Exchanging Dual-Encoder–Decoder: A New Strategy for Change Detection With Semantic Guidance and Spatial Localization

Sijie Zhao, Xueliang Zhang, *Member, IEEE*, Pengfeng Xiao, *Senior Member, IEEE*, and Guangjun He

***Abstract*— Change detection is a critical task in earth observation applications. Recently, deep-learning-based methods have shown promising performance and are quickly adopted in change detection. However, the widely used multiple encoders and single decoder (MESD) as well as dual-encoder–decoder (DED) architectures still struggle to effectively handle change detection well. The former has problems of bitemporal feature interference in the feature-level fusion, while the latter is inapplicable to intraclass change detection (ICCD) and multiview building change detection (MVBCD). To solve these problems, we propose a new strategy with an exchanging DED (EDED) structure for binary change detection with semantic guidance and spatial localization. The proposed strategy solves the problems of bitemporal feature inference in MESD by fusing bitemporal features in the decision level and the inapplicability in DED by determining changed areas using bitemporal semantic features. We build a binary change detection model based on this strategy and then validate and compare it with 18 state-of-the-art change detection methods on six datasets in three scenarios, including ICCD datasets (CDD and SYSU), single-view building change detection (SVBCD) datasets (WHU, LEVIR-CD, and LEVIR-CD+), and an MVBCD dataset (NJDS). The experimental results demonstrate that our model achieves superior performance with high efficiency and outperforms all benchmark methods with F1-scores of 97.77%, 83.07%, 94.86%, 92.33%, 91.39%, and 74.35% on CDD, SYSU, WHU, LEVIR-CD, LEVIR-CD+, and NJDS datasets, respectively. The code of this work will be available at https://github.com/NJU-LHRS/official-SGSLN.**

***Index Terms*— Change detection, deep learning, high-spatial-resolution remote sensing, semantic guidance, spatial localization.**

I. Introduction

CHANGE detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times [1]. It is crucial in applications such as urban expansion investigations [2], land-use planning [3], and disaster damage assessments [4]. Binary change detection is the process of identifying changed objects of interest given binary labels, which is basic but of great significance in change detection. Binary change detection can be classified into intraclass change detection (ICCD) and specific-class change detection, where the former detects all categories of changed objects, and the latter detects specific categories of changed objects. Since binary labels only provide the change information of changed objects rather than their semantic information, ICCD faces great challenges in detecting multiple categories of changed objects. Building change detection occupies an important position in specific-class change detection, which is important for urban planning and monitoring illegal construction [5]. According to the imaging angles of multitemporal remote-sensing images, building change detection can be classified into single-view building change detection (SVBCD) with similar imaging angles and multiview building change detection (MVBCD) with large differences in imaging angles, where the latter is common and faces great challenges in very high-resolution remote-sensing images. In SVBCD, the edge parts of changed buildings are difficult to accurately detect due to factors such as building shadows and the dense distribution of buildings. In MVBCD, because of the different imaging angles of multitemporal remote-sensing images, the same building has large spatial differences in the bitemporal images, leading to confusion about real changes and thus false positives.

In recent years, deep-learning methods have been quickly adopted in remote sensing due to the advent of massive remote-sensing data and the rapid development of deep learning [6], [7]. A large number of deep-learning-based change detection methods have been developed for binary change detection [8], [9], [10], [11]. There are two types of neural networks widely used in binary change detection: a multiple-encoder and single-decoder (MESD) network [8], [12] and a dual-encoder–decoder (DED) network [11], [13].

MESD consists of multiple encoders with shared weights and a single decoder for change detection. Bitemporal semantic features are extracted in multiple encoders and fused in the feature level in the single decoder to identify the changed areas, as shown in Fig. 1(a). This network suffers from problems in the feature-level fusion: when fusing bitemporal encoder features, the changed object features in one temporal

Manuscript received 12 July 2023; revised 25 September 2023; accepted 23 October 2023. Date of publication 26 October 2023; date of current version 6 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297 and in part by the Fundamental Research Funds for the Central Universities under Grant 020914380119. *(Corresponding author: Xueliang Zhang.)*

Sijie Zhao, Xueliang Zhang, and Pengfeng Xiao are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: zsj@smail.nju.edu.cn; zxl@nju.edu.cn; xiaopf@nju.edu.cn).

Guangjun He is with the State Key Laboratory of Space-Ground Integrated Information Technology, Space Star Technology Company Ltd., Beijing 100095, China (e-mail: hgjun_2006@163.com).

Digital Object Identifier 10.1109/TGRS.2023.3327780

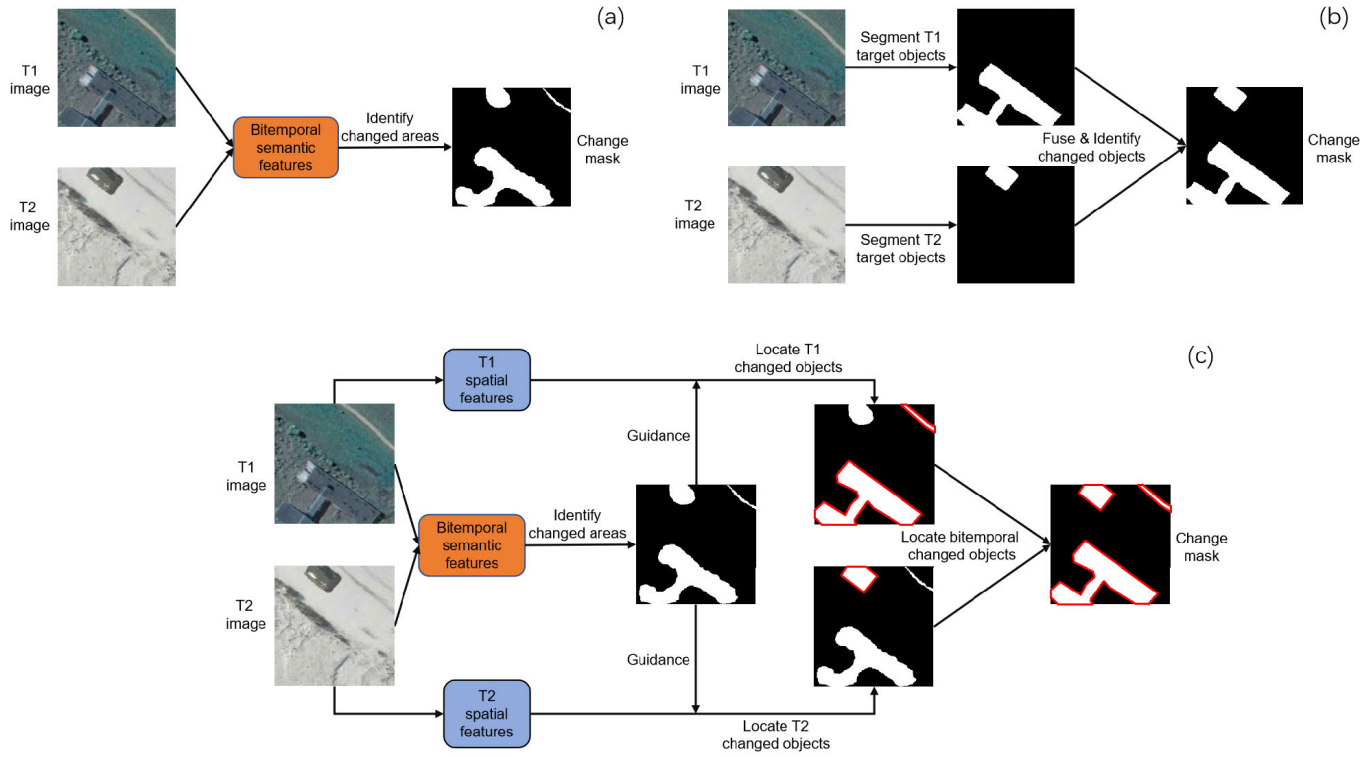


Fig. 1. Illustration of the main idea of MESD, DED, and EDED. (a) MESD: changed areas can be roughly identified by bitemporal semantic features. (b) DED: the specific types of changed objects can be identified by comparing the segmentation results of bitemporal target objects. (c) EDED: bitemporal changed objects can be identified by bitemporal semantic features and located by bitemporal spatial features, where the red edges denote changed objects in each temporal phase.

phase could be contaminated by the background features in another phase at the same spatial location, leading to inferior performance of the network [11]. As an example, Fig. 1(a) shows that the MESD can only identify the rough changed areas of changed objects.

DED consists of a DED with shared weights and a single decoder for change detection. The input bitemporal remote-sensing images are fed into the DED to segment the target objects in each image. The bitemporal features are then fused in the decision level in the single decoder to detect changed objects, as shown in Fig. 1(b). DED solves the problems of bitemporal feature contamination in MESD by segmenting the target objects in each image and fusing them in the decision level to detect changed objects.

However, DED has two assumptions: the target objects in the bitemporal images can be segmented accurately, and the changes can be retrieved correctly by comparing the target objects [11]. Therefore, DED faces great challenges in ICCD and MVBCD. In ICCD, there are multiple types of changes occurring in different classes of objects, while the binary labels only indicate the presence of changes without specifying the change types. Therefore, it is difficult for DED to segment the target objects without semantic labels of bitemporal images. As an example, Fig. 1(b) shows that DED detects the changed buildings while missing the changed roads. In MVBCD, due to the different imaging views, there are significant spatial discrepancies for the same object in bitemporal images. Since DED aims to segment the target objects in bitemporal images accurately, the spatial discrepancies of the same object will be mistaken as changed areas, leading to false positives.

Current neural networks also have other limitations for change detection: 1) most change detection networks focus on the important parts within features in each temporal when fusing bitemporal features, neglecting the important parts across the bitemporal features and 2) most change detection networks have a large number of parameters and require huge computational resources, resulting in time-consuming training and inference.

To address the aforementioned challenges, we propose a new strategy with exchanging DED (EDED) structure for binary change detection, as shown in Fig. 1(c). EDED has the same structure as DED except for a channel exchange module, which leads to a new strategy for change detection. In EDED, spatial features in each temporal phase are extracted in the shallow layers of the dual-encoder and half-exchanged, which makes features in each temporal branch both contain bitemporal features. Therefore, changed areas can be determined as guidance using bitemporal semantic features in the deep layers of the dual-encoder. Next, based on the changed areas, the T1 changed objects are located accurately using T1 spatial features. The changed objects in phase T2 are located in the same way. Finally, all changed objects can be located accurately when fusing bitemporal decoder features at the decision level. As an example, Fig. 1(c) shows that EDED can successfully detect the changed buildings and roads in the T1 image and the changed buildings in the T2 image.

EDED solves the problem of bitemporal feature concatenation in MESD by separately locating the changed objects in each image and fusing them at the decision level. Moreover, EDED can overcome the limitations in DED by determining the changed areas to identify all types of changed objects in the ICCD and distinguish the false change caused by view differences by using bitemporal semantic features in the MVBCD.

We also design a temporal fusion attention module (TFAM) and a half-convolution unit (HCU), in which the former focuses on the important parts across bitemporal features using temporal information, and the latter reduces the parameters and computation of conventional convolution to 1/4.

Based on these works, we propose a semantic guidance and spatial localization network (SGSLN) for binary change detection. The main contributions of this study are as follows.

1) We propose an EDED backbone as a new strategy for binary change detection. It uses bitemporal semantic features to determine changed areas and spatial features in each temporal phase to locate changed objects in the corresponding phase, thus locating all changed objects by fusing bitemporal changed object features. In this way, it addresses the issues of bitemporal feature contamination in MESD and overcomes the inapplicability in ICCD and MVBCD scenarios in DED.
2) We design a TFAM to fuse bitemporal features effectively. By comparing bitemporal features in the temporal dimension, the model can identify the important part across bitemporal features and achieve effective feature fusion.
3) We design an HCU to reduce the parameters and computation of the network. It has 1/4 the number of parameters and computation of conventional convolution and can facilitate feature reuse, allowing the model to achieve fast and robust training and inference.

II. Related Works

A. Classical Change Detection Methods

Classical change detection algorithms mainly include layer arithmetic, postclassification change, direct classification, transformation, change vector analysis (CVA), and hybrid change detection [14].

The layer arithmetic method compares image radiance or derivative features numerically to identify changes. For example, Coulter et al. [15] utilized regionally normalized difference vegetation index (NDVI) measures to detect changes in vegetative land cover. While this approach is straightforward to apply, it often offers limited insights into the detected changes.

Postclassification change method is the process of overlaying thematic maps from different time periods to pinpoint changes. One of the most established and extensively employed change detection techniques is directly comparing land cover maps derived from satellite data [16], [17]. This method offers a comprehensive thematic approach that can address specific queries about changes, making it applicable across various domains. Nevertheless, any error present in the input maps could be directly reflected in the resultant change map.

The direct classification method utilizes a multitemporal data stack as input, classifying it using supervised or unsupervised techniques to establish a set of consistent land cover classes and detect changes in land cover transitions. For example, Chehata et al. [18] implemented a forest change detection system by employing unsupervised classification on multitemporal imagery. This method only needs one classification stage and can provide an effective framework to mine a complicated time series. However, constructing training datasets for such a classification can be highly demanding, and unsupervised methods might not effectively capture subtle changes in magnitude [19].

Data transformations, such as principal component analysis (PCA) and multivariate alteration detection (MAD), can be employed on a multitemporal stack of remotely sensed images to emphasize variance between images and facilitate change identification. For example, Doxani et al. [20] found that implementing the MAD transformation on image objects effectively highlighted changed objects in very high-resolution (VHR) imagery. Similarly, Chen et al. [21] utilized the MAD transformation on image objects to accentuate change. These transformations offer a useful approach for assessing changes in complex time series of images. However, their primary function is often to highlight changes, thus they should ideally be integrated into a hybrid change detection workflow. Notably, due to scene-specific features, locating changes within multiple components may prove challenging, particularly if the change is not distinctly represented [14].

CVA is a technique for interpreting change by considering both its magnitude and direction. For example, Bruzzone and Prieto [22] computed the change magnitude across all six Landsat spectral bands to evaluate the apparent extent of change. Analyzing the magnitude and direction of change vectors can provide insights into the types of changes. However, this approach can also introduce ambiguity because the change vector itself can be repositioned within the feature space while preserving the same magnitude and direction measurements [23]. Consequently, there is a possibility that various thematic changes might yield identical measures of magnitude and direction.

The hybrid change detection method employs multiple comparison methods simultaneously to enhance the comprehension of detected changes. At a fundamental level, it can be conceptualized as a two-stage process: change localization and change identification. For example, Doxani et al. [20] tackled urban change detection in VHR imagery by first utilizing a MAD transform to highlight changed areas and then applying a knowledge-based classification to filter and classify the results. This methodology reflects a research trend that incorporates multiple stages of change comparison to address specific challenges [14].

B. Deep-Learning-Based Change Detection Method

Deep learning offers the capability to extract useful features and make accurate decisions by leveraging extensive sets of remote-sensing images, which allows deep-learning-based methods to outperform traditional methods in many remote-sensing applications. There are mainly three widely used

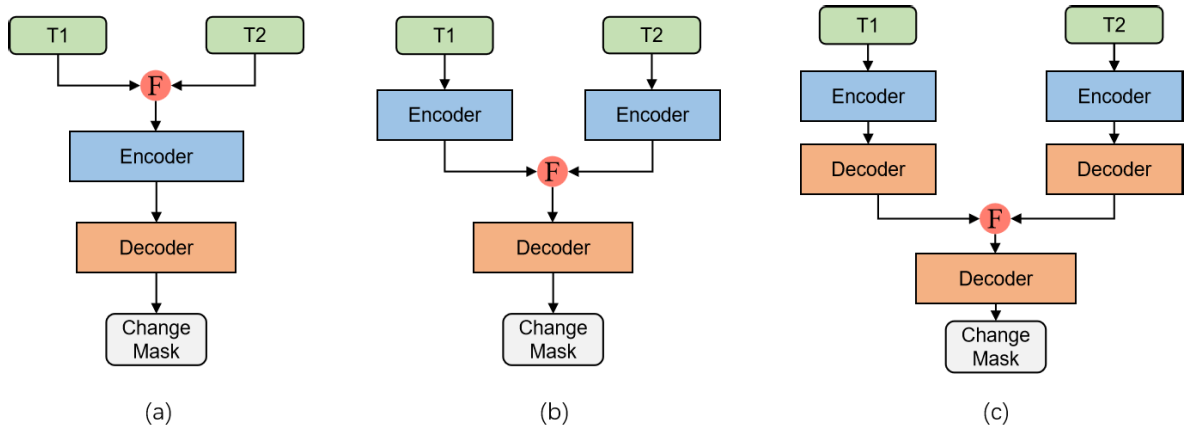


Fig. 2. Illustration of the architecture of SED, MESD, and DED. T1 and T2 denote bitemporal images, and F denotes a fusion module. (a) SED architecture. (b) MESD architecture. (c) DED architecture.

architectures in deep-learning-based change detection models: single encoder–decoder, MESD networks, and DED.

1) Single Encoder–Decoder: Single encoder–decoder (SED) uses a single encoder–decoder architecture to generate a change map from concatenated or difference images of bitemporal images, as shown in Fig. 2(a). Papadomanolaki et al. [24] proposed a deep-learning framework for urban change detection, which combines U-Net [25] for feature extraction and LSTMs [26] for temporal modeling. Peng et al. [27] used UNet++_MSOF [28] as the backbone, in which the spatial and channel attention strategies are used in the upsampling unit and the difference features are used to refine the results. Zheng et al. [29] proposed a convolutional neural network in which the U-Net [25] structure is used as the backbone and cross-layer blocks are embedded to incorporate multiscale features and multilevel context information. Most SED networks are modified from networks for single-image semantic segmentation. Since bitemporal images are fused as one input before being fed into the networks, early layers fail to provide informative deep features of individual raw images, which consequently results in change maps with broken object boundaries and poor object internal compactness [8].

2) Multiple Encoders and Single Decoder: MESD feeds bitemporal images and their difference image if necessary into multiple encoders to extract features, which are then merged and upsampled in a single decoder to generate the change map, as shown in Fig. 2(b). MESD can extract the information of images in each temporal by the Siamese network structure while maintaining a similar number of parameters and computation as SED by weight-sharing. Hou et al. [30] proposed a change detection method that uses triple encoders and multiscale modules to extract features, which are then fused with upsampling and distance computation to produce the change map. Zhu et al. [3] proposed a change detection network using a global hierarchical sampling mechanism to address the imbalanced training sample problem with insufficient samples. As bitemporal features may contaminate each other in the feature-level fusion, how to fuse them effectively becomes a challenge. Zhang et al. [8] concatenated and fused bitemporal features with channel and spatial attention strategies. Zhang et al. [31] fused bitemporal features with a convolution enhancement approach and self-attention in spatial and channel dimensions. Chen et al. [5] fused bitemporal features with a feature differential enhancement module, in which both local and global information is exploited and beneficial for bitemporal feature fusion. However, these fusion methods focus on the enhancement of the features themselves, ignoring the temporal information between bitemporal features. At the same time, bitemporal features interfere with each other at the positions of changed objects in the feature-level fusion, making it difficult to detect the changed objects accurately.

3) Dual-Encoder–Decoder: DED feeds bitemporal images into a DED to segment target objects in each image, which are then fused in a single change decoder to generate a change map, as shown in Fig. 2(c). DED networks are commonly used in multitask change detection and semantic change detection, where both segmentation labels and change labels are needed for training the model. However, DED networks are rarely applied in binary change detection since DED depends on the supervision of two segmentation branches. Chen et al. [11] demonstrated that the DED structure with only binary change labels supervised outperforms MESD in change detection and further improved the performance of DED by utilizing a self-supervised learning (SSL) strategy. SSL enables supervision of a dual-segmentation branch by making bitemporal segmentation results in the pseudo-labels for each other with a specific loss function, in which the unchanged part should be similar, and the changed part should be different between bitemporal segmentation maps. Liang et al. [13] adopted the same network structure and SSL strategy as [11], with additional deep supervision modules to train the network better and relation-aware modules to enhance features.

However, DED depends on the accurate segmentation of bitemporal images and obtains change maps by comparing segmentation maps [11], making it challenging to adapt to ICCD and MVBCD. ICCD with multiple change types and binary labels makes it difficult for DED to segment target objects of bitemporal images. MVBCD with different imaging angles between bitemporal images makes DED mistakes in the

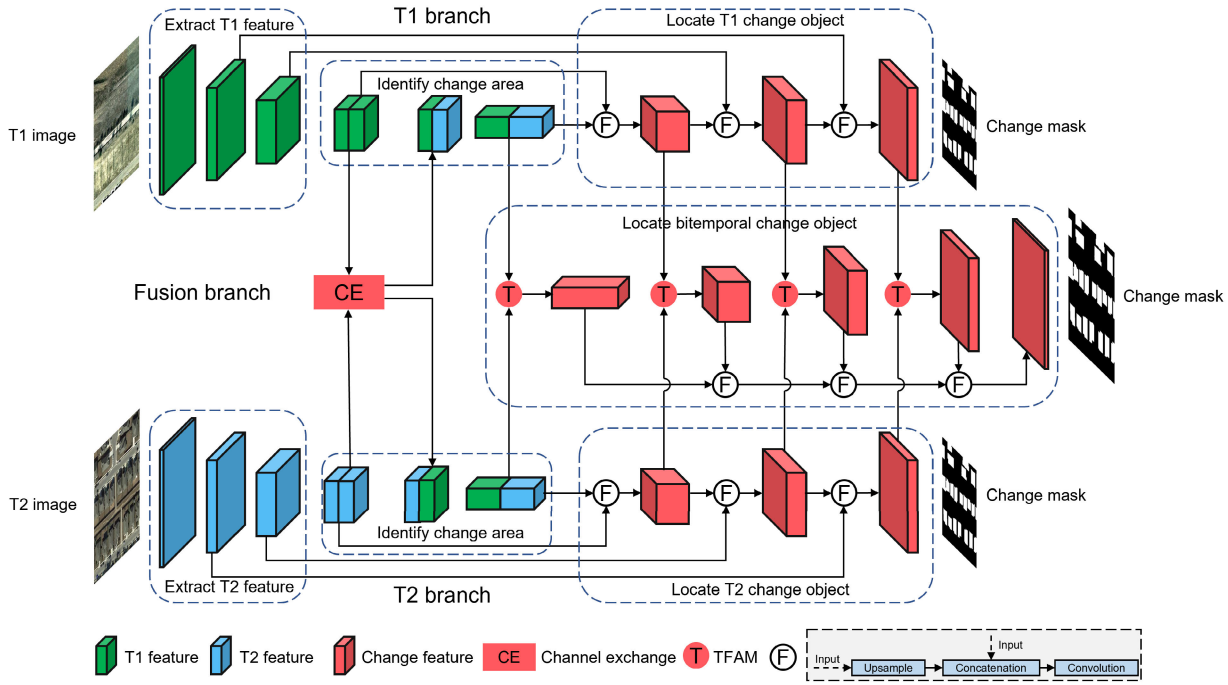


Fig. 3. Overall structure of the SGSLN. CE denotes the channel exchange module and TFAM denotes the temporal fusion attention module.

differences caused by different imaging angles of bitemporal target objects as changed areas.

III. Methodology

A. Overall Structure of the Proposed Network

SGSLN consists of an EDED backbone with two weight-sharing encoders, two weight-sharing decoders, and a fusion decoder, as shown in Fig. 3. Each encoder and decoder is composed of a series of HCUs and convolutional block attention modules (CBAMs [32]) for effective feature extraction.

The dual-weight-sharing encoder extracts spatial features of bitemporal images in the shallow layers. Bitemporal encoder features are then half-exchanged, which means features in each encoder both contain bitemporal features. After that, bitemporal encoder features are passed to deep layers of the dual-encoder so that the changed areas can be roughly identified in each encoder by exploiting the bitemporal semantic features, which guide subsequent changed object localization.

Based on the changed areas, the decoder in the T1 branch can precisely locate T1 changed objects by using the spatial features in T1 encoder features when fusing T1 decoder features and T1 encoder features through skip connections. The decoder in the T2 branch can locate T2 changed objects precisely in the same way. The dual-decoders in the bitemporal branches both generate a change mask with half the size of the input images, which are supervised with change labels to reduce the path length of gradient backpropagation and train the model effectively. As bitemporal changed objects are located in bitemporal decoder features, TFAMs in the fusion branch are designed to determine the relatively important parts between bitemporal features, thus effectively fusing the bitemporal changed object features and locating all changed objects. The decoder in the fusion branch generates a change mask with the same size as the input images, which is the result of the SGSLN.

B. EDED Backbone

EDED and DED have the same structure except for a channel exchange module, which completely changes the strategy of the model for change detection, as shown in Fig. 4. EDED and DED both have a DED and a single decoder. The dual-encoder–decoder in DED is used to segment the bitemporal changed objects separately, which means that the features in each encoder–decoder branch only contain single temporal information and ignore the connection between the bitemporal changed objects. In contrast, with the channel exchange module between the DED in EDED, each branch contains bitemporal information after channel exchange, which means that bitemporal features are connected and each branch can determine the changed areas itself. Based on the changed areas, the features with only single temporal information before channel exchange contain rich spatial features, which can be used to refine changed areas and accurately locate the changed objects in the same temporal phase. Finally, all the changed objects can be precisely located by fusing bitemporal features.

In EDED, bitemporal remote-sensing images are fed into dual-encoder blocks 1–3 to extract bitemporal features in each branch. The bitemporal features are then half-exchanged in the channel exchange module, which alternately exchanges half of the input bitemporal features in the channel dimension, as shown in the top right corner of Fig. 4. The process of the channel exchange module can be formulated as

$$T_1', T_2' = M * T_1 + (1 - M) * T_2 \tag{1}$$

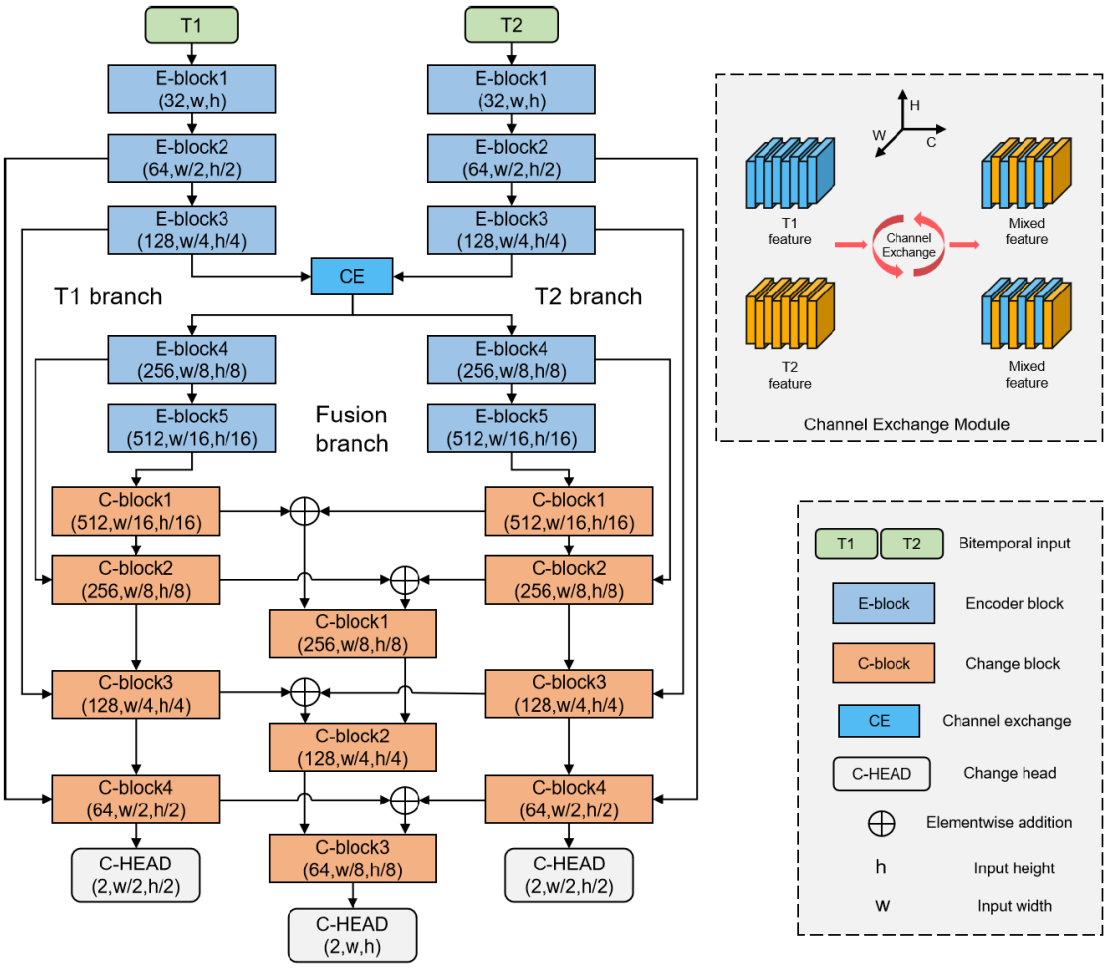


Fig. 4. EDED backbone for change detection. T1 and T2 denote the bitemporal remote-sensing image inputs and CE denotes the channel exchange module, which is shown in the top right corner.

where T_1 and T_2 denote bitemporal features, T_1' and T_2' denote exchanged bitemporal features, and M denotes the 1-D exchange mask, in which the length is equal to the channel dimension size of bitemporal features and the values are filled with 0 and 1 alternately. In this way, each exchanged feature contains half of the bitemporal features, which means that each exchanged feature contains bitemporal semantic features of bitemporal remote-sensing images. Therefore, dual-encoder blocks 4–5 in the bitemporal branch can determine rough changed areas using the bitemporal semantic features.

Since the changed areas only contain part of the changed objects, they can only determine the approximate location of the changed objects but cannot completely detect the changed objects. Therefore, we use the spatial features of each temporal image to refine the changed areas. Taking the changed areas as guidance, the decoder in the T1 branch fuses the encoder features with the decoder features through skip connections, and uses the spatial features of the T1 changed objects in the encoder features to refine the changed areas, thus completely detecting the T1 changed objects. The decoder in the T2 branch completely detects the T2 changed objects in the same way. The dual-decoders in bitemporal branches both generate change masks and are supervised by the change labels to train the dual branches better. Based on the bitemporal changed objects located in the bitemporal decoder, the decoder in the fusion branch can accurately locate all changed objects when fusing bitemporal decoder features and generate a change map as the result of the model.

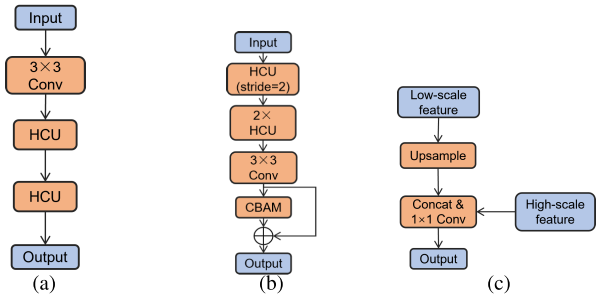


Fig. 5. Details of encoder blocks and change blocks. HCU denotes the half convolution unit, 3 × 3 Conv refers to the convolution layer with kernel size = 3, 1 × 1 Conv refers to the convolution layer with kernel size = 1, and concat refers to the concatenation of two features in the channel dimension. (a) Encoder block 1. (b) Encoder blocks 2–5. (c) Change block.

The blocks in EDED are carefully designed to obtain a strong feature extraction ability while keeping lightweight. Encoder block 1 extracts bitemporal features of input bitemporal remote-sensing images without downsampling, making the bitemporal features contain rich original information, as shown in Fig. 5(a). Encoder blocks 2–5 first use an HCU with

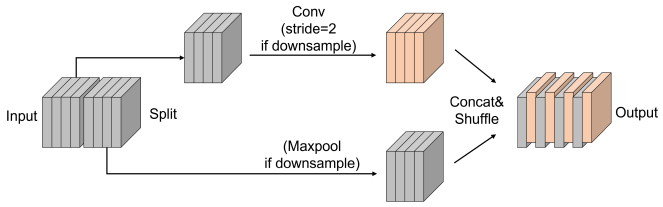


Fig. 6. Illustration of the HCU. Half-features passed to convolution layers are concatenated and shuffled with other residual half-features.

a stride equal to 2 to downsample the input features, then use 3 × 3 convolution after two HCUs to ensure sufficient cross-channel interaction of features, and finally use CBAM to enhance the features, as shown in Fig. 5(b). In this way, encoder blocks can achieve effective feature extraction with feature reuse of lightweight HCUs and feature enhancement of CBAMs so that encoder blocks 1–3 have rich spatial features to locate bitemporal changed objects, and encoder blocks 4–5 have rich semantic features to determine the changed areas. Change blocks upsample the input low-resolution features and fuse them with high-resolution features, thus identifying multiscale changed objects using multiscale features, as shown in Fig. 5(c).

EDED locates the changed objects separately and fuses them in the decision level, so that the changed object features in one temporal image will not be interfered with by the background feature in another temporal image at the same position when fusing bitemporal features, thus solving the bitemporal feature interference problems in MESD. Moreover, based on the semantic features of bitemporal images, EDED can determine changed areas of all categories of changed objects in ICCD and distinguish real changes and pseudo-changes caused by viewing angle differences in MVBCD, thus overcoming the inapplicability in ICCD and MVBCD scenarios in DED.

C. Half-Convolutional Unit

We propose an HCU to replace the conventional convolution, which reduces the parameters and computation of the model and achieves effective feature extraction, as shown in Fig. 6. The input features are split into two halves in the channel dimension, one of which is passed to convolution layers to be enhanced and the other serves as residual features. Features in two branches are then concatenated in the channel dimension and shuffled in an alternating arrangement, generating the output features. If the input features need to be downsampled, the stride of convolution is set to 2 and the residual features are downsampled using maxpool with a stride equal to 2.

The shuffle operation can ensure sufficient cross-channel interaction, and retaining half of the input features is beneficial to gradient back-propagation and feature reuse [33]. In this way, HCU has only 1/4 of the parameters and computation of conventional convolution while maintaining a strong feature extraction ability, thus making the model lightweight and achieving effective feature extraction.

D. Temporal Fusion Attention Module

Bitemporal feature fusion methods in change detection models can be classified into simple fusion, convolution enhancement, and attention enhancement [8], [11], [12], [31], [34]. The simple fusion method directly performs elementwise addition, subtraction, or concatenation on bitemporal features to fuse them [12], [34]. This method is susceptible to noise interference in bitemporal features, and it is difficult to achieve effective feature fusion. The convolution enhancement method enhances bitemporal features of multiple scales and semantic levels by applying various convolution operations, which reduces noise interference in bitemporal features. Then, it fuses bitemporal features using addition, subtraction, or concatenation [11]. The attention enhancement method usually concatenates bitemporal features in the channel dimension and then achieves effective fusion using attention mechanisms [8], [31]. However, the convolution enhancement method focuses on the enhancement of bitemporal features before fusion, and the attention enhancement method focuses on the enhancement of bitemporal features after simple fusion. They both ignore the temporal information between bitemporal features.

To solve the above issues, we propose a TFAM to utilize temporal information for effective feature fusion, which is shown in Fig. 7. It uses channel and spatial attention to determine the important parts of features and uses temporal information to determine the important parts between bitemporal features. In the channel branch, the input bitemporal features are passed through global pooling across the spatial dimension to aggregate spatial information. The aggregated process can be formulated as

$$S_c = \text{Concat}(\text{Avg}(T_1), \text{Max}(T_1), \text{Avg}(T_2), \text{Max}(T_2)) \quad (2)$$

where S_c denotes the aggregated spatial features, T_1 and T_2 denote bitemporal features, and $\text{Avg}(\cdot)$ and $\text{Max}(\cdot)$ denote global average pooling and global max-pooling across spatial dimension, respectively. The aggregated spatial features are passed to two 1-D convolutions, which are the same as ECA modules [35] to determine the bitemporal channel weights of the input bitemporal features. The two channel weights can be formulated as

$$W_{c1}, W_{c2} = \text{Conv}_1(S_c), \text{Conv}_2(S_c) \quad (3)$$

where W_{c1} and W_{c2} denote bitemporal channel weights, and $\text{Conv}_1(\cdot)$ and $\text{Conv}_2(\cdot)$ denote one-dimension convolutions. Softmax is then used in bitemporal channel weights to make their summation equal to 1, which means comparing bitemporal weights to determine the higher value between them, thus determining the important parts between bitemporal features in the channel dimension. The softmax approach can be formulated as

$$W'_{c1}, W'_{c2} = \frac{e^{W_{c1}}}{e^{W_{c1}} + e^{W_{c2}}}, \frac{e^{W_{c2}}}{e^{W_{c1}} + e^{W_{c2}}} \quad (4)$$

where W'_{c1} and W'_{c2} denote output bitemporal channel weights. The bitemporal spatial weights W'_{s1} and W'_{s2} are determined in the same way in the spatial branch, thus determining the important parts between bitemporal features in the spatial dimension. Bitemporal channel weights and bitemporal spatial weights are summarized to obtain bitemporal weights, which determine the important parts between bitemporal features.

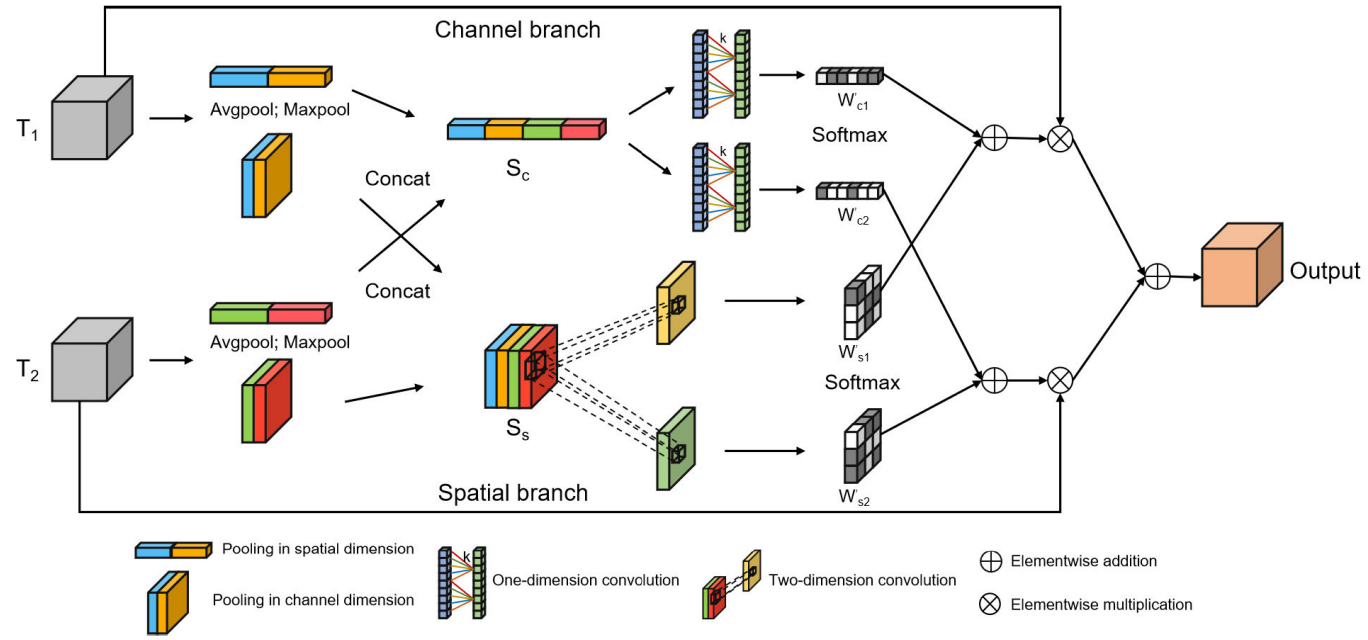


Fig. 7. Illustration of the structure of TFAM. T1 and T2 refer to bitemporal features. Avgpool and Maxpool denote average pooling and max-pooling in the channel and spatial dimensions, respectively. Concat denotes the concatenation of two features in the channel dimension.

Finally, bitemporal weights are multiplied with bitemporal features and summarized to effectively fuse bitemporal features. The output can be formulated as

$$\text{Output} = \left(W'_{c1} + W'_{s1}\right) * T_1 + \left(W'_{c2} + W'_{s2}\right) * T_2 \qquad (5)$$

where Output denotes the fused features. As the summation of bitemporal weights is equal to 1, the useful parts between bitemporal features are retained while the useless parts are discarded, thus achieving effective feature fusion.

IV. Experimental Settings and Results

We conducted experiments on three scenarios of six datasets to evaluate whether SGSLN is applicable for binary change detection: two for ICCD (CDD [36] and SYSU [37] datasets), three for SVBCD (WHU [38], LEVIR-CD [39], and LEVIR-CD+ datasets), and one for MVBCD (NJDS [40] dataset). We compared three versions of SGSLN with 18 state-of-the-art change detection methods. Three versions of SGSLN are SGSLN/128, SGSLN/256, and SGSLN/512. The number in the name indicates the maximum channel size of the encoder features in the model. The model with a larger channel number has a larger number of parameters and computations.

A. Datasets

We offer a brief description of the experimental binary change detection datasets in Table I.

1) Intraclass Change Detection: The CDD dataset [36] consists of 11 pairs of season-varying Google Earth images covering various objects (such as buildings, roads, and vehicles) that change. The dataset excludes the changes caused by seasonal differences and brightness, which makes it challenging for the change detection algorithm. The dataset is cropped into patches of 256 × 256 pixels, with 10 000 patches for training, 3000 patches for validation, and 3000 patches for testing.

The SYSU dataset [37] contains images that capture various types of complex change scenes, such as road expansion, new urban buildings, vegetation change, suburban growth, and groundwork before construction. We split the data into training, validation, and testing sets at a ratio of 6:2:2, following the same approach as [37].

2) Single-View Building Change Detection: The WHU dataset [38] consists of two-period aerial images acquired in 2012 and 2016, which contain various buildings with large-scale changes. Following the splitting approach used in [27], we crop the dataset into nonoverlapping patches of 256 × 256 pixels and randomly split them into training/validation/test sets with a ratio of 7:1:2.

The LEVIR-CD dataset [39] is a large-scale change detection dataset that contains VHR (0.5 m/pixel) Google Earth images. The images capture various types of buildings that have changed for 5–14 years. The dataset focuses on building-related changes, such as building growth and decline. The bitemporal images are labeled by experts using binary masks (1 for change and 0 for unchanged). The dataset has a total of 31 333 individual change-building instances. Following [11], we crop the images into patches of 256 × 256 pixels with an overlap of 128 pixels on each side (horizontal and vertical) and split the samples into training/validation/test sets with a ratio of 7:1:2.

The LEVIR-CD+ dataset is an extension of the LEVIR-CD dataset. It contains 985 pairs of images acquired from 2002 to 2020, with approximately 80 000 building instances. We follow the same splitting and cropping approach as the LEVIR-CD dataset on the LEVIR-CD+ dataset.

3) Multiview Building Change Detection: The NJDS dataset [40] addresses the building height displacement issue in change detection. It contains bitemporal images of Nanjing City in 2014 and 2018, obtained from Google Earth. The images include different types of low-, middle-, and high-rise buildings. Following the same approach as [40], we crop

TABLE I
BRIEF INTRODUCTION OF THE EXPERIMENTAL DATASETS

Name	Resolution (m)	Image pairs	Image size (pixels)
CDD	0.03-1	16000	256×256
SYSU	0.5	20000	256 × 256
WHU	0.3	1	32207×15354
LEVIR-CD	0.5	637	1024×1024
LEVIR-CD+	0.5	985	1024×1024
NJDS	0.3	1	14231×11381

the images into nonoverlapping patches of 256 × 256 pixels and randomly split them into training (540 pairs), validation (152 pairs), and testing sets (1827 pairs).

B. Benchmark Methods

We compare the proposed method with change detection networks based on SED, MESD, and DED architectures to verify its effectiveness. The benchmark methods tested on the same dataset are based on the same splitting of the dataset and use the same data, except that SFCCD [40] additionally uses segmentation labels of bitemporal images.

In SED architecture-based networks, bitemporal remote-sensing images are concatenated in the channel dimension and fed into a fully convolution-based network to obtain a change map. The compared change detection networks based on SED include U-Net [25], AttU-Net [41], FC-EF [12], UNet++_MSOF [28], and Intelligent-BCD [42].

In MESD architecture-based networks, bitemporal remote-sensing images and their addition or subtraction are fed into multiple encoders to extract features and fused in a single decoder to obtain a change map. The compared change detection networks based on MESD include FC-Siam-Diff [12], FC-Siam-Conc [12], DTCDSTN [43], IFN [8], SNUNet [34], STANet [39], TransUNetCD [44], and DARNet [45].

In DED architecture-based networks, bitemporal remote-sensing images are fed into a DED to segment bitemporal target objects. Bitemporal target object features are then fused in the change decoder to obtain a change map. The compared change detection networks based on DED include BiT [46], FCCDN [11], MTU-Net [47], and SFCCD [40].

C. Implementation Details

1) Data Augmentation: We apply various data augmentation techniques in the training stage to enhance the generalization ability of the models. These techniques include random flipping (probability = 0.5), transposing (probability = 0.5), random shifting (probability = 0.3), random scaling (probability = 0.3), random rotation (probability = 0.3), and one of the following transformations with probability = 0.3: HSV shifting, Gaussian noise, brightness and contrast adjustment, gamma noise, embossing, and motion blur. We use Albumentations [48] to implement all data augmentation methods with the default settings. Moreover, we randomly exchange the input order of the bitemporal images with probability = 0.5.

2) Training and Inference: We use PyTorch [49] to implement the SGSLN and train it on 1 RTX A5000 GPU (24 GB memory). The batch size is 64 for our network. We adopt the binary cross-entropy loss and dice coefficient loss as the loss function. AdamW [50] is used as the optimizer with an initial learning rate of 0.001 and a weight decay of 0.001. For the learning rate adjustment scheduler, we reduce the learning rate by 0.1 if the F1-score of the validation set does not increase within 12 epochs. We train the network for 250 epochs and save the checkpoints with the highest F1-scores on the validation sets for testing. The choice of 250 epochs was made to ensure that the model receives sufficient training and has reached convergence. The first 30 epochs are skipped in the validation as the model is far from converging. On the LEVIR-CD, CDD, SYSU, and NJDS datasets, we initialize the models following PyTorch's default settings to keep the same parameter initial method with other change detection methods. As the pretrained model can improve robustness and accelerate the model to converge [51], following the parameter initialized method in [11], we use the pretrained model trained on the LEVIR-CD dataset to initialize the SGSLN in the experiments on the WHU and LEIVR-CD+ datasets.

3) Evaluation Metrics: We use precision (P), recall (R), F1-score, and intersection over union (IoU) as the evaluation metrics for change detection. These metrics are widely used to measure the performance of change detection models. Precision measures the false positives in results while recall measures the false negatives. It is difficult to achieve high precision and recall simultaneously. The F1 score is the harmonic mean of precision and recall, which can balance the tradeoff by taking both metrics into account. The IoU is the ratio of the overlapping area between the predicted changed pixels and the changed pixels to the area of their union.

D. Ablation Study

To verify the effectiveness and superiority of EDED, we conduct comparative experiments among MESD, DED, and EDED on three change detection datasets of different scenarios (the SYSU dataset for ICCD, the LEVIR-CD dataset for SVBCD, and the NJDS dataset for MVBCD). In addition, to verify the effectiveness of HCU and TFAM, we conduct ablation experiments on HCU and TFAM on the LEVIR-CD dataset using EDED as the backbone.

EDED outperforms MESD and DED in all experimental change detection scenarios and achieves good performance, as shown in Table II. The results indicate that in ICCD and SVBCD, DED performs better than MESD, but in the MVBCD scenario, DED performs worse than MESD. In the above three change detection scenarios, EDED performs better than MESD and DED, especially in ICCD and MVBCD. The results of MESD, DED, and EDED further support the argument in Section III-B.

EDED outperforms DED in ICCD and MVBCD since DED cannot segment all types of changed objects in the former scenario and confuses real changes and false positives of spatial differences in the latter scenario, as shown in Fig. 8. The first row shows that EDED can correctly identify the

TABLE II

ABLATION STUDY OF EDED BACKBONE ON THE SYSU, LEVIR-CD, AND NJDS DATASETS. THE BEST VALUES ARE HIGHLIGHTED IN BOLD IN EACH DATASET

Backbone	Dataset	P (%)	R (%)	F1 (%)	IoU (%)
MESD	SYSU	85.12	74.56	79.49	65.96
DED	SYSU	**85.66**	75.84	80.45	67.30
EDED	SYSU	85.46	**78.24**	**81.69**	**69.05**
MESD	LEVIR-CD	93.01	89.44	91.19	83.80
DED	LEVIR-CD	92.88	90.67	91.76	84.78
EDED	LEVIR-CD	**93.09**	**91.32**	**92.20**	**85.52**
MESD	NJDS	78.04	63.57	70.07	53.92
DED	NJDS	76.77	63.72	69.64	53.42
EDED	NJDS	**80.21**	**67.94**	**73.57**	**58.19**

TABLE III

ABLATION STUDY OF THE HALF-CONVOLUTION UNIT AND TFAM ON LEVIR-CD. WE USE EDED AS THE BACKBONE. COMMON DENOTES THE BASIC CONVOLUTION UNIT

Convolution Unit	Fusion	P (%)	R (%)	F1 (%)	IoU (%)
Common	Add	92.12	90.27	91.19	83.80
HCU	Add	93.09	91.32	92.20	85.52
Common	TFAM	92.87	91.26	92.16	85.46
HCU	TFAM	93.07	91.61	92.33	85.76

spatial differences caused by different imaging angles as unchanged areas, while DED incorrectly identifies these spatial differences as changed areas, resulting in false positives. In the second row, EDED can detect most of the changed areas that contain multiple categories of objects accurately, while the result of DED has many false negatives.

Table III shows the ablation experiment results of HCU and TFAM on the LEVIR-CD dataset. Note that when comparing the model using conventional convolution and the model using HCU, the channel size of features in the latter model is twice that of the former model, making the number of parameters and computation of the two models consistent. The results show that HCU and TFAM are effective for binary change detection from different aspects. The first and second rows show that with consistent parameters and computation, using HCU can make the model more efficient and improve the performance; the first and third rows show that using TFAM can focus on the relatively important parts between bitemporal features and effectively fuse bitemporal features, thus improving the performance; the fourth row shows that using HCU and TFAM at the same time can further improve the performance. The above results indicate that HCU and TFAM can improve performance, and using them together can further improve the change detection ability. This ablation study shows that HCU and TFAM are better choices for feature extraction and feature fusion on change detection tasks, respectively.

E. Experimental Results

1) Intraclass Change Detection: This section presents the results of comparing SGSLN with other change detection models on the ICCD task with two datasets (CDD and SYSU datasets).

The accuracy comparison results on the CDD dataset are presented in Table IV. It shows that SGSLN/512 surpasses all the compared models and achieves the highest IoU (0.9563) and F1-score (0.9777) on the CDD dataset. Compared with the second-best method (TransUNetCD), SGSLN/512 increases the F1-score by 0.6%.

TABLE IV

ACCURACY COMPARISON ON THE CDD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [12]	60.90	58.30	59.57	42.42
FC-Siam-Diff [12]	76.20	57.30	65.41	48.60
FC-Siam-Conc [12]	70.90	60.30	65.17	48.34
UNet++_MSOF [28]	86.68	76.53	81.29	68.48
IFN [8]	90.56	70.18	79.08	65.40
BiT [46]	96.19	93.99	95.08	90.62
SNUNet [34]	96.30	96.20	96.25	92.77
TransUNetCD [44]	96.93	**97.42**	97.17	94.50
SGSLN/128	94.79	92.76	93.76	88.26
SGSLN/256	96.66	95.82	96.24	92.75
SGSLN/512	**98.25**	97.29	**97.77**	**95.63**

TABLE V

ACCURACY COMPARISON ON THE SYSU DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

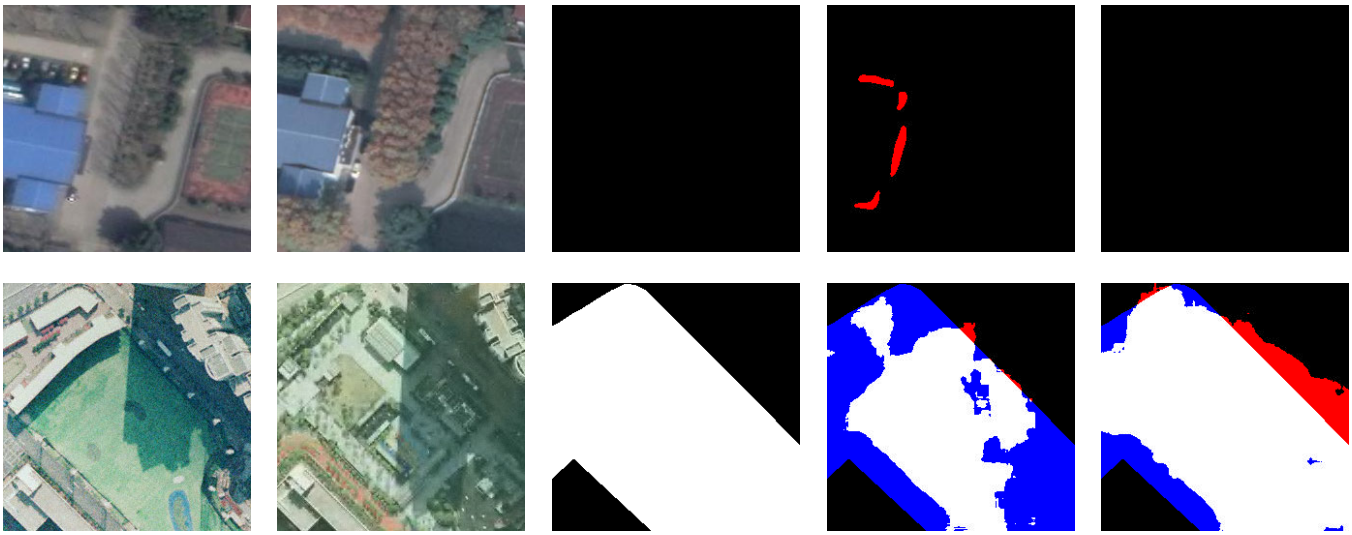
Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [12]	74.32	75.84	75.07	60.09
FC-Siam-Diff [12]	**89.13**	61.21	72.58	56.96
FC-Siam-Conc [12]	82.54	71.03	76.35	61.75
IFN [8]	79.59	75.58	77.53	63.31
STANet [39]	70.76	**85.33**	77.36	63.09
BiT [46]	81.14	76.48	78.74	64.94
SNUNet [34]	78.26	76.30	77.27	62.96
DARNet [45]	83.04	79.11	81.03	68.11
SGSLN/128	82.48	79.77	81.10	68.21
SGSLN/256	83.28	81.50	82.38	70.04
SGSLN/512	84.76	81.45	**83.07**	**71.05**

The accuracy comparison results on the SYSU dataset are summarized in Table V. Since the semantic information of the changed objects on the SYSU dataset is vague and contains objects of multiple categories, other change detection methods fail to accurately detect the changed objects. In contrast, SGSLN can detect the changed objects of all categories by using bitemporal semantic features to determine the changed areas of all changed objects. The accuracy comparison results show that SGSLN/512 achieves the highest IoU (0.7105) and F1-score (0.8307) on this dataset, surpassing all other models by a large margin. SGSLN/512, SGSLN/256, and SGSLN/128 all outperform other change detection models, increasing the F1-score by 2.04%, 1.35%, and 0.07% compared with the second-best model (DARNet), respectively.

The inference results of the test set of the CDD and SYSU datasets are shown in Fig. 9. The results show that SGSLN performs excellently on the two datasets. Under the strong interference of seasonal changes and illumination in CDD and the unclear semantic information of the changed objects in SYSU, SGSLN/512 still detects various types of changed objects with binary labels.

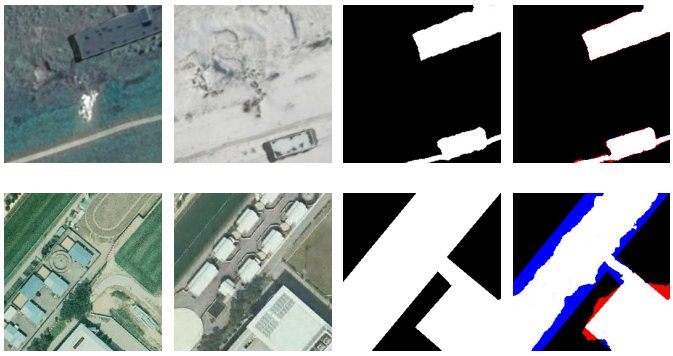
2) Single-View Building Change Detection: This section presents the results of comparing SGSLN with other change detection models on the SVBCD task with three datasets (WHU, LEVIR-CD, and LEVIR-CD+ datasets).

The accuracy comparison results on the WHU dataset are shown in Table VI. It shows that SGSLN/512 achieves the highest IoU (0.9022) and F1-score (0.9486) on this dataset,



(a) (b) (c) (d) (e)

Fig. 8. Change detection results of DED and EDED on the NJDS in the first row and SYSU in the second row. Red areas denote false positives and blue areas denote false negatives. (a) T1 image. (b) T2 image. (c) Ground-truth image. (d) DED result. (e) EDED result.



(a) (b) (c) (d)

Fig. 9. Sample inference results of SGSLN/512 on the ICCD. The results on the CDD and SYSU datasets are shown in the first and second rows, respectively. Red areas denote false positives and blue areas denote false negatives. (a) T1 image. (b) T2 image. (c) Ground-truth image. (d) SGSLN/512 result.

TABLE VI

ACCURACY COMPARISON ON THE WHU DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-Siam-Diff [12]	84.73	87.31	86.00	75.44
FC-Siam-Conc [12]	78.86	78.64	78.75	64.95
UNet++_MSOF [28]	91.96	89.40	90.66	82.92
DTCDSCN [43]	63.92	82.30	71.95	56.19
IFN [8]	91.44	89.75	90.59	82.79
STANet [39]	79.37	85.50	82.32	69.95
SNUNet [34]	85.60	81.49	83.49	71.67
BiT [46]	86.64	81.48	83.98	72.39
TransUNetCD [44]	93.59	89.60	91.55	84.42
FCCDN [11]	**96.39**	91.24	93.74	88.23
SGSLN/128	93.52	89.91	91.68	84.64
SGSLN/256	96.28	93.11	94.67	89.88
SGSLN/512	96.11	**93.64**	**94.86**	**90.22**

outperforming all other change detection models. At the same time, SGSLN/256 also outperforms all other models with a smaller size. Compared with the second-best method (FCCDN), SGSLN/512 and SGSLN/256 increase the F1-score by 1.12% and 0.93%, respectively.

The accuracy comparison results on the LEVIR-CD dataset are summarized in Table VII. It shows that SGSLN/512 outperforms all other models, achieving the highest IoU (0.8576) and F1-score (0.9233) on this dataset. Compared with the second-best method (FCCDN), SGSLN/512 increases the F1-score by 0.08%.

TABLE VII

ACCURACY COMPARISON ON THE LEVIR-CD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

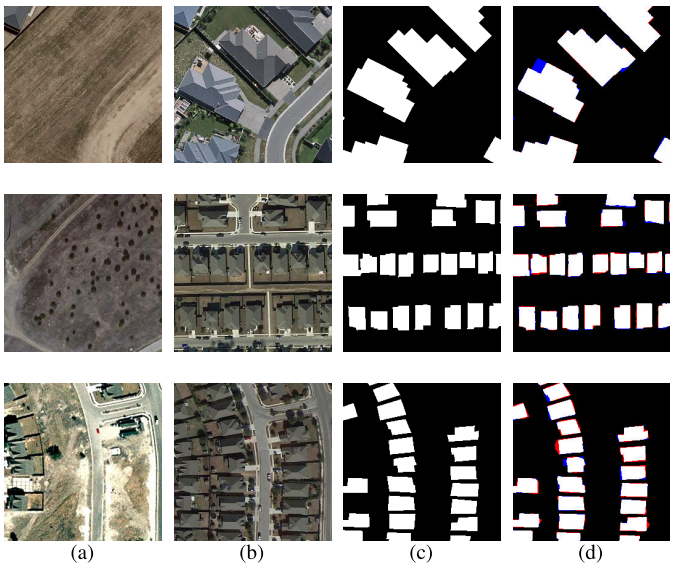
Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [12]	86.91	80.17	83.40	71.53
FC-Siam-Diff [12]	89.53	83.31	86.31	75.91
FC-Siam-Conc [12]	91.99	76.77	83.69	71.96
DTCDSCN [43]	88.53	86.83	87.67	78.05
IFN [8]	**94.02**	82.93	88.13	78.77
STANet [39]	83.81	91.00	87.26	77.39
BiT [46]	89.24	89.37	89.30	80.68
SNUNet [34]	89.18	87.17	88.16	78.83
TransUNetCD [44]	92.43	89.82	91.11	83.67
FCCDN [11]	92.96	91.55	92.25	85.61
SGSLN/128	91.79	90.21	91.00	83.48
SGSLN/256	92.71	91.17	91.93	85.07
SGSLN/512	93.07	**91.61**	**92.33**	**85.76**

Table VIII presents the accuracy comparison results on the LEVIR-CD+ dataset. It shows that SGSLN/512 achieves the highest IoU (0.8414) and F1-score (0.9139) on this dataset, surpassing other models by a large margin. SGSLN/512, SGSLN/256, and SGSLN/128 all outperform other change detection models, increasing the F1-score by 5.12%, 4.16%, and 3.66% compared with the second-best model (Intelligent-BCD), respectively. Since LEVIR-CD+ adds more hard samples on the basis of LEIVR-CD, the performance of the same model on LEVIR-CD+ has a significant decline compared with that on LEVIR-CD, such as BiT and STANet with declines of 6.51% and 7.99% in the F1-score metric, while SGSLN/512, SGSLN/256, and SGSLN/128 only have declines of 0.94%, 1.03%, and 1.05% in the F1-score metric. This shows that SGSLN can resist the interference of building shadows and dense building distribution better than other change detection methods and thus achieve superior performance in the more difficult SVBCD task.

We show some inference results of the test set of the WHU, LEVIR-CD, and LEVIR-CD+ datasets in Fig. 10.

TABLE VIII
ACCURACY COMPARISON ON THE LEVIR-CD+ DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	P (%)	R (%)	F1 (%)	IoU (%)
U-Net [25]	93.20	79.60	85.86	75.23
AttU-Net [41]	93.50	79.60	85.99	75.43
FC-EF [12]	61.30	72.61	66.48	49.79
FC-Siam-Diff [12]	74.97	72.04	73.48	58.07
FC-Siam-Conc [12]	66.24	81.22	72.97	57.44
UNet++_MSOF [28]	85.90	67.10	75.34	60.44
DTCDSCN [43]	80.36	75.03	77.60	63.40
STANet [39]	74.62	84.54	79.27	65.66
BiT [46]	82.74	82.85	82.79	70.64
Intelligent-BCD [42]	**93.80**	79.90	86.29	75.89
SGSLN/128	90.74	89.18	89.95	81.74
SGSLN/256	91.30	90.50	90.90	83.32
SGSLN/512	92.20	**90.59**	**91.39**	**84.14**



(a) (b) (c) (d)

Fig. 10. Sample inference results of SGSLN/512 on the SVBCD. The results on the WHU, LEVIR-CD, and LEVIR-CD+ datasets are shown in the first, second, and third rows, respectively. Red areas denote false positives and blue areas denote false negatives. (a) T1 image. (b) T2 image. (c) Ground-truth image. (d) SGSLN/512 result.

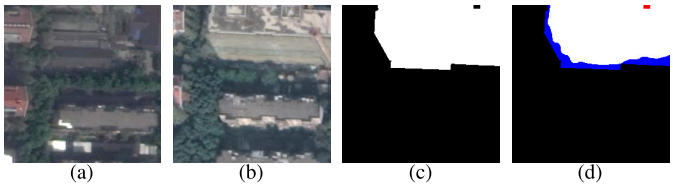
SGSLN/512 detects almost all the changed buildings in the three datasets. With the influence of building shadows and dense distribution of changed buildings, SGSLN/512 can still detect the regions and edges of changed buildings well. Note that there is a false change in the change mask of SGSLN/512 in the WHU dataset, as indicated by the blue part of the change mask in the first line. However, the building in the posttemporal image is under construction, which indicates building changes in this area. This is a frequent issue in this dataset as discussed in [52].

3) Multiview Building Change Detection: This section presents the results of comparing SGSLN with other change detection models on the MVBCD task with the NJDS dataset.

Table IX summarizes the accuracy comparison results on the NJDS dataset. It shows that SGSLN/512 achieves the highest IoU (0.5582) and F1-score (0.7165) on this dataset, surpassing all other models by a large margin. SGSLN/512 and SGSLN/256 both outperform other change detection models, increasing the F1-score by 4.82% and 3.76% compared with the second-best one (SFCCD), respectively.

TABLE IX
ACCURACY COMPARISON ON THE NJDS DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	P (%)	R (%)	F1 (%)	IoU (%)
U-Net [25]	46.45	52.64	49.35	32.76
AttU-Net [41]	55.57	44.60	49.48	32.88
PSPNet [53]	50.57	58.21	54.12	37.10
DTCDSCN [43]	51.92	62.78	56.84	39.70
IFN [8]	49.44	14.35	22.24	12.51
MTU-Net [47]	65.29	62.82	64.03	47.09
SFCCD [40]	74.49	65.19	69.53	53.29
SGSLN/128	71.01	59.40	64.69	47.81
SGSLN/256	79.44	68.03	73.29	57.85
SGSLN/512	**79.92**	**69.51**	**74.35**	**59.18**



(a) (b) (c) (d)

Fig. 11. Sample inference results of SGSLN/512 on the MVBCD. Red areas denote false positives and blue areas denote false negatives. (a) T1 image. (b) T2 image. (c) Ground-truth image. (d) SGSLN/512 result.

Fig. 11 illustrates the inference results on the test set of the NJDS dataset. This shows that under the strong interference of spatial differences caused by the multiviews of both low-rise and high-rise buildings, SGSLN/512 can still accurately detect the change in high-rise and low-rise buildings and identify the spatial differences as unchanged.

4) Efficiency Test: This section reports the results of comparing SGSLN with other models in terms of parameters, computation, and accuracy. We conduct an efficiency test on the WHU dataset using the same implementation details as described in Section IV-C2. Table X summarizes the results of the efficiency comparison. SGSLN/128 achieves an F1-score of 0.9168 with only 0.381 M parameters, 0.8045 FLOP computation, and 96 s of training time, surpassing other change detection models except FCCDN on the F1-score. SGSLN/256 and SGSLN/512 surpass all the compared change detection models in F1-score with relatively low parameters, computation costs, inference time, and training time. This demonstrates that SGSLN can achieve superior performance with relatively low computation complexity and high efficiency.

V. DISCUSSION

A. Exchanging Position

The channel exchange module between the dual-encoders lightens the EDED backbone. Without the channel exchange module, the model structure would resemble the DED architecture. It is not solely responsible for fusing bitemporal features as TFAM, its key contribution lies in the position of feature exchange. The channel exchange module makes the encoder features after the exchange have bitemporal semantic features of bitemporal images to determine changed areas, while the encoder features before the exchange retain the spatial features of bitemporal images for subsequent localization of changed objects. Therefore, the position of the channel exchange module is a key point for the EDED backbone. The position of exchanging bitemporal features should ensure that the encoder

TABLE X

EFFICIENCY COMPARISON ON THE WHU DATASET. THE BEST VALUES ARE MARKED WITH BOLD FONT. PARAMS, FLOPS, IT, TB, TT, AND F1 DENOTE THE NUMBER OF PARAMETERS, COMPUTATION COSTS, INFERENCE TIME WITH BATCH SIZE = 1, TRAINING BATCH SIZE WITH 12-GB MEMORY, TRAINING TIME IN 1 EPOCH, AND F1-SCORE, RESPECTIVELY

Name	Params (M)	FLOPs (G)	IT (s)	TB	TT (s)	F1 (%)
FC-EF [12]	1.35	3.56	**35**	90	140	78.75
FC-Siam-Diff [12]	1.54	5.30	**35**	65	140	86.00
FC-Siam-Conc [12]	1.35	4.70	**35**	60	140	83.47
UNet++_MSOF [28]	9.05	34.01	40	14	170	90.66
IFN [8]	35.7	82.3	75	14	255	90.59
BiT [46]	3.55	67.8	40	45	145	83.98
FCCDN [11]	6.25	12.4	60	28	150	93.74
SGSLN/128	**0.38**	**0.80**	50	**92**	**96**	91.68
SGSLN/256	1.51	2.98	58	48	128	94.67
SGSLN/512	6.04	11.5	65	25	165	**94.86**

TABLE XI

ACCURACY OF DIFFERENT EXCHANGING POSITIONS FOR SGSLN ON WHU. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Position	P (%)	R (%)	F1 (%)	IoU (%)
1	92.96	92.47	92.71	86.42
2	95.06	92.12	93.57	87.91
3	**96.11**	**93.64**	**94.86**	**90.22**
4	94.78	92.89	93.83	88.37
5	93.67	92.58	93.12	87.13

TABLE XII

ABLATION STUDY OF THE SUPERVISION BRANCH OF SGSLN/512 ON LEVIR-CD+. FUSION BRANCH MEANS ONLY FUSION BRANCH IS SUPERVISED, AND TRIPLE BRANCHES MEANS FUSION BRANCH AND BITEMPORAL BRANCHES ARE ALL SUPERVISED

Supervision Branch	P (%)	R (%)	F1 (%)	IoU (%)
Fusion branch	91.14	90.05	90.59	82.80
Triple branches	92.20	90.59	91.39	84.14

features before this position have rich spatial features, while the encoder features after this position have rich semantic features. We choose to exchange bitemporal features at the output position of encoding block 3, as in this position, the EDED backbone has a better performance.

Table XI shows the results of SGSLN/512 with different exchange positions, where exchanging bitemporal features at the position of encoding block 3 can make SGSLN/512 perform best. We argue that exchanging bitemporal features at other positions makes the model perform worse due to the following reasons: 1) if exchanging bitemporal features at the position of encoding block 1 or 2, there are too few spatial features for subsequent localization of changed objects, leading to inaccurate detection of changed objects and 2) if exchanging bitemporal features at the position of encoding block 4 or 5, the encoder features after the exchange have lost too much information due to downsampling many times, which makes it difficult for the model to detect all changed areas, and the encoder features are too small to detect small-scale changed areas. Therefore, exchanging features at the position of encoding block 3 can ensure that encoder features before the exchange have sufficient spatial features, while encoder features after the exchange have sufficient semantic features to detect all changed areas.

B. Bitemporal Branches

All triple branches in SGSLN generate change masks and are supervised by the same change label. What are the differences between the change masks produced by the bitemporal branches and the change mask generated by the fusion branch? How does the supervision of bitemporal branches affect the model's performance? We will discuss these two points in the following.

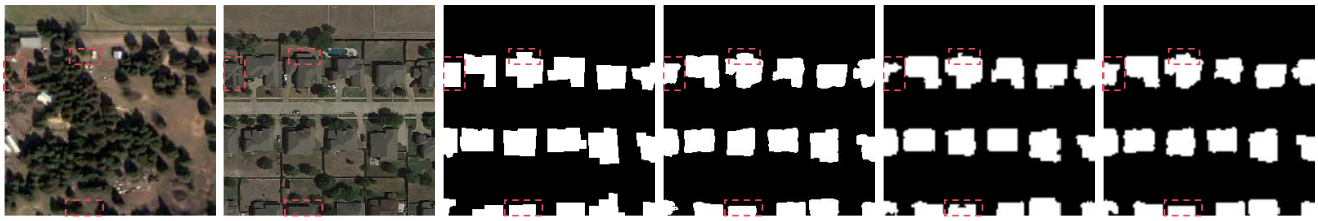
T1 branch utilizes the semantic features of bitemporal images to identify the changed area and spatial features of the T1 image to accurately localize T1 changed objects. Thus, the change mask generated by the T1 branch can detect T1 changed objects and the coarse area of T2 change objects. Similarly, the T2 branch can detect T2 changed objects and the coarse area of T1 changed objects. The fusion branch then fuses bitemporal features and precisely detects bitemporal changed objects.

Fig. 12 illustrates the inference results of SGSLN/512 on the LEVIR-CD+ dataset. Numerous new buildings have been constructed in the T2 image, which means all the changed objects are distributed in the T2 image. The enclosed area within the red box demonstrates disparities among the change masks of triple branches. The change mask of the T1 branch can only identify the rough area of changed buildings, while the T2 branch can precisely locate the changed buildings utilizing spatial features of the T2 image. The fusion branch result fuses features in bitemporal branches and precisely detects all changed buildings.

Supervising the bitemporal branches can enhance the training of the shallow layers. Since the gradient propagates from the final classification layer to the initial feature extraction layer in the deep network, the problem of gradient vanishing or exploding may occur in the propagation process, leading to the inefficiency of gradient propagation and poor features learned from lower layers [8], [54], [55], [56], [57]. Supervision of bitemporal branches can reduce the distance of gradient back propagation and enable sufficient training of the shallow layers. Table XII presents the ablation of the supervision branch on the LEVIR-CD+ dataset. A model with the supervision of triple branches improves 0.8% on the F1-score and 1.34% on IoU compared with a model with supervision only on the fusion branch, which demonstrates that supervision of bitemporal branches can enhance model training and improve the performance of the model.

C. Transferability

We conducted a comparative analysis of the transferability of SGSLN alongside other benchmark methods. All models



(a) (b) (c) (d) (e) (f)

Fig. 12. Sample inference results of SGSLN/512 with results of triple branches on the LEVIR-CD+ dataset. (a) T1 image. (b) T2 image. (c) Ground-truth image. (d) SGSLN/512 results in a fusion branch. (e) SGSLN/512 result in T1 branch. (f) SGSLN/512 result in T2 branch.

TABLE XIII

ACCURACY COMPARISON ON THE CROSS-DOMAIN CHANGE DETECTION FROM LEVIR-CD TO WHU

Methods	P (%)	R (%)	F1 (%)	IoU (%)
IFN [8]	**78.92**	21.73	34.08	20.54
SNUNet [34]	69.84	38.29	49.46	32.86
BiT [46]	64.38	33.83	44.35	28.50
TransUNetCD [44]	72.28	38.47	50.21	33.52
FCCDN [11]	70.91	41.24	52.15	35.27
SGSLN/512	73.46	**42.83**	**54.11**	**37.09**

were trained on the LEVIR-CD dataset, and then their accuracy was assessed on the WHU dataset. The resulting accuracy metrics are presented in Table XIII.

The model trained on LEIVR-CD exhibits a significant decrease in performance when applied to WHU, primarily due to disparities in imaging conditions and image scenes between the LEIVR-CD and WHU datasets. SGSLN/512 outperforms other models in terms of F1-score and IoU on WHU, indicating that SGSLN/512 has superior transferability compared to other methods. However, all pretrained models perform poorly on WHU. While SGSLN/512 achieves an F1-score of 0.9233 on LEIVR-CD, its F1-score on WHU drops to 0.5411. Training SGSLN/512 with WHU training data yields an F1-score of 0.9486 on the WHU test set, which means SGSLN/512 has considerable room for improvement in terms of transferability. In the future, we will focus on adjusting the model structure, enhancing training methods, and extending training data to bolster its transferability.

D. Expectations and Limitations

Binary change detection is a fundamental task in change detection that aims to identify the changes of interest by comparing the remote-sensing images of different time periods with binary labels. Various change detection methods have been proposed for this task, among which the MESD and DED architectures are widely adopted in network design. However, MESD suffers from the interference of bitemporal features in the fusion process, while DED is not suitable for ICCD and MVBCD scenarios because it fails to segment all types of changed objects in the former and confuses real changes with false positives of spatial differences in the latter.

Therefore, we propose an EDED backbone as a new strategy for binary change detection. EDED outperforms MESD and DED in the ICCD, SVBCD, and MVBCD scenarios, as shown in Table II. In ICCD, EDED can use bitemporal semantic features to determine changed areas, which are part of changed objects of all categories, and then use bitemporal spatial features to detect all types of changed objects. EDED increases the F1-score by 2.20% and 1.24% compared with MESD and DED on the SYSU dataset, respectively. In SVBCD, EDED can locate the changed buildings in each temporal using the bitemporal spatial features and accurately detect the region and edge parts of changed buildings. EDED increases the F1-score by 1.01% and 0.44% compared with MESD and DED on the LEVIR-CD dataset, respectively. In MVBCD, EDED can distinguish real changes from false changes caused by different imaging angles by using the bitemporal semantic features and accurately determining all the changed objects. EDED increases the F1-score by 3.50% and 3.93% compared with MESD and DED on the NJDS dataset, respectively. We expect the EDED backbone to be a new backbone for multiple binary change detection scenarios and achieve superior and robust performance in binary change detection.

Although EDED has superior performance compared with MESD and DED in binary change detection, it still has some drawbacks. EDED requires a large amount of multitemporal remote-sensing images and binary labels for supervised training, in which the annotation of changed labels has high labor and time costs, leading to the inadaptability of EDED in binary change detection with few or no labeled data. At the same time, EDED is oriented to binary change detection and cannot determine the category of changed objects in multitemporal remote-sensing images, limiting its use in semantic change detection.

VI. CONCLUSION

We propose an SGSLN model for binary change detection, which consists of an EDED backbone, HCUs, and TFAMs. Specifically, we propose an EDED backbone as a new strategy for binary change detection, which solves the bitemporal feature interference problem in MESD by locating changed objects in each temporal separately and overcomes the limitations in ICCD and MVBCD in DED by using bitemporal semantic features to detect all types of changed objects in the former and distinguish spatial differences pseudo-changes with real changes in the latter. We also propose a TFAM to fuse bitemporal features effectively by identifying the important parts between bitemporal features using temporal information and an HCU with 1/4 the number of parameters and computation of conventional convolution to achieve effective convolution and a lightweight model.

Experiments of SGSLN on the ICCD, SVBCD, and MVBCD scenarios show that SGSLN achieves superior performance with high efficiency and outperforms all compared

models. In ICCD scenarios, SGSLN can accurately determine various types of changed objects using bitemporal semantic features of bitemporal remote-sensing images. In SVBCD scenarios, SGSLN can accurately locate changed buildings by spatial localization of changed objects in each temporal and temporal fusion of bitemporal features. In MVBCD scenarios, SGSLN can use bitemporal semantic features to distinguish spatial differences caused by multiviews of objects with real changes. We expect SGSLN to serve as a baseline for binary change detection, exploring its potential in more diverse change detection scenarios and achieving accurate and robust performance in change detection.

ACKNOWLEDGMENT

The authors are grateful to the High-Performance Computing Center, Nanjing University, Nanjing, China, for their help with GPU resources. They would also like to thank the editor and the anonymous reviewers for their constructive comments.

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] F. Wang and Y. J. Xu, "Comparison of remote sensing change detection techniques for assessing hurricane damage to forests," *Environ. Monit. Assessment*, vol. 162, nos. 1–4, pp. 311–326, Mar. 2010.

[3] Q. Zhu et al., "Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.

[4] S. Jin, L. Yang, P. Danielson, C. Homer, J. Fry, and G. Xian, "A comprehensive change detection method for updating the national land cover database to circa 2011," *Remote Sens. Environ.*, vol. 132, pp. 159–175, May 2013.

[5] Z. Chen et al., "EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 203–222, Sep. 2022.

[6] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[7] H. Zhai, H. Zhang, P. Li, and L. Zhang, "Hyperspectral image clustering: Current achievements and future lines," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 35–67, Dec. 2021.

[8] 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.

[9] H. Cheng, H. Wu, J. Zheng, K. Qi, and W. Liu, "A hierarchical self-attention augmented Laplacian pyramid expanding network for change detection in high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 182, pp. 52–66, Dec. 2021.

[10] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, Jan. 2022.

[11] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 101–119, May 2022.

[12] 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.

[13] Y. Liang, C. Zhang, and M. Han, "RaSRNet: An end-to-end relation-aware semantic reasoning network for change detection in optical remote sensing images," *IEEE Trans. Instrum. Meas.*, early access, Feb. 22, 2023, doi: 10.1109/TIM.2023.3243680.

[14] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A critical synthesis of remotely sensed optical image change detection techniques," *Remote Sens. Environ.*, vol. 160, pp. 1–14, Apr. 2015.

[15] L. L. Coulter, A. S. Hope, D. A. Stow, C. D. Lippitt, and S. J. Lathrop, "Time–space radiometric normalization of TM/ETM+ images for land cover change detection," *Int. J. Remote Sens.*, vol. 32, no. 22, pp. 7539–7556, Nov. 2011.

[16] O. R. A. El-Kawy, J. K. Rød, H. A. Ismail, and A. S. Suliman, "Land use and land cover change detection in the western Nile delta of Egypt using remote sensing data," *Appl. Geogr.*, vol. 31, no. 2, pp. 483–494, Apr. 2011.

[17] L. Dingle Robertson and D. J. King, "Comparison of pixel- and object-based classification in land cover change mapping," *Int. J. Remote Sens.*, vol. 32, no. 6, pp. 1505–1529, Mar. 2011.

[18] N. Chehata, C. Orny, S. Boukir, and D. Guyon, "Object-based forest change detection using high resolution satellite images," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 49–54, Apr. 2013.

[19] T. A. Warner, A. Almutairi, and J. Y. Lee, "Remote sensing of land cover change," in *The SAGE Handbook of Remote Sensing*. London, U.K.: SAGE, 2009, pp. 459–472.

[20] G. Doxani, K. Karantzalos, and M. Tsakiri-Strati, "Monitoring urban changes based on scale-space filtering and object-oriented classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 15, pp. 38–48, Apr. 2012.

[21] G. Chen, G. J. Hay, L. M. T. Carvalho, and M. A. Wulder, "Object-based change detection," *Int. J. Remote Sens.*, vol. 33, no. 14, pp. 4434–4457, 2012.

[22] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[23] 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.

[24] 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.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[26] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[27] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

[28] 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.

[29] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 247–267, May 2021.

[30] X. Hou, Y. Bai, Y. Li, C. Shang, and Q. Shen, "High-resolution triplet network with dynamic multiscale feature for change detection on satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 103–115, Jul. 2021.

[31] L. Zhang, X. Hu, M. Zhang, Z. Shu, and H. Zhou, "Object-level change detection with a dual correlation attention-guided detector," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 147–160, Jul. 2021.

[32] 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.

[33] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[34] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[36] 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.

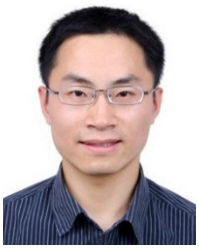
[37] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.
[38] S. Ji, S. Wei, and M. Lu, "Fully convolutional 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.
[39] 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.
[40] Q. Shen, J. Huang, M. Wang, S. Tao, R. Yang, and X. Zhang, "Semantic feature-constrained multitask Siamese network for building change detection in high-spatial-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 189, pp. 78–94, Jul. 2022.
[41] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[42] H. Zhang, G. Ma, and Y. Zhang, "Intelligent-BCD: A novel knowledge-transfer building change detection framework for high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5065–5075, 2022.
[43] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[44] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.
[45] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
[46] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[47] S. Tsutsui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Semantic segmentation and change detection by multi-task U-Net," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 619–623.
[48] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[49] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran, 2019, pp. 8024–8035.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[51] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2712–2721.
[52] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? Roll the dice and demand attention," *Remote Sens.*, vol. 13, no. 18, p. 3707, Sep. 2021.
[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[54] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[55] T. Mao, W. Liu, Y. Zhao, and J. Huang, "Change detection in semantic level for SAR images," in *Proc. IEEE 3rd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2018, pp. 633–636.
[56] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.
[57] J. Liu, M. Gong, A. K. Qin, and K. C. Tan, "Bipartite differential neural network for unsupervised image change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 876–890, Mar. 2020.

Sijie Zhao received the B.S. degree in geographic information science from the School of Geography and Ocean Science, Nanjing University, Nanjing, China, in 2023, where he is currently pursuing the M.S. degree in cartography and geographic information systems.

His research interests include change detection and earth surface forecasting for remote sensing.

Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information systems and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visiting Student with the Informatics Institute, University of Missouri, Columbia, MO, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University. He is currently an Associate Professor with the Department of Geographic Information Science, Nanjing University. His research interests include high-resolution remote-sensing image analysis, semantic segmentation, and deep learning for remote sensing.



Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Hesse, Germany, from 2011 to 2012; and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. Since 2019, he has been a Professor with Nanjing University. He has authored four books and over 160 articles. His research interests include high-resolution remote-sensing image analysis, remote sensing of snow cover, and land use and land cover change.

Guangjun He was born in Hubei, China, in 1987. He received the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2015.

From 2015 to 2017, he was an Assistant Researcher Fellow with the State Key Laboratory of Space-Ground Integrated Information Technology, the Chinese Academy of Space Technology (CAST), Beijing, China, where he has been an Associate Researcher Fellow, since 2018. He is the author of more than 40 articles. His research interests include remote-sensing image processing, multisensor information fusion, and machine learning.