL) L_{BCL} is calculated according to the distance map D and the label y.

3) In the meantime, the absolute values of the two features $F_{\text{abs}} = \{\text{feat}^i_{T1} - \text{feat}^i_{T2}\}, i = 1, 2$ are input into the DS layers, yielding two intermediate change maps Cmap^1, Cmap^2, after which a dice loss L_{Dice} is calculated according to Cmap^1, Cmap^2, and the label y.

4) Finally, L_{BCL} and L_{Dice} are summed up together to facilitate a more accurate model training.

B. Feature Extractor

The encoder adopts a Siamese architecture with two weight-sharing branches to extract features from bitemporal

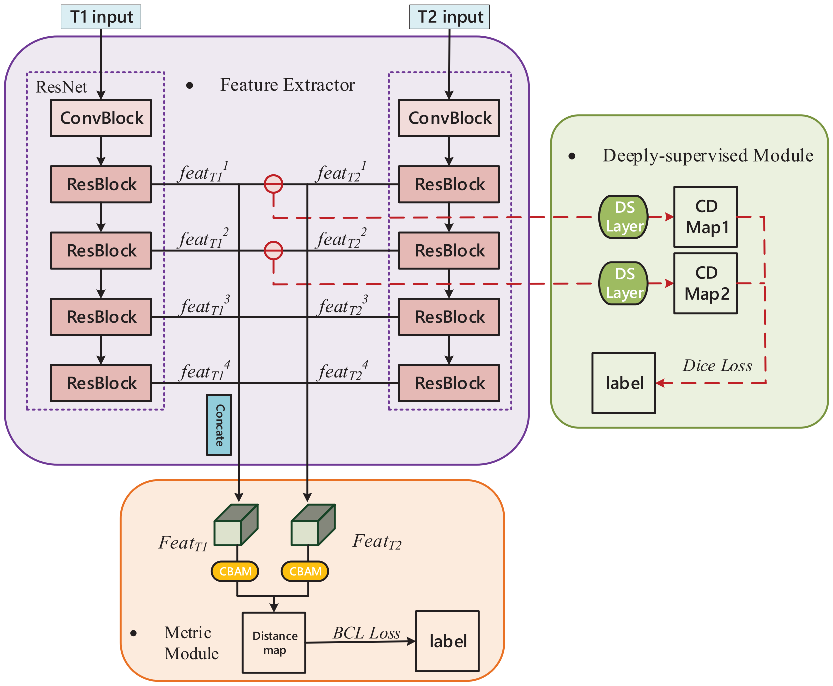


Fig. 2. Overview of the proposed DSAMNet. The feature extractor is extended to the Siamese structure to extract features of bitemporal image pairs. Features of different scales are stacked into the metric module after being resized to the same size, where two CBAMs are adopted to make them more discriminative. Two DS layers are employed in the DS module to assist the learning of the hidden layer.

image pairs. Many previous studies have demonstrated that borrowing a CNN feature extractor with pretrained parameters is conducive to model convergence [29], [50]. We therefore build our feature extractor based on ResNet [51] and load parameters from the pretrained ResNet-18 on ImageNet [14].

An increase in the depth of traditional CNNs is often accompanied by the problem of gradient disappearance or gradient explosion, which leads to network degradation. Theoretically, if deep layers learn an identical mapping, that is, the outputs are consistent with the inputs, then the network will not degrade with increasing depth. Therefore, taking an original block stacked with two layers [Fig. 3(a)] as an example, the output of the original block is as follows:

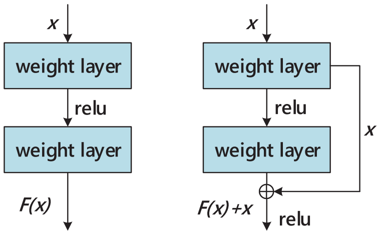
$$F(x) = W_2\sigma(W_1(x)) \tag{1}$$

where W_1 and W_2 represent the weight of the two layers, respectively, whereas σ is a rectified liner unit (ReLU) function.

If our goal is to achieve identity mapping in the original blocks, the parameters need to be adjusted to implement $F(x) = x$. However, it is comparatively more difficult to learn a nonlinear mapping. ResNet utilizes residual blocks [Fig. 3(b)], in which an identity shortcut connection is employed to transfer the input x directly to the output. Thus, the output of the residual block become:

$$F(x) + x = W_2\sigma(W_1(x)) + x. \tag{2}$$

In this case, the network needs to adjust its internal parameters to make $F(x) + x = x$, that is, $F(x) = 0$, which greatly simplifies the learning difficulty compared to the original blocks and thus improves the model performance.



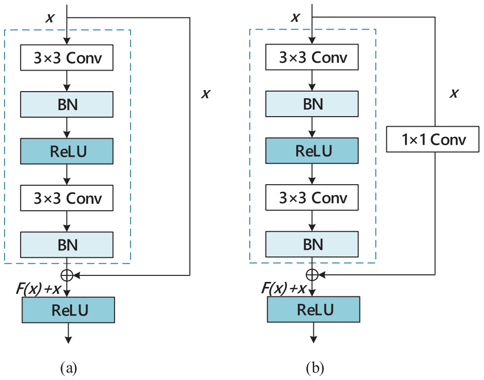
(a) (b)

Fig. 3. Structure of different blocks: (a) Original blocks and (b) Residual blocks.

After removing the global pooling layer as well as the fully connected layer from the initial ResNet-18, a variant of ResNet with 18 layers is obtained. Our feature extractor consists of five stages. First, a convolutional layer with a kernel size of 7 × 7 is used to extract low-level features from the input images; next to these are a batch normalization (BN) layer [52] and a ReLU [14]. To avoid the smoothing effect of the convolution operation and increase the reception field, a max-pooling layer with a stride of 2 is utilized to resize the features to half the size of the input image. Four basic blocks make up the remaining four stages, each of which consists of two layers. Basic blocks also take two forms, as Fig. 4 shows: one is to transfer the initial input x, while the formula is the same as (2). The other utilizes a 1 × 1 convolutional layer for dimension increase and downsampling purposes. At this point, the output of the residual block is expressed as follows:

$$F(x) + x = W_2\sigma(W_1(x)) + W_s x \tag{3}$$

where W_s denotes the weight of the 1 × 1 layer.



(a) (b)

Fig. 4. Structure of basic blocks in ResNet-18: (a) Simple connection and (b) Connection with 1 × 1 conv.

Note that the stride of the first two basic blocks is 2; this means that the feature map would be reduced by half, whereas that of the last two basic blocks is 1, meaning that the feature size is consistent with the input size. The depths of the four basic blocks are 64, 128, 256, and 512 in turn. Except for the output feature $feat4$ at the end of the feature extractor, which is 1/8 of the original input image size, the three intermediate feature maps of other basic blocks ($feat1$, $feat2$ and $feat3$) are then output for subsequent processing.

C. Metric Module

The metric module learns the change map by comparing the embedding eigenvectors of bitemporal images. Therefore, the construction of the bitemporal feature pairs for comparison is critical for the outcome. To make the best of sufficient spatial information in low-level features and rich semantic information in high-level features, a multiscale feature fusion strategy is exploited to reconstruct the paired features with the multilevel features from each branch of the feature extractor, where a 1 × 1 convolutional layer is first applied to unify the channel of each feature to 96, after which a bilinear interpolation is applied to resize their sizes to half of the input image. The features from the same branch are then sequentially concatenated and fed into a convolutional block, which consists of two convolutional layers with kernel sizes of 3 and 1, respectively. This reconstruction results in feature pairs with a channel size of 64.

As discussed above, feature pairs with inconspicuous representations might be difficult for the metric module to distinguish. Consequently, we introduce the CBAM to further fuse the multilevel features and make these feature pairs more distinguishable, which is a very lightweight module that does not incur excessive memory and computing overhead. As Fig. 5 shows, CBAM comprises two submodules, a channel attention submodule [Fig. 5(b)] and a spatial attention submodule [Fig. 5(c)], to help strengthen useful information in the extracted features in different dimensions. The channel attention submodule aims to capture channel-wise long-range contextual information through a channel attention map, which can be calculated using the following formula:

$$M_c(F) = \sigma(\mathrm{MLP}^r(\mathrm{AvgPool}(F)) + \mathrm{MLP}^r(\mathrm{MaxPool}(F))). \tag{4}$$

The F denotes an input feature of size C × H × W, on which an average pooling layer and a max pooling layer are utilized to generate two aggregated vectors of size C × 1 × 1. A weight sharing multilayer perception (MLP) module with a channel reduction ratio r is then applied to each vector to give weights to each channel. The MLP layer contains two 1 × 1 convolutional layers; the first of these reduces the channel of the input feature by r times, whereas the latter restores the channel number to the same size as the original input. The two layers are linked by a ReLU layer. The channel attention map $M_c(F)$ is obtained using the elementwise sum of the above two vectors and a sigmoid function σ. Finally, the original features will be multiplied with the channel attention map to obtain a channel-refined feature F'.

Similarly, the spatial attention submodule also adopts average pooling and max pooling for the first-step process in order to squeeze the input channel-refined feature F' to two 1 × H × W matrixes, which are concatenated together and forwarded into a convolutional layer with a kernel size of k. Finally, a sigmoid function is used to obtain the final spatial attention map $M_s(F')$. The formula can be denoted as follows:

$$M_s(F') = \sigma(f^{(k\times k)}(\mathrm{AvgPool}(F'); \mathrm{MaxPool}(F'))). \tag{5}$$

A more discriminative feature F'' will be obtained the channel sub-module and spatial sub-module are passed to CBAM successively. This process can be expressed as follows:

$$F' = M_c(F) \otimes F \tag{6}$$

$$F'' = M_s(F') \otimes F'. \tag{7}$$

Thereafter, a Euclidean distance map $Dist$ would be calculated based on the refined feature pairs according to the following formula:

$$\mathrm{Dist} = \sqrt{\left(F''_{T1} - F''_{T2}\right)\left(F''_{T1} - F''_{T2}\right)^T} \tag{8}$$

where F_{T1} and F_{T2} denote the CBAM-refined feature map of $T1$ and $T2$, respectively.

In the training stage, the distance map $Dist$ would be compared with the ground truth to obtain the contrastive loss for optimization, whereas in the prediction stage, a simple threshold segmentation would be applied on the distance map to obtain the change map.

D. DS Module

Generally speaking, most traditional end-to-end CNN networks only provide supervision at the output layer to train the entire network. However, because the training of hidden layers in the middle of the deep convolutional networks is non-transparent and lacks supervision, the hidden layer cannot efficiently learn the effective features, especially in a deep network, which affects the subsequent prediction [31]. Previous works have demonstrated the excellent ability of deep supervision to improve the effectiveness of hidden layers [32], [53], [54]; thus, we introduce deeply supervised nets (DSNs) [31] into our network. Rather than providing oversight only to the output layer, DSN aims to supply direct supervision

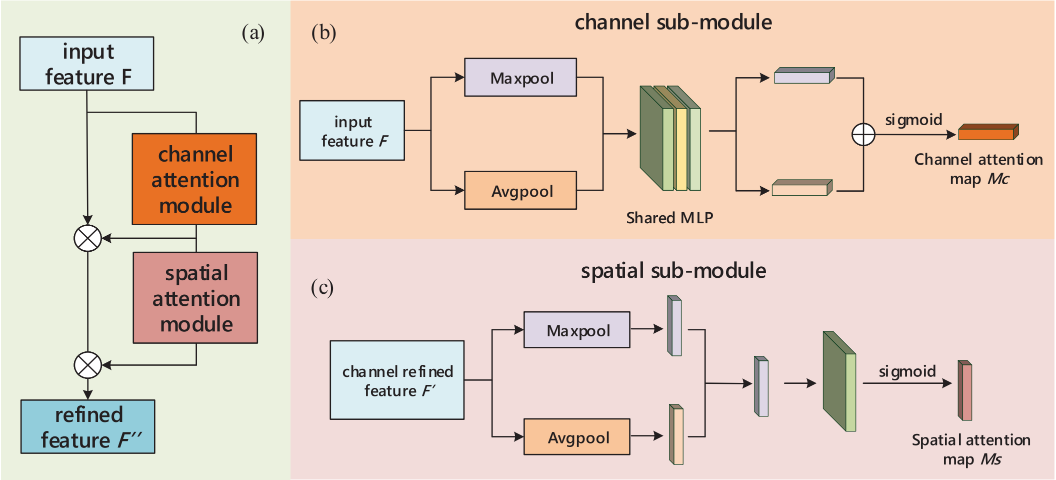


Fig. 5. Architecture of CBAM block: (a) Overview of CBAM block, (b) channel sub-module, and (c) spatial submodule.

for hidden layers by introducing a companion classifier, along with a companion objective, as a soft constraint for hidden layers during the training process, with the goal of aiding the learning of hidden layers and obtaining more robust and discriminative features.

More specifically, for a normal deep convolutional network with a total of M layers, we can denote the weights of all layers as $W = \{W^{(1)}, \ldots, W^{(M-1)}\}$, and the weight of the output layer as $w^{(\text{out})}$. The objective function of the present network is as follows:

$$P(W) = L(W, w^{(\text{out})}) \tag{9}$$

where $L(W, w^{(\text{out})})$ is the overall loss directly decided by $w^{(\text{out})}$, which depends on the weight of all layers W.

We now apply $M-1$ classifiers to all hidden layers to obtain additional outputs for deep supervision, then write the weights of the classifiers as $w = \{w^{(1)}, \ldots, w^{(M-1)}\}$. The companion objective function of the hidden layers is thus

$$Q(W) = \sum_{m=1}^{M-1} \alpha_m \ell(W, w^{(m)}) \tag{10}$$

where m marks the hidden layers, $\ell(W, w^{(m)})$ denotes the companion loss of the mth hidden layer, to which α_m provides a weight.

Different from the overall loss, each companion loss $\ell(W, w^{(m)})$, $w^{(m)}$ depends on the mth hidden layer as well as the previous hidden layers. Finally, the objective function of the DS network can be defined as follows:

$$F(W) = P(W) + Q(W). \tag{11}$$

As there is no decoder with an upsampling structure in our metric-based network, the DS network is integrated into the hidden layers of the feature extractor to conduct supervision. Considering that the supervision of all hidden layers will reduce the computational efficiency of the model, only two DS layers are added on the second- and the third-layer features of the feature extractor (referred as feat^1 and feat^2 in Fig. 2). In view of the properties of hierarchical features, these two intermediate features with both rich spatial and semantic information are more suitable to be selected for supervision. Each DS layer contains two 3 × 3 deconvolutional layers to upsample the feature maps to the size of the input image, along with a sigmoid layer to obtain the probability distribution of each pixel's category. To screen more effective change information for supervision and learning during the training process, the absolute differences between the selected features at the same level from each branch are forwarded into the DS layers. The outputs of these supervised layers were used to supervise parameter optimization along with the output of the metric model, which leads to the hidden layers having a better feature learning ability.

E. Loss Function

Given a set of training image pairs $X_n = \{(x_n^{t1}, x_n^{t2}), n = 1, \ldots, N\}$ and the ground truth $Y_n = \{y_n^{t1}, n = 1, \ldots, N\}$, our goal is to optimize the objective function for an accurate CD network. Note that a batch of distance maps and a batch of DS maps are obtained through our network, to which two different loss functions are applied.

Contrastive loss [55] is employed in the metric module to measure the similarity between the distance map and ground truth, which is commonly applied to paired data in the Siamese network. The BCL in the metric module can be expressed as follows:

$$L_{\text{BCL}}(X_n, Y_n) = \sum_{i,j=0}^{M} \frac{1}{2}\left[(1 - y_{i,j})d_{i,j}^2 + y_{i,j}\max(d_{i,j} - m)^2\right] \tag{12}$$

where $d_{i,j}$ represents the value of the distance map at point (i, j); here, $y_{i,j}$ represents the value of the label map at point (i, j), whereas M is the size of the distance map. Moreover, 0 denotes unchanged while 1 denotes changed, and m is the margin to filter out pixel pairs with a distance greater than this value.

Algorithm 1 Flowchart of DSAMNet

```
Input:
    A set of image pairs of the $n$th batch $\{x_n^{t1}, x_n^{t2}\}$ and corresponding ground truth $y_n$;
    Parameters of learning rate $lr$, number of training epochs $N$, batch size $batch$, reduction rate in CBAM $r$, and kernel size in CBAM $k$
Output:
    a distance map $Dist_n$; two change maps $Cmap_n^1$, $Cmap_n^2$

    for each $m \in [1, N]$ do
        extract multiscale features $feat_n^{t1}$ from $x_n^{t1}$;
        extract multiscale features $feat_n^{t2}$ from $x_n^{t2}$;

        obtain refined feature map $F''_{t1n}$ and $F''_{t2n}$ by (4)–(7);

        calculate a distance map $Dist_n$ by (8);
        calculate the contrastive loss $L_{BCL-n}$ by (12);

        obtain two intermediate CD maps, $Cmap_n^1$, $Cmap_n^2$
        calculate the dice loss $L_{Dice-n}$ by (13);

        propagate back the overall loss $L$ calculated by (14)
    end for
```

According to this formula, the loss of the unchanged pixel pairs ($y = 0$) depends on the distance between the pixel pairs $d_{i,j}$. A large $d_{i,j}$ will result in a large loss at (i, j), which will help to reduce the distance between unchanged samples. A similar strategy is used for changed pixel pairs ($y = 1$) to increase the distance between change samples. The contrastive loss can therefore well express the matching degree of paired samples, which is greatly helpful for achieving accurate change extraction.

Under the influence of class imbalance in CD, the model's training direction will be dominated by the majority of "unchanged" pixels, neglecting the minority "change" information and thus leading to a less-efficient model. The dice loss is adopted in the supervision module to overcome class imbalance; this approach uses a dice coefficient to measure the degree of similarity between the prediction map and the ground truth and ranges between 0 and 1. The objective function in the deep supervision module can be denoted as follows:

$$L_{\text{Dice}}(X_n, Y_n) = 1 - \frac{1}{m}\sum_{j=1}^{m} \frac{2\sum_{i=1}^{N} \hat{y}_{i,j} y_{i,j}}{\sum_{i=1}^{N} \hat{y}_{i,j} + \sum_{i=1}^{N} y_{i,j}} \qquad (13)$$

where $\hat{y}$ and y represents prediction map and target label, respectively.

Finally, the objective function of the network is as follows:

$$L(X_n, Y_n) = L_{\text{BCL}}(X_n, Y_n) + \lambda L_{\text{Dice}}(X_n, Y_n) \qquad (14)$$

where λ is a factor used to regulate the influence of the DS module.

The Algorithm 1 outlines the optimization process of DSAMNet.

V. Experimental and Results

In this section, the datasets and comparison algorithms employed in the following experiments are first illustrated. We then provide a brief description of the implementation details and evaluation metrics. Finally, the experimental results are analyzed in detail.

A. Datasets

Two groups of experiments were designed to verify the effectiveness of our proposed DSAMNet. The first group of experiments was conducted on the CDD dataset, which is a widely used CD dataset, to evaluate the validity of DSAMNet. All baselines were then compared on the SYSU-CD dataset to verify the validity of our dataset and to further validate our model. The details of the two datasets are as follows:

1) CDD Dataset: The CDD dataset [34] consists of 16 000 pairs of real season-varying remote sensing images, each with an image size of 256 × 256 and a spatial resolution of 0.03–1 m, with 10 000 and 3000 pairs used for training and validation, respectively, whereas the remaining 3000 pairs are used for testing. The main types of changes in CDD are summarized and shown in Fig. 6.

2) SYSU-CD Dataset: As discussed in Section III, the SYSU-CD dataset contains 20 000 pairs of orthographic aerial image patches with a size of 256 × 256 and resolution of 0.5 m. In our experiments, the ratio of samples used for the training, validation, and test sets is set to 6:2:2. Multiple changes types in relatively complex scenarios are provided in this dataset as shown in Fig. 1.

B. Comparative Methods

To demonstrate the superiority of DSAMNet, the following five state-of-the-art CD methods are selected for comparison purposes and introduced in brief:

1) Fully Convolutional-Early Fusion (FC-EF): FC-EF [22] is proposed based on U-Net architecture, in which the bitemporal images are cascaded as a multiband image for input. Skip connections are used to progressively transport the multiscale features from the encoder to the decoder to recover spatial information.

2) Fully Convolutional-Siamese-Concatenation (FC-Siam-conc): As a variation of the FC-EF model, FC-Siam-conc [22] extracts the features of bitemporal images with a Siamese encoder rather than EF. The features at the same level from the encoder are then concatenated to the decoder.

3) Fully Convolutional-Siamese-Difference (FC-Siam-diff): Different from FC-Siam-conc, the skip connections of FC-Siam-diff [22], which is another variety of FC-EF model, transport the absolute difference between bitemporal features.

4) BiDateNet: BiDateNet [27] is a FCN with a U-Net structure for CD, which introduces Long Short-Term Memory (LSTM) convolutional blocks into the skip connection for improved temporal pattern investigation between the bitemporal images.

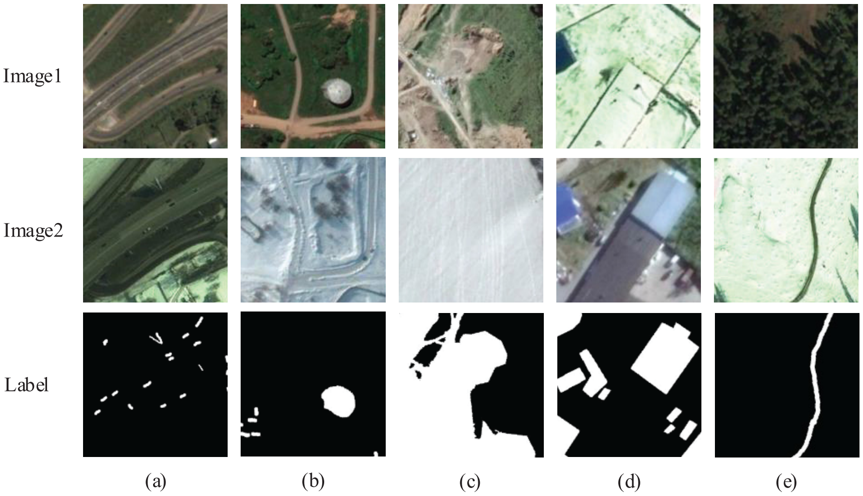


Fig. 6. Example samples (256 × 256) from CDD. Each column represents different types of change: (a) Car change; (b) tank change; (c) land change; (d) building change; and (e) road change.

5) Spatial-Temporal Attention-Based Network (STANet): STANet [28] is a metric-based CD method, which learns the change map based on the distances between the features. A spatial-temporal attention module is utilized to learn the spatial-temporal relationships between the bitemporal images to generate more discriminative features.

In summary, four U-Net-based methods and two metric-based methods (including DSAMNet) would be compared in the experiment to compare two different ways of obtaining change maps: up-sampling and metric learning. Additionally, the effectiveness of different attention mechanisms on CD can also be tested from the three attention-integrated methods, including BiDateNet, STANet, and DSAMNet.

C. Implementation Details

We use the PyTorch library for all experiments. The parameters used in comparison methods are as consistent as possible with the original literatures. However, because of memory limitations, the batch size of STANet is set to 4 in our experiment. As for the DSAMNet, the Adam optimizer was adopted with an initial learning rate of 0.0001 on CDD dataset and that of 0.0005 on SYSU-CD dataset. A batch size of 8 sample pairs was utilized to facilitate faster model convergence. The reduction ratio r and kernel size k in the CBAM blocks were 8 and 7, respectively. The margin m in the BCL took a value of 2, and the threshold to segment the distance map was set to 1. All of our experiments are conducted on the GeForce RTX 2080ti to accelerate model training.

To evaluate the performance of the proposed methods, we utilize four typical metrics: namely, precision, recall, F1-score, and IoU. More specifically, precision reflects the false alarm rate, recall reflects the miss alarm rate of the model, whereas F1 takes both indices into account; therefore, a larger F1 score indicates a better model. IoU represents the overlap rate of the change class on the detection map and the ground truth. In general, an IoU larger than 0.5 denotes a good result. These four metrics can be calculated as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{F1} = \frac{2\text{presicion} \cdot \text{recall}}{\text{presicion} + \text{recall}} \tag{17}$$

$$\text{IoU} = \frac{\text{DetectionResult} \cap \text{GroundTruth}}{\text{DetectionResult} \cup \text{GroundTruth}} \tag{18}$$

where TP, FP, TN, and FN refer to true positives, false positives, true negatives, and false negatives, respectively.

D. Comparisons on CDD Dataset

The quantitative results for the precision, recall, F1, and IoU of all methods are summarized in Table II. As Table II shows, FC-EF obtains the lowest F1 and IoU of 78.65% and 64.81% among the compared methods, which is followed by FC-Siam-diff, which obtains an F1 of 82.93% and an IoU of 70.84%. This shows that the Siamese encoder can slightly improve the model accuracy here. The FC-Siam-conc scores 85.90% in terms of F1 and 75.28% on IoU, which denotes that more useful information can be maintained through concatenation than difference and transferred to the decoder. The BiDateNet obtains the highest precision of 95.98%, as well as an F1 and an IoU of 90.01% and 81.83%, respectively; this demonstrates that LSTM blocks are well able to capture the temporal change pattern and improve the accuracy. The STANet based on metric learning obtains an F1 of up to 91.44% and an IoU of up to 84.23%, higher than the scores obtained by the above UNet-based methods. Our method achieved the highest F1 of 93.69% and an IoU of 88.13% among all the compared methods; these figures are 2.25% and 3.90% higher than those obtained by STANet.

Fig. 7 provides a more intuitive picture of each method's performance on the CDD datasets. Generally speaking, FC-EF and its two variants, FC-Siam-conc and FC-Siam-Diff, can identify relatively obvious changes; however, many small

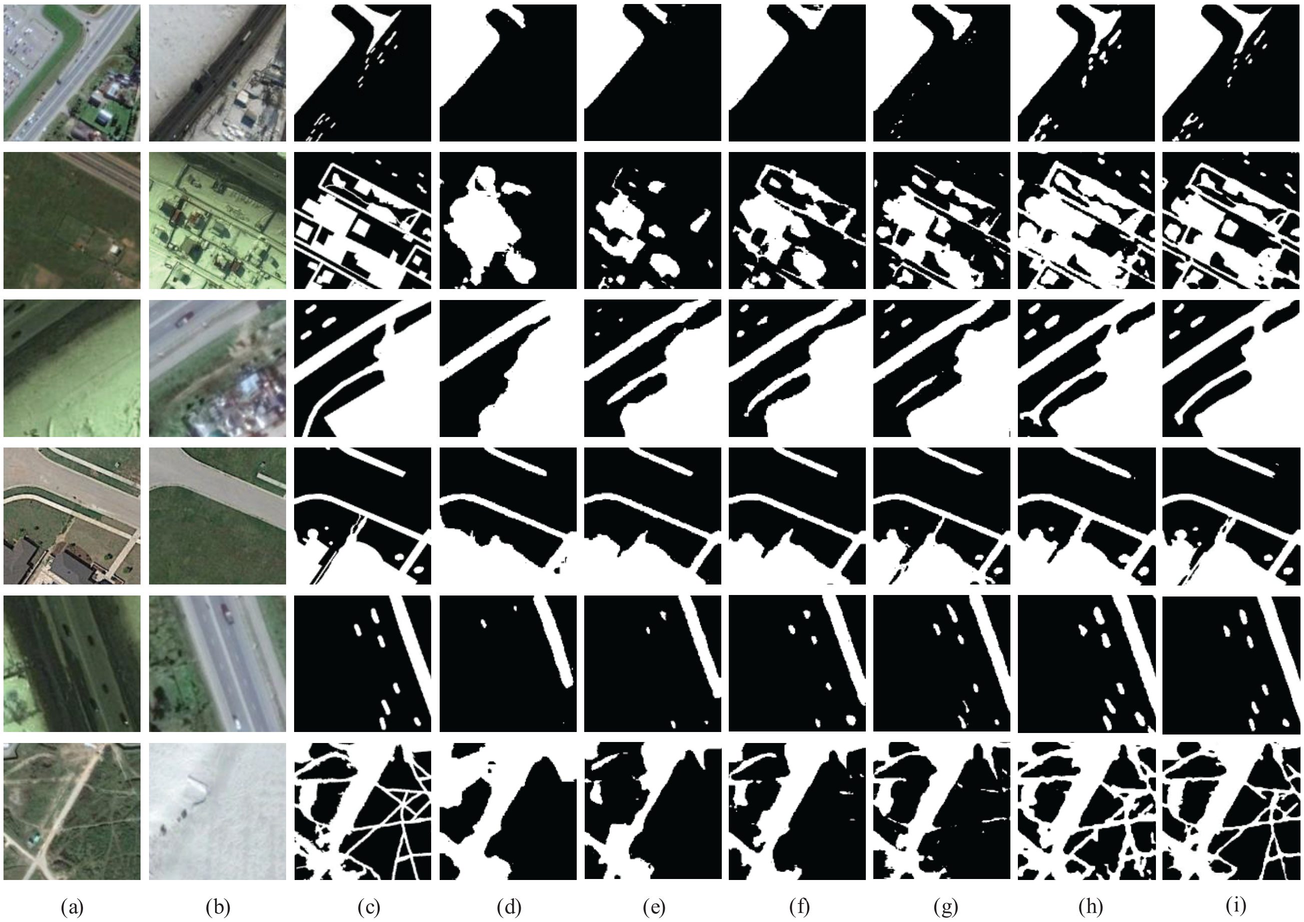


Fig. 7. Visual comparisons on the CDD dataset: (a) Image at time1; (b) Image at time2; (c) Ground truth; (d) FC-EF; (e) FC-Siam-diff; (f) FC-Siam-conc; (g) BiDateNet; and (h) STANet. (i) DSAMNet.

TABLE II
EXPERIMENTAL RESULTS ON CDD DATASET

Method	Pre(%)	Rec(%)	F1(%)	IoU(%)
FC-EF	90.41	69.60	78.65	64.81
FC-Siam-diff	94.76	73.71	82.93	70.84
FC-Siam-conc	92.15	80.44	85.90	75.28
BiDateNet	**95.98**	84.74	90.01	81.83
STANet	89.28	**93.71**	91.44	84.23
DSAMNet	94.54	92.77	**93.69**	**88.13**

changes, such as cars and alleys, are missed in the change map, leading to the low recall rate of these three methods. Among the UNet-based methods, the BiDateNet has the best performance in the extraction of cars and alleys; this indicates that LSTM blocks can help to capture small changes. The metric-based STANet can successfully extract changes at different scales, including cars, roads, and buildings. However, because of the degradation of spatial context information during feature extraction, the STANet are hard to keep precise boundaries of these small objects, which are relatively rough and supersaturated compared to the ground truth. According to Fig. 7, our proposed method is able to recognize scale-variance changes with finer boundaries. This demonstrates that the integration of DS module can help to restore the spatial information and improve the CD accuracy.

E. Comparisons on SYSU-CD Dataset

As can be seen from Table III, the proposed method also outperforms all baselines on the SYSU-CD dataset, achieving the highest F1 and IoU scores of 78.18% and 64.18%, respectively. The second-ranked STANet obtains an F1 of 77.37% and an IoU of 63.09%, which further proves the advancement of metric learning for CD. The BiDateNet achieves best performance among the UNet baselines with an F1 of 76.94% and an IoU of 62.52%, which are 0.59% and 0.77% higher than those of the FC-Siam-conc. Meanwhile, different from the results on the CDD dataset, FC-EF performs better than FC-Siam-diff on the SYSU-CD dataset, which obtains an F1 of 75.07% and an IoU of 60.09%. Despite achieving the highest precision of 89.13%, the FC-Siam-diff has the lowest F1 and IoU; this may attribute to that in relatively complex scenarios of SYSU-CD dataset, feature difference can lead to excessive filtering of useful change information and missed alarms.

Fig. 8 further demonstrates the behavior of different methods on the SYSU-CD dataset. As can be seen from Row 1 of Fig. 8, the FC-EF has a deficiency in identifying newly built urban buildings compared with other methods. According to the ground truth, there are many omissions in the result of FC-Siam-diff, which is consistent with its high precision and low recall in Table II. The FC-Siam-conc and the BiDateNet are able to extract major changes but show relatively poor

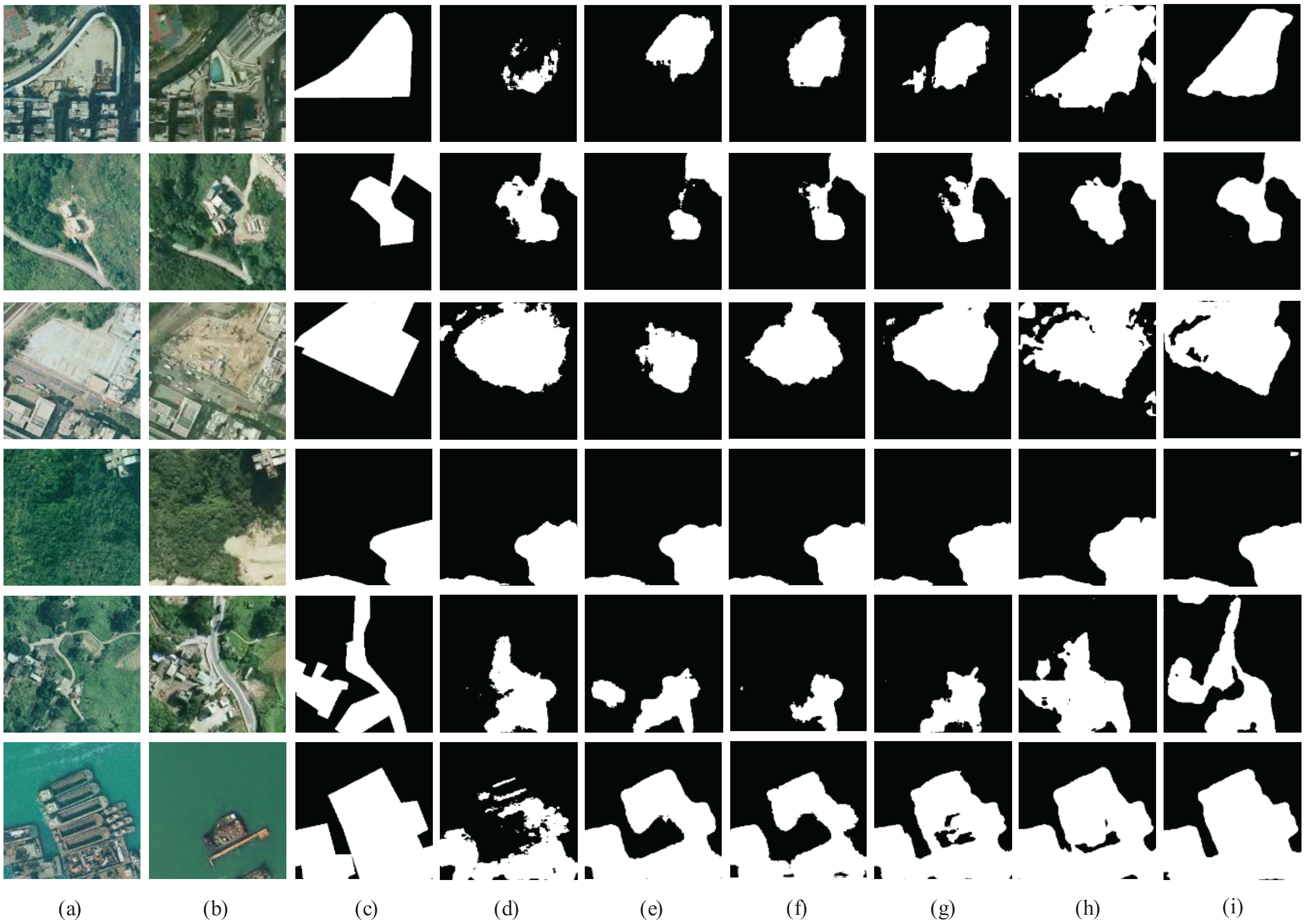


Fig. 8. Visualization comparisons on SYSU-CD dataset: (a) Image at time1, (b) Image at time2, (c) Ground truth, (d) FC-EF, (e) FC-Siam-diff, (f) FC-Siam-conc, (g) BiDateNet, (h) STANet, and (i) DSAMNet.

TABLE III
EXPERIMENTAL RESULTS ON SYSU-CD DATASET

Method	Pre(%)	Rec(%)	F1(%)	IoU(%)
FC-EF	74.32	75.84	75.07	60.09
FC-Siam-diff	**89.13**	61.21	72.57	56.96
FC-Siam-conc	82.54	71.03	76.35	61.75
BiDateNet	81.84	72.60	76.94	62.52
STANet	70.76	**85.33**	77.37	63.09
DSAMNet	74.81	81.86	**78.18**	**64.18**

TABLE IV
ABLATION STUDY ON DSAMNET

Method	CDD		SYSU-CD	
	F1(%)	IoU(%)	F1(%)	IoU(%)
Base	92.74	86.46	75.02	60.03
Base+DS	92.90	86.74	75.78	61.01
Base+CBAM	93.08	87.05	76.62	62.10
DSAMNet	**93.69**	**88.13**	**78.18**	**64.18**

performance in some complex scenarios. The STANet and the DSAMNet can capture more complete change areas in most cases, including village and road dilation (Row 5 of Fig. 8). Moreover, they also work well at recognizing changes with less disparate appearance, such as the newly built urban buildings (Row 1 of Fig. 8) and the groundwork before construction (Row 3 of Fig. 8). Similar to the results of CDD datasets, the DSAMNet does a best job of maintaining boundary information of changes among all the baselines.

F. Ablation Study of DSAMNet

On the basis of metric learning, the DSAMNet integrates both CBAM blocks and DS module for accurate CD. We therefore design ablation experiments on DSAMNet to verify both the validity of CBAM and DS layers. In the following experiment, the "Base" baseline denotes the basic model without CBAM and deep supervision. The "Base + DS" model is adopted as the second baseline by introducing DS module into the "Base" model, whereas the "Base + CBAM" represents model with CBAM integration.

As can be seen from Table IV, the incorporation of both DS layers and CBAM blocks can improve the model performance on both datasets. More specifically, the F1 and IoU can be improved by 0.16% and 0.28% on the CDD dataset and by 0.76% and 0.98% on the SYSU-CD dataset after adding DS layers. It shows that deep supervision on the hidden layers do enhance the ability of the model. Besides, the CBAM blocks can improve the F1 and IoU of the CDD dataset by 0.34% and 0.59% and those of the SYSU-CD dataset by 1.60% and 2.07%, respectively; this indicates that CBAM blocks can contribute substantially to the subsequent metric learning

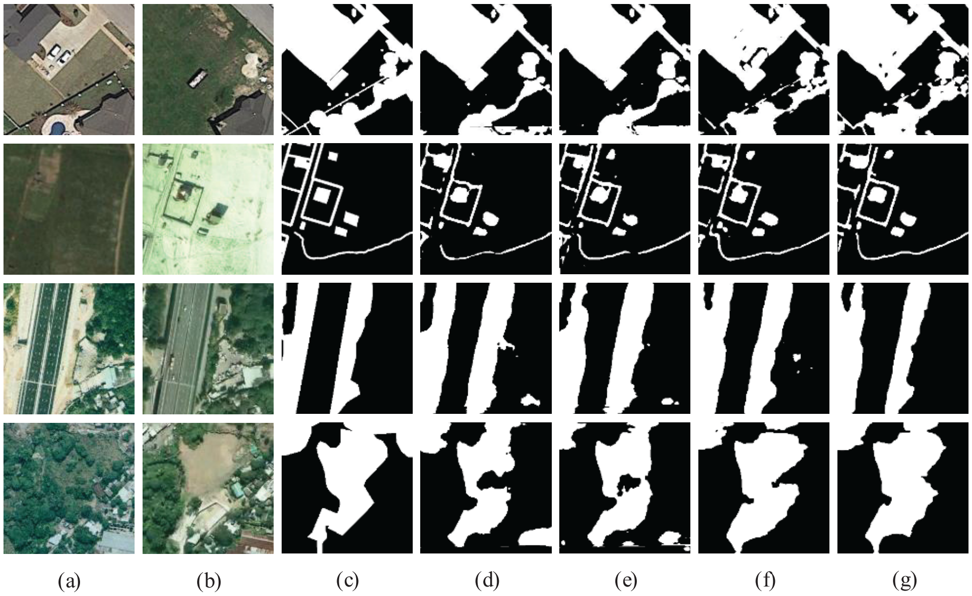


Fig. 9. Examples of the ablation experiment on the proposed method. The first two rows present the samples from CDD dataset, whereas the last two rows are from the SYSU-CD dataset: (a) Image at time1; (b) Image at time2; (c) Ground truth; (d) Base; (e) Base + DS; (f) Base + CBAM; and (g) DSAMNet.

by making the feature pairs more distinguishable from each other. Notably, compared to the "Base" model, the F1 and IoU of the DSAMNet with both DS layer and CBAM blocks integrated are increased by 0.95% and 1.67% on CDD dataset, and 3.16% and 4.15% on SYSU-CD dataset, respectively. The great improvement of the DSAMNet not only further proves the effectiveness of CBAM blocks and DS layers, but also proves the gain effect of their combination.

Fig. 9 presents the ablation result on the two datasets. Compared with the Base model, both the second and third baselines can reduce the missed and false alarms in the change maps. Therefore, with both the DS layers and CBAM blocks, the DSAMNet can largely improve the completeness and accuracy of the change results; this further verifies that through the DS layers to improve the feature extraction ability of hidden layers, and combined with CBAM blocks to further enhance the expression of effective information, the DSAMNet can effectively extract more accurate change maps.

VI. Discussion

In this section, a group of sensitivity experiment is designed for parameter settings on the hybrid loss function and a discussion on the CDD and SYSU-CD datasets is conducted at the end.

A. Sensitivity Experiments on the Hybrid Loss

TABLE V
SENSITIVITY EXPERIMENTS ON LOSS FUNCTION

λ[1]	CDD		SYSU-CD	
	F1(%)	IoU(%)	F1(%)	IoU(%)
0	93.08	87.05	76.62	62.10
0.1	**93.69**	**88.13**	77.54	63.31
0.3	93.45	87.71	**78.18**	**64.18**
0.5	93.30	87.44	76.60	62.08
0.7	93.25	87.36	76.88	62.45
0.9	93.33	87.49	76.60	62.07

[1] λ is the coefficient of dice loss in the hybrid loss function.

The loss function is adopted to evaluate the degree to which the prediction differs from the ground truth; this metric plays a decisive role in ensuring that the model converges fast and stably during the training process, which has a highly significant influence on the final model performance. As described in Section IV, we employ a hybrid loss, which combines the contrastive loss and the dice loss, for the training of our proposed network; in this approach, a parameter is introduced to balance the effect of the dice loss. To explore the influence of different values on the training of DSAMNet, we conduct comparative experiments on the two datasets by setting different λ values. In the training process, the deeply supervised module is adopted as an auxiliary module to improve the feature extraction ability of the model through the supervision of the hidden layer, so as to help the metric module to extract more accurate changes. Therefore, we set the value of λ between (0, 1) to distinguish the different role of the DS module and the metric module in the model training. These results are collected and presented in Table V. Note that when λ is set to 0, the network is equivalent to the third baseline "Base + CBAM" in Section V-F.

On the CDD dataset, the accuracies of all models with the DS module are improved to a certain extent. Therefore, the proposed network achieves the highest F1 and IoU when $\lambda = 0.1$, which are 0.61% and 1.08% higher than those when $\lambda = 0$. Then, when λ continues to increase, the gain of the DS module to the model gradually decreases. To be more specific, when $\lambda = 0.7$, the gain on the F1 and IoU are reduced to 0.17% and 0.31% compared to those of $\lambda = 0$. On the SYSU-CD dataset, the highest F1 and IoU are achieved when

$\lambda = 0.3$, representing an improvement of 1.19% and 1.69% compared to when $\lambda = 0$. Similarly, as λ continues to increase, the accuracy of the model is only slightly improved or even decreased a bit; this indicates that the influences of the λ value on different datasets are not the same because of the nature of the dataset itself, and the SYSU-CD dataset may be more sensitive to the λ value in the hybrid loss.

B. Discussion on the SYSU-CD Dataset

The experiments on the SYSU-CD dataset have proven the validity of SYSU-CD by means of a comprehensive comparison that takes both quantitative and visual form. However, we also note that the best F1 performance by DSAMNet on SYSU-CD is about 15.51% lower than that of the CDD dataset. In this regard, we analyze the reasons for this as follows:

1) Challenges brought about by complex scenes in the SYSU-CD dataset. One highlight of the SYSU-CD dataset is that the images contain many high-rises and dense buildings, which are insufficient in many existing datasets. In this case, the CD accuracy is to a certain extent limited by the complex environments.
2) Difficulties brought about by multiple change types in the SYSU-CD dataset. It goes without saying that CDD is an excellent CD dataset. For its part, the SYSU-CD dataset contains more complex and confusing types of change, as mentioned above, which also contributes to the relatively lower detection rate of SYSU-CD.

In summary, despite complicated detection scenes and multiple change types, we also achieved relatively good CD results on the SYSU-CD dataset. Moreover, there is also an urgent need to detect changes under such complex scenarios (such as megacities in China); for these situations, the SYSU-CD can provide an effective benchmark.

VII. Conclusion

In this article, a new DL-based method called DSAMNet is proposed for bitemporal remote sensing CD. We also provide a new benchmark dataset, SYSU-CD, which largely complements existing CD datasets in terms of image resolution, change type, and dataset volume. DSAMNet contains a CBAM-integrated metric module to learn a change map directly from features obtained using the feature extractor, as well as an auxiliary deep supervision module used to generate change maps with more spatial information. A hybrid loss is adopted to combine these two modules in the training process. Experimental results demonstrate that the proposed DSAMNet outperforms other state-of-the-art methods on both the CDD and SYSU-CD datasets. The CBAM blocks in the metric module can effectively make features more discriminative, thereby assisting with the learning of the metric module. Moreover, the DS module can make good use of the information contained in intermediate features, thereby further improving the change maps learned by the metric module. In the future, we will explore CDDs with semantic change information and develop effective semantic CD algorithms that can meet the needs of more diversified scenarios.

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.-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.

[3] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998.

[10] T. Habib, J. Inglada, G. Mercier, and J. Chanussot, "Support vector reduction in SVM algorithm for abrupt change detection in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 606–610, Jul. 2009.

[11] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS J. Photogramm. Remote Sens.*, vol. 61, no. 2, pp. 125–133, Nov. 2006.

[12] J. Im and J. Jensen, "A change detection model based on neighborhood correlation image analysis and decision tree classification," *Remote Sens. Environ.*, vol. 99, no. 3, pp. 326–340, Nov. 2005.

[13] K. Wessels *et al.*, "Rapid land cover map updates using change detection and robust random forest classifiers," *Remote Sens.*, vol. 8, no. 11, p. 888, Oct. 2016.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

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

[16] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, Apr. 2019.

[17] Q. Shi, X. Liu, and X. Li, "Road detection from remote sensing images by generative adversarial networks," *IEEE Access*, vol. 6, pp. 25486–25494, 2018.

[18] G. Cheng, Z. Li, J. Han, X. Yao, and K. Li, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.

[19] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[20] 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.

[21] Y. Long *et al.*, "DiRS: On creating benchmark datasets for remote sensing image interpretation," 2020, *arXiv:2006.12485*. [Online]. Available: http://arxiv.org/abs/2006.12485

[22] R. C. Daudt, B. L. 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.

[23] R. C. Daudt, B. L. 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.

[24] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.

[25] 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.

[26] 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.

[27] 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.

[28] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[29] 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.

[30] 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.

[31] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Statist.*, 2015, pp. 562–570.

[32] Q. Zhu, B. Du, B. Turkbey, P. L. Choyke, and P. Yan, "Deeply-supervised CNN for prostate segmentation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 178–184.

[33] C. Benedek and T. Sziranyi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.

[34] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 4176–4179.

[35] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. 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, pp. 565–571, May 2018.

[36] 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.

[37] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura, "Damage detection from aerial images via convolutional neural networks," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 5–8.

[38] J. López-Fandiño, A. S. Garea, D. B. Heras, and F. Argüello, "Stacked autoencoders for multiclass change detection in hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1906–1909.

[39] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2019.

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

[41] P. Hofmann, J. Strobl, T. Blaschke, and H. Kux, "Detecting informal settlements from QuickBird data in Rio de Janeiro using an object based approach," in *Object-Based Image Analysis*. Berlin, Germany: Springer, 2008, pp. 531–553.

[42] P. R. Coppin and M. E. Bauer, "Digital change detection in forest ecosystems with remote sensing imagery," *Remote Sens. Rev.*, vol. 13, nos. 3–4, pp. 207–234, Apr. 1996.

[43] R. Jackson, "Spectral indices in N-space," *Remote Sens. Environ.*, vol. 13, no. 5, pp. 409–421, Nov. 1983.

[44] E. P. Crist, "A TM tasseled cap equivalent transformation for reflectance factor data," *Remote Sens. Environ.*, vol. 17, no. 3, pp. 301–306, Jun. 1985.

[45] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[46] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[47] F. Rahman, B. Vasu, J. Van Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 958–962.

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

[49] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From W-Net to CDGAN: Bitemporal change detection via deep learning techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1790–1802, Mar. 2020.

[50] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional Siamese networks for change detection of high resolution satellite images," 2020, *arXiv:2003.03608*. [Online]. Available: http://arxiv.org/abs/2003.03608

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

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[53] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," 2015, *arXiv:1505.02496*. [Online]. Available: http://arxiv.org/abs/1505.02496

[54] C. Zhang *et al.*, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[55] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

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, and transfer learning.

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.

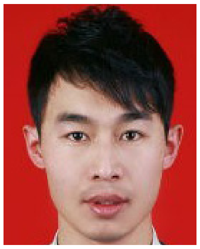
Her research interests include intelligent understanding of remote sensing images, change detection, and domain adaptation.

Shengchen Li is pursuing the B.S. degree with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China.

His research interests include urban remote sensing and deep learning.

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.



Fei Wang received the Ph.D. degree in cartography and geographic information system from the State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, China, in 2014.

After his Ph.D. degree he worked with the University of Xinjiang, Urumqi, as an Associate Professor in land resource management, geographic information system (GIS), and remote sensing for six years. He has undertaken a number of national-level scientific research projects, including information mining of remote sensing images, 3-D soil apparent conductivity, safety paradigm of water use, carbon cycle process, integration of remote sensing and hydrological models, optimization of village layout under rural revitalization, etc. More than 30 academic articles have been published. He served as a Reviewer for internationally renowned journals such as *Geoderma*, *Catena*, *Science of The Total Environment*, and *Remote Sensing*.

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 a 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 Highly Cited Author in the Institute for Scientific Information (ISI). 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 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 the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter.