DS-HyFA-Net: A Deeply Supervised Hybrid Feature Aggregation Network with Multi-Encoders for Change Detection in High-Resolution Imagery

Zilu Ying[1]*, Tingfeng Xian[1]*, *Student Member, IEEE*, Yikui Zhai[1]*, *Senior Member, IEEE*, Xudong Jia[2]*, Hongsheng Zhang[3], *Senior Member, IEEE*, Jiahao Pan[1], *Student Member, IEEE*, Pasquale Coscia[4], *Senior Member, IEEE*, Angelo Genovese[4], *Senior Member, IEEE*, Vincenzo Piuri[4], *Fellow, IEEE*, Fabio Scotti[4], *Senior Member, IEEE*.

***Abstract*—With the advancement of deep learning (DL) technologies, remarkable progress has been achieved in change detection (CD). Existing DL-based methods primarily focus on the discrepancy in bitemporal images, while overlooking the commonality in bitemporal images. However, one of the reasons hindering the improvement of CD performance is the inadequate utilization of image information. To address the above issue, we propose a Deeply Supervised Hybrid Feature Aggregation Network (DS-HyFA-Net). This network predicts changes by integrating the distinctness and the commonality in bitemporal images. Specifically, the DS-HyFA-Net primarily consists of a set of encoders and a Hybrid Feature Aggregation (HyFA) module. It uses a Siamese encoder (or Encoder I) and a specialized encoder (or Encoder II) to extract distinct and common features in bitemporal images, respectively. The HyFA module efficiently aggregates distinct and common features (or hybrid features) and generates a change map using a predictor. In addition, a common feature learning strategy (CFLS) is introduced, based on deeply supervised (DS) techniques, to guide Encoder II in learning common features. Experimental results on three well-recognized datasets demonstrate the effectiveness of the innovative DS-HyFA-Net, achieving F1-Scores of 93.33% on WHU-CD, 90.98% on LEVIR-CD, and 81.14% on SYSU-CD. Our code is available at https://github.com/yikuizhai/DS-HyFA-Net.**

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), Guangdong Science and Technology Planning Project (No. 2021A0505030080, No. 2021A0505060011), Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031, No. 2022ZDZX1032, No. 2023ZDZX1029, 2024ZDZX1009), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390). *(Corresponding author: Yikui Zhai and Hongsheng Zhang.)*

Zilu Ying, Tingfeng Xian, Yikui Zhai and Jiahao Pan are with the College of Electronics and Information Engineering, Wuyi University, Jiangmen, 529020, China (e-mail: ziluy@163.com, tingfengxian21@163.com, yikuizhai@163.com, 939946626@qq.com).

Xudong Jia is the College of Engineering and Computer Science, California State University, Northridge, 18111, America (e-mail: Xudong.Jia@csun.edu).

Hongsheng Zhang is the Assistant Professor with the Department of Geography, The University of Hong Kong, Hong Kong, China (e-mail: zhanghs@hku.hk).

Pasquale Coscia, Angelo Genovese, Vincenzo Piuri and Fabio Scotti are with the Department of Computer Science and Dipartimento di Informatica, Universita Degli Studi di Milano, 20133, Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it).

*Contributed to this work equally.

***Index Terms*—Change detection (CD), Multi-Encoder, hybrid feature aggregation module (HyFA), common feature learning strategy (CFLS), deeply supervised (DS).**

I. Introduction

CHANGE detection (CD), specifically from remote sensing images, refers to the process of recognizing alterations or changes in bitemporal images of the identical area [1]. CD has been utilized in numerous domains, including urban development assessment [2], environmental monitoring [3], land management [4], and disaster monitoring [5].

Advancements in remote sensing (RS) imaging technology have made possible to acquire insightful information from images efficiently and cost-effectively in the past few years [6] . These advancements have, in turn, propelled the field of CD. However, to date, CD still faces numerous challenges. One of the most prevalent issues is how to extract and utilize information from remote sensing images (RSIs) fully and efficiently.

Bitemporal RSIs introduce more complex and voluminous information [7], [8]. Additionally, the bitemporal RSIs used for CD are often captured under various light conditions, angles, and environmental factors. Consequently, objects with identical semantic information exhibit varying spectral characteristics across different time instances and spatial locations [9]. Therefore, how to extract information from bitemporal RSIs adequately and effectively is a topic worthy of research.

Traditional methods for CD primarily rely on algebraic or manual operations to identify changed regions. These methods include algebraic-based approaches [10], [11], [12], image transformation-based methods [13], [14], and post-classification techniques [15], [16]. No matter which methods they are, they have certain drawbacks. The algebraic-based methods suffer from the drawback that the segmentation threshold is a manual choice and is often difficult to determine. This often leads to false detections, that is, unchanged areas are often mistakenly identified as changed. The image transformation-based methods require a manual design of feature space. However, determining a universal and applicable feature space is challenging. The post-classification methods rely on their own classification algorithm. Their ability to generalize features across different scenarios is often

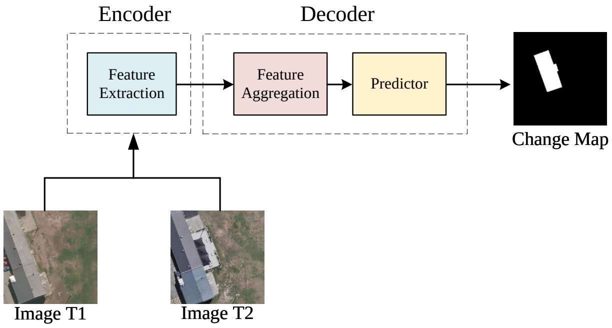


Fig. 1. A general framework of change detection. Image T1 and T2 were captured at the same location but at different times.

limited. In summary, traditional CD methods often rely on manual intervention, thus having high false detection rates and exhibiting poor generalization capabilities.

Nowadays, deep learning (DL) has gained renown for its powerful feature representation and generalization capabilities. Notably, convolutional neural networks (CNNs) were employed in widespread applications including CD studies. This is primarily because CNNs have strong modeling capabilities for image features. As an example, Zhan et al. [17], addressing the input of bitemporal RSIs in CD, proposed a deep Siamese convolutional network structure. Furthermore, Daudt et al. [18] introduced three fully convolutional U-Net-based network for CD, providing a valuable reference for exploring various structures of CD network. Wang et al. [19] introduced a cross-fusion network.

The emergence of attention mechanisms has offered more effective tools for feature fusion and interaction in CD. Chen et al. [20] introduced a pyramid spatiotemporal attention mechanism based on self-attention to integrate features embedded in bitemporal images. Shi et al. [21] incorporated convolutional based attention modules into the fusion of bitemporal image features. With the advent of Transformer [22] and their powerful global modeling capabilities, researchers have been inspired to explore their applications in CD. Chen et al. [9] introduced Transformer into CD. Zhang et al. [23] presented a cross-temporal difference (CTD) Transformer to model changes in bitemporal RSIs.

Despite the promising results achieved by existing DL-based CD methods, there remains ample room for exploration and further research. DL, as a branch of artificial intelligence, aims to design deep neural networks (DNNs) that can match or even surpass human performance. When observing how humans perform CD, we notice that they do more than just assess differences in bitemporal RSIs. They also consider the commonalities between RSIs. In this process, humans use the differences between bitemporal RSIs to initially detect changes and then focus on challenging areas, such as building edges and regions affected by lighting variations. Additionally, they utilize the commonalities between bitemporal RSIs to further confirm the change regions, leading to better CD. By utilizing the commonalities between bitemporal RSIs to confirm changed regions, a more efficient utilization of RSI information is achieved, thereby improving CD performance. Therefore, we argue that effectively utilizing both the differences and the commonalities between bitemporal RSIs to fully exploit the information in the images is a critical factor in achieving superior results in CD tasks. This concept provides a valuable perspective for further research in the field of DL-based CD.

By reviewing existing supervised learning CD methods [24], [25], [26], [27], [28], we have discovered that a framework for CD can be summarized by three modules, as depicted in Fig. 1. The feature extraction module extracts features from bitemporal RSIs. The feature aggregation module utilizes the extracted features to generate features that encapsulate information about the changed regions. The feature predictor then uses the fused features to generate a binary change map. It's worth mentioning that the feature extraction module is also considered as an encoder, while the feature aggregation module and the predictor combined are viewed as a decoder.

However, this framework primarily focuses on discrepancies in bitemporal images while neglecting their commonalities. This oversight leads to insufficient utilization of image information and consequently limits model performance. While exploring differential information in bitemporal images aligns with the objectives of CD tasks and has achieved some success, relying solely on differential information is insufficient for fully utilizing image information. To our knowledge, the exploration of common information in bitemporal images remains largely unexplored in the field of CD. Theoretically, relying solely on common information is also insufficient, a theory substantiated by subsequent ablation experiments. In contrast, integrating both differential and common information presents a novel approach for fully leveraging bitemporal images information, enhancing model performance.

Two limitations remain challenging in the field of RSI CD. The first limitation is the lack of attention to the commonality in bitemporal RSIs in the CD framework. We believe that CD models can be enhanced by understanding the discrepancy and commonality together in RSIs. We argue that incorporating the commonality in bitemporal RSI can enhance models' performance of CD. The second limitation is that existing models conducted CD by looking at the distinctness in bitemporal RSIs only. We believe that focusing on both the distinctness and commonality in bitemporal RSIs can effectively increase the CD performance. Therefore we suggest that approaching CD from aggregating the distinctness and commonality in bitemporal RSIs is a new research direction.

We have reexamined all existing CD frameworks and their related algorithms. We propose a new CD framework, that is, a DL-based network which simultaneously utilizes both the distinctness and commonality in bitemporal RSIs. Specifically, the DL-based network takes advantage of the existing CD frameworks and the established methods when it comes to detecting the distinctness in bitemporal RSIs. A Siamese encoder based on the pre-trained ResNet-18 [29] is used in our network to extract features from bitemporal RSIs. This process generates distinct features from each pair of RSIs. Regarding the commonality in bitemporal RSI, an efficient encoder was created to extract common features from bitemporal RSIs that are concatenated along the channel dimensions. As a result, the

DL-based network generates common features in bitemporal RSIs and hybridizes these common features with distinct features to form hybrid features. Finally, the hybrid features are passed through a hybrid feature aggregation (HyFA) module to produce aggregated features. The change map is generated by a predictor utilizing these aggregated features. The primary contributions of this work are outlined below:

1) An innovative CD method with multi-encoders is suggested for RSIs. In contrast to existing CD methods, the suggested approach addresses CD by aggregating both distinctive and common features in bitemporal RSIs. The combined utilization of distinct and common features allows for a more comprehensive exploitation of image information, enhancing model performance. Specifically, a Siamese encoder (or Encoder I) is employed to extract distinct features in bitemporal RSIs, while a specific encoder (or Encoder II) is developed for extracting common features. To the best of our knowledge, we are the first to aggregate both distinctive and common features in bitemporal RSIs.

2) A common feature learning strategy (CFLS) has been proposed. This strategy, based on the deeply supervised (DS) technique, uses auxiliary labels to mark non-change regions during the supervised training process. Specifically, DS utilizes intermediate layer features of the network for supervised learning(SL) to improve the network's feature learning capability. SL with auxiliary labels (representing non-change regions) obtained by taking the inverse values of normal labels enables Encoder II to capture common information in bitemporal images. This strategy significantly enhances the common feature extraction capability of Encoder II.

3) A Hybrid Feature Aggregation (HyFA) module is proposed to fully utilize hybrid features. After obtaining distinct features through Encoder I and common features through Encoder II, the HyFA module efficiently integrates these hybrid features, combining differential and common information from bitemporal images. The HyFA module progressively fuses multi-scale hybrid features, significantly enhancing model performance. This is the first method developed for dealing with the aggregation of hybrid features.

4) Our proposed method underwent quantitative and qualitative experiments on three well-recognized datasets. The quantitative experimental results(achieving F1-Scores of 93.33% on WHU-CD, 90.98% on LEVIR-CD, and 81.14% on SYSU-CD) and qualitative experimental results(Fig.9, Fig.10, Fig.11) illustrate that our suggested approach has superior performance.

The subsequent sections of this article are organized as follows. An overview of CD methods are provided in Section II. Section III describes the DS-HyFA-Net in detail. Section IV presents the comparison and ablation study results. Finally, Section V concludes this article while outlining directions for future research.

II. Related Work

A. *Traditional CD methods*

Traditional CD primarily relies on algebraic operations [30], image transformations [31], post-classification [15], and other methods. These approaches have made significant contributions to the advancement of remote sensing CD. Algebra-based methods require the selection of appropriate thresholds to determine changed and unchanged regions. These methods heavily rely on manually selecting a threshold to differentiate changed and unchanged pixels. Image transformation methods, such as independent component analysis (ICA) [32] and principal component analysis (PCA) [33], implement CD by creating feature space. However, finding a generic feature space is often a challenging task. Post-classification methods heavily depend on classification algorithms, thus exhibiting poor generalization performance. The emergence of machine learning (ML) algorithms injected new vitality into CD. Habib et al. [34] applied support vector machines (SVM) in CD and proposed an accelerated SVM algorithm for CD. Touati et al. [35], using Bayesian statistical methods, introduced a CD model based on a Random Markov Field. However, these ML-based methods face challenges in extracting and utilizing high-density information in RSIs due to the sophisticated settings of RS devices.

B. *DNN-Based CD methods*

The emergence of DNNs has ushered CD into a new stage of development. The powerful data representation and feature extraction capabilities of DNNs have significantly improved the performance of CD, opening new and promising research directions.

CNNs are commonly used in CD thanks to their powerful feature extraction capabilities. Given the nature of CD with bitemporal inputs, CNNs often utilize a weight-sharing Siamese architecture to extract distinct features from bitemporal RSIs individually. For example, three fully CNNs for CD were proposed by Daudt et al. [18]. Liu et al. [36] put forward a pyramid CNN for building CD. Wang et al. [37] put froward a Siamese spatial-spectral CNN. Han et al. [38] proposed a hierarchical attention network to address the issue of pixel class imbalance. Wang et al. [39] proposed a CD method that integrates superpixels with CNN. Tan et al. [40] applied a mixed interleaved group CNN for multi-sensor CD. Wang et al. [41] integrated spatial location information with graph convolution to detect urban changes.

Transformer equipped with self-attention mechanisms excel in a wide range of computer vision tasks. This success has inspired researchers to apply Transformer for CD. Li et al. [42] put forward a method to combine CNN with Transformer in parallel for CD. Liu et al. [43] employed Transformer for context aggregation in CD. A Siamese Transformer network multi-attention was put forward by Zhang et al. [44]. Tang et al. [45] combined CNN and Transformer to propose a W-shaped Net to resolve the problem of obscured long-range contexts.

The DL-based model significantly improved its performance by introducing attention mechanisms to mimic human behavior. However, the self-attention mechanism in Transformers often introduces a large set of parameters and computational complexity. Therefore, many researchers have turned to convolutional attention mechanisms, such as convolutional block

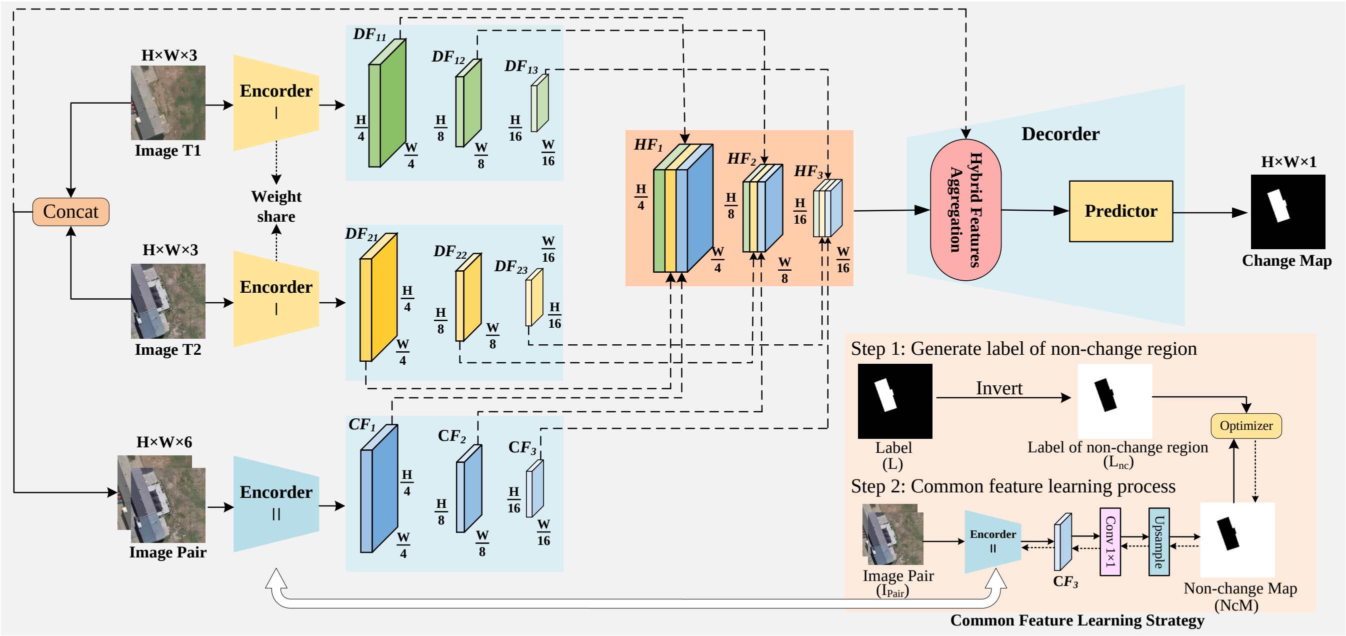


Fig. 2. Architecture of the DS-HyFA-Net. Common Feature Learning Strategy (CFLS) is based on Deep supervise (DS) technology and is used during the model training process. In CFLS, dashed arrow lines represent gradient backpropagation.

attention module [46] (CBAM), as a replacement for self-attention mechanisms and have applied them for CD. Liu et al. [47] developed a stacked attention module (SAM) based on CBAM to enhance effective information in multi-scale features. An ensemble channel attention module (ECAM) based on CBAM was put forward by Fang et al. [48] .

Furthermore, to enhance the feature discrimination capability of CD networks and improve network training, some researchers have employed DS techniques [49]. Zhang et al. [50] put forward a DS image fusion network (DSIFN). Ding et al. [51] combined DS and attention for CD. Lin et al. [52] put forward a CD network that combines video understanding techniques with DS techniques to improve the network's ability in distinguishing spatiotemporal features. Wang et al. [53] applied DS to retinal vessel segmentation, achieving improved segmentation results by dynamically analyzing and concurrently learning from both ”Easy” and ”Hard” regions. Zhang et al. [54] utilized DS for multi-sensor fusion.

Overall, existing methods primarily focus on detecting changes based on the distinctiveness in bitemporal RSIs, but they do not consider the significance of the commonality in bitemporal RSIs. Therefore, we propose the DS-HyFA-Net to extract both the distinct and common features in bitemporal RSIs. To enhance the ability to learn about common features, we introduce the DS technique. In addition, we also introduce CBAM to effectively combine distinct and common features. The overall performance of our network in CD tasks can be improved by these techniques.

III. Proposed Method

Firstly,the general framework of the DS-HyFA-Net will be presented in this section. We then proceed to introduce the multi-encoders, the CFLS, the HyFA module, and the hybrid loss functions used in the framework. Finally, the detection of changes in bitemporal images is obtained by training DS-HyFA-Net.

A. *Overall Architecture*

The architecture of DS-HyFA-Net is illustrated in Fig 2. Two sets of distinct features with different spatial resolutions, $[DF_{11}, DF_{12}, DF_{13}]$ and $[DF_{21}, DF_{22}, DF_{23}]$, are extracted from bitemporal images T1 and T2 using the weight-sharing Siamese encoder or Encoder I. Images T1 and T2 form an image pair $I_{pair} \in \mathbb{R}^{(H)\times(W)\times 6}$ by connecting them along the channel dimension. I_{pair} is then used for feature extraction by a specific encoder (or Encoder II), resulting in a set of common features $[CF_1, CF_2, CF_3]$. In addition, to enhance the common feature extraction capability of Encoder II, a CFLS was introduced. As shown in Fig 2, during model training, CF_3 is served as producing the non-change map and improve common feature learning using loss optimization and gradient backpropagation. Here, H and W represent the height and width, respectively, and the channels of images or features are denoted by C. After feature extraction, $[DF_{11}, DF_{21}, CF_1]$ are concatenated, denoted as $HF_1 \in \mathbb{R}^{(\frac{H}{4})\times(\frac{W}{4})\times 3C}$. Similarly, $HF_2 \in \mathbb{R}^{(\frac{H}{8})\times(\frac{W}{8})\times 3C}$ is obtained from $[DF_{12}, DF_{22}, CF_2]$, and $HF_3 \in \mathbb{R}^{(\frac{H}{16})\times(\frac{W}{16})\times 3C}$ is obtained from $[DF_{13}, DF_{23}, CF_3]$. Finally, $[HF_3, HF_2, HF_1, I_{pair}]$ are sequentially fed into the decoder, which includes a HyFA module and a predictor. The predictor comprises a 3x3 convolutional layer followed by a batch normalization (BN) layer [45]. This process generates the change map and achieves CD.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3471075

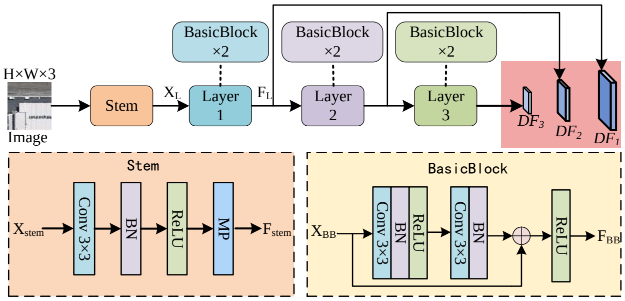


Fig. 3. Structure of Encoder I. Layers 1, 2, and 3 have the same structure, each consisting of two BasicBlocks, but their parameters differ.

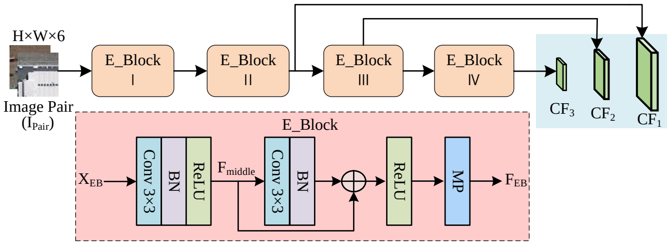


Fig. 4. Structure of Encoder II. E_Block I, II, III and IV have the same structure, but their parameters differ.

B. ***Multi-Encoders***

Existing DL methods primarily focus on discrepancies in bitemporal images, often overlooking their commonalities. One reason hindering the improvement of CD performance is the inadequate utilization of image information. We believe that access to more comprehensive information in RSIs is crucial for achieving excellent performance in CD tasks. Integrating both differential and common information presents a novel approach to fully leveraging bitemporal image information, thereby enhancing model performance. Specifically, a Siamese encoder (Encoder I) is employed to extract distinct features for obtaining differential information, while a specific encoder (Encoder II) is developed to extract common features for obtaining common information.

1) ***Encoder I****:* Inspired by [55], we utilize ResNet-18 [29] (a pre-trained model on ImageNet [56]) as an encoder I to extract the distinct features, enhancing the efficiency and simplicity of the network. Encoder I consists of Stem, Layer1, Layer2, and Layer3, as shown in Fig 3. For a given RSI $I \in \mathbb{R}^{(H)\times(W)\times 3}$, the distinct features $[DF_1, DF_2, DF_3]$ are computed as follows.

$$DF_1 = Layer1\,(Stem\,(I)) \tag{1}$$

$$DF_2 = Layer2\,(DF_1) \tag{2}$$

$$DF_3 = Layer3\,(DF_2) \tag{3}$$

Given the input X_{Stem}, the output feature F_{stem} of Stem in Encoder I is calculated as follows.

$$F_{stem} = MP(ReLU(BN(Conv_{3\times3}(X_{Stem})))) \tag{4}$$

where, $Conv_{3\times3}$ indicates a 3×3 convolutional layer, $ReLU$ represents a rectified linear unit (ReLU) activation function, MP represents a max-pooling layer.

Layer1, Layer2, and Layer3 have the same structure, each consisting of two BasicBlocks [29]. Specifically, take Layer 1 as an example given the input X_L, the output F_L of Layer 1 in Encoder I is computed as follows:

$$F_L = BasicBlock(BasicBlock(X_L)) \tag{5}$$

BasicBlock is mainly composed of $Conv_{3\times3}$, BN, and $ReLU$ activation function. Given the input X_{BB}, the output feature F_{BB} in Encoder I is computed as follows:

$$\begin{aligned} F_{BB} = &ReLU(BN(Conv_{3\times3}(\\ &ReLU(BN(Conv_{3\times3}(X_{BB})))))+X_{BB}) \end{aligned} \tag{6}$$

2) ***Encoder II****:* To extract the common features in bitemporal images efficiently and concisely, and to reduce the model parameters and complexity, we have designed an encoder called Encoder II for common feature extraction. As illustrated in Fig 4, Encoder II primarily comprises four Efficient Blocks (E_Blocks). Specifically, for Image Pair $I_{pair} \in \mathbb{R}^{(H)\times(W)\times 6}$ obtained by splicing bitemporal images $T1 \in \mathbb{R}^{(H)\times(W)\times 3}$and $T2 \in \mathbb{R}^{(H)\times(W)\times 3}$ along channel dimensions, the common features $[CF_1, CF_2, CF_3]$ are computed as follows:

$$CF_1 = E\_Block_{II}\,(E\_Block_{I}\,(I_{pair})) \tag{7}$$

$$CF_2 = E\_Block_{III}\,(CF_1) \tag{8}$$

$$CF_3 = E\_Block_{IV}\,(CF_2) \tag{9}$$

Inspired by [52], the design of the E_Block is like the residual module in ResNet [29]. In addition, the max-pooling helps extract features that contain information at various scales. Given an input image pair or feature x_{EB}, the output feature F_{EB} in Encoder II is computed as follows:

$$F_{middle} = ReLU(BN(Conv_{3\times3}(X_{EB}))) \tag{10}$$

$$\begin{aligned} F_{EB} = MP(ReLU(&F_{middle}+ \\ &Conv_{3\times3}\,(BN(F_{middle})))) \end{aligned} \tag{11}$$

C. ***Common Feature Learning Strategy***

Unlike distinct features that focus on the characteristics of RS images, common features need to focus on the shared attributes of RS images. DS technique is utilized to introduce an additional label to mark a non-change region. Significantly, the regular label for CD is a binary map in black and white, the additional label is a binary map with pixel values that are the opposite of the regular label.

Specifically, the proposed CFLS is divided into two steps (see Fig 2). In Step 1, the labels of non-change regions are generated. Given a set of data $\{I_{T1}, I_{T2}, L\}$, the pixel values for label L are inverted to a new set of data $\{I_{T1}, I_{T2}, L, L_{nc}\}$, where I_{T1} and I_{T2} are a pair of images, L denotes the regular label and L_{nc} denotes the label of the non-change region. In Step 2, CF_3 is utilized to generate a prediction of non-change region as shown in Eq 12:

$$NcM = Up\,(Conv_{1\times1}\,(CF_3)) \tag{12}$$

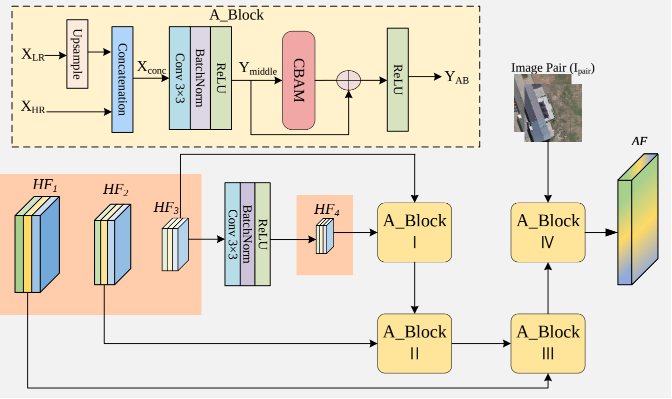


Fig. 5. Structure of the HyFA Module. A_Block I, II, III and IV have the same structure, but their parameters differ.

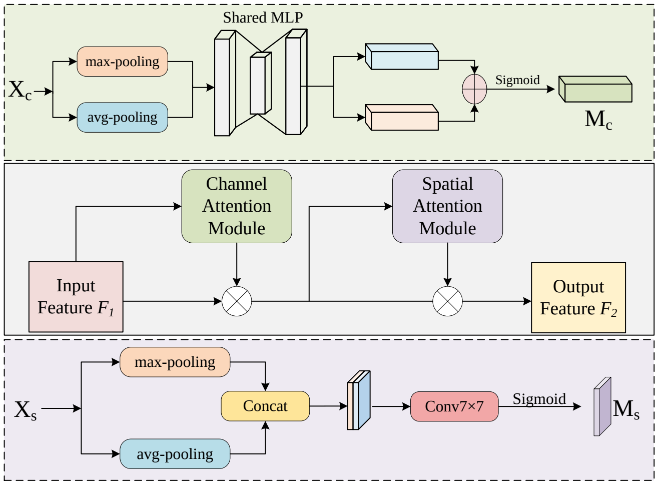


Fig. 6. Structure of CBAM.

where, $Conv_{1\times1}$ indicates a 1×1 convolutional layer. Up indicates the upsample operation based on nearest interpolation.

Ultimately, CFLS enhances the ability of Encoder II to extract common features by minimizing the gap between NcM and L_{nc} through a hybrid loss function and backpropagation method [58]. For comprehensive information about the hybrid loss function, please consult Section III-E .

D. Hybrid Feature Aggregation

How to efficiently utilize hybrid features is a topic worth studying. Shallow hybrid features contain more detailed information about changed regions, including the boundaries and spatial structure. Deeper hybrid features contain more abstract semantic information. Therefore, we first identify change regions by utilizing deep semantic information, and then use the detailed information from shallow features to enhance the boundary and structure of the change region, thereby achieving a better CD. Based on this, a HyFA module is developed which comprises four aggregation blocks (A_Blocks). As depicted in Fig 5, the aggregated feature $AF \in \mathbb{R}^{(H)\times(W)}$ is computed as follows for a set of hybrid features $[HF_1, HF_2, HF_3]$:

$$HF_4 = ReLU(BN(Conv_{3\times3}(HF_3))) \tag{13}$$

$$\begin{aligned} AF \ &= A\_Block_{IV}(I_{pair}, A\_Block_{III}(HF_1, \\ &\quad A\_Block_{II}(HF_2, A\_Block_{I}(HF_3, HF_4)))) \end{aligned} \tag{14}$$

As shown in Fig 5, A-BlockI, A-BlockII, A-BlockIII, and A-BlockIV have the same structure. For two given inputs, X_{LR} and X_{HR}, with different spatial resolutions, the A-Block computes the output Y_{AB} using the following equations:

$$X_{conc} = Concat\left(X_{HR}, Up\left(X_{LR}\right)\right) \tag{15}$$

$$Y_{middle} = ReLU\left(BN\left(Conv_{3\times3}\left(X_{conc}\right)\right)\right) \tag{16}$$

$$Y_{AB} \ = ReLU(CBAM(Y_{middle}) + \ Y_{middle}) \tag{17}$$

where, X_{LR} represents inputs with lower spatial resolution and X_{HR} represents inputs with higher spatial resolution, $Concat$ refers to concatenate.

Fig. 6 illustrates the CBAM. Specifically, given an input feature F_1, the output feature F_2 is computed according to the following equation:

$$F_2 \ = (F_1 \otimes CAM(F_1)) \otimes SAM(F_1 \otimes CAM(F_1)) \tag{18}$$

where, $\otimes$ denotes pixel-level multiplication, CAM denotes the channel attention module, and SAM denotes the spatial attention module.

Specifically, given an input X_c, the channel attention map M_c is computed according to the following equation:

$$M_c = \sigma(MLP(AP(X_c)) \ + \ MLP(MP(X_c))) \tag{19}$$

where, σ denotes the Sigmoid activation function. AP denotes an avg-pooling layer. MLP represents the multi-layer perceptron with one hidden layer.

For a given input X_s, M_s is computed as the following equation:

$$M_s = \sigma(\ Conv_{7\times7}(Concat(AP(X_s), MP(X_s)))) \tag{20}$$

where, $Conv_{7\times7}$ represents a 7×7 convolutional layer.

E. Hybrid Loss Function

During model training, the binary cross-entropy (BCE) is used as the fundamental loss function, denoted by Equation 21. Based on the DS technology, we introduce an auxiliary output to train the common feature extraction capability of Encoder II during model training. Consequently, the overall loss function is represented by Equation 22:

$$\begin{aligned} &\mathrm{L_{BCE}}\left(\mathrm{P},\mathrm{L}\right) = \\ &\frac{1}{H\times W}\sum_{i,j}\left[-Llog\left(P\right) - \left(1-L\right)log(1-P)\right] \end{aligned} \tag{21}$$

$$L_{HL} = L_{BCE}\left(P_c, L_c\right) + L_{BCE}\left(P_{nc}, L_{nc}\right) \tag{22}$$

where, P_c denotes the change map, P_{nc} refers to the non-change map, L_c signifies the label of change regions, and L_{nc} signifies label of non-change regions.

IV. Experiments

A. Datasets

1) WHU-CD [59]: A high-resolution building CD dataset, proposed by Wuhan University, was captured in Christchurch, New Zealand. Pairs of images captured in 2012 and 2016 formed the dataset. These images have the dimensions of 32,507 x 15,354 pixels and a spatial resolution of 0.2 meters per pixel. In this experiment, we followed the method described in the BIT [9] to crop the overlapping images into 256×256 patches. The training/validation/test set was randomly divided into 6096/762/762.

2) LEVIR-CD [20]: A large-scale, high-definition CD dataset, proposed by Beihang University, was captured at 20 different locations in Texas, USA. The dataset includes 637 pairs of images spanning 5-14 years, with the image dimensions of 1024×1024 and a spatial resolution of 0.5 meters per pixel. In this experiment, we followed the original dataset division method and cropped the images into non-overlapping 256x256 patches. The ratio of training, validation and test sets are 7120, 1024 and 2048 respectively.

3) SYSU-CD [21]: A high-resolution CD dataset, proposed by Sun Yat-sen University, includes changes in buildings, ships, roads, and vegetation. The dataset was captured in Hong Kong, China, and contains 20,000 pairs of 256x256 images with a spatial resolution of 0.5 meters per pixel. For our experimental tests, we followed the original dataset division method of SYSU-CD, which uses a training/validation/test set ratio of 12000/4000/4000.

B. Performance Metrics

We utilized five metrics to assess the performance of CD methods on the aforementioned datasets: Precision (Pre.), Recall (Rec.), F1-score (F1), Intersection over Union (IoU), and Overall Accuracy (OA). The five metrics are computed as follows:

$$Pre. = TP/\left(TP + FP\right) \tag{23}$$

$$Rec. = TP/\left(TP + FN\right) \tag{24}$$

$$F_1 = \frac{2 \times Pre. \times Rec.}{Pre. + Rec.} \tag{25}$$

$$IoU = TP/\left(TP + FP + FN\right) \tag{26}$$

$$OA = \left(TP + TN\right)/\left(TP + TN + FP + FN\right) \tag{27}$$

where, TPs - True Positives, TNs - True Negatives, FPs - False Positives, and FNs - False Negatives

C. Baseline Methods and Implementation Details

The existing state-of-the-art methods including FC-EF [18], FC-Siam-Conc [18], FC-Siam-Diff [18], STANet [20], BIT [9], DSIFN [50], SNUNet [48], P2V [52], and GeSANet [60] are used to be the baseline methods for comparing with DS-HyFA-Net. More details on the existing methods are provided below:

1) FC-EF: This is a U-shaped CD network that utilizes a single encoder and a single decoder structure.

2) FC-Siam-Conc: This is a U-shaped Siamese CD network that utilizes a dual-encoder single-decoder structure. It employs a feature fusion method for channel-wise concatenation and integrates the distinct features generated by the Siamese encoder before feeding the distinct features into the decoder.

3) FC-Siam-Diff: This is a U-shaped Siamese CD network that utilizes a dual-encoder single-decoder structure. This network calculates the pixel-wise differences between distinct features and passes the results to the decoder.

4) STANet: This is an Encoder-Decoder structured CD network that utilizes a Siamese ResNet-based network as the encoder. It introduces a pyramid spatial-temporal attention for the aggregation of distinct features and utilizes a contrastive loss-based measurement module as the decoder.

5) BIT: This is a CNN-Transformer CD network that utilizes a Siamese ResNet as the encoder for feature extraction. It incorporates a Transformer to retrieve interactions between distinct features.

6) DSIFN: Using VGG16 to extract distinct features and applying DS technique for CD, this approach introduces a deep supervision image fusion network.

7) SNUNet: Combining the Siamese network with Nested-UNet, this method introduces an ECAM based on CBAM.

8) P2V: Combining CD with the video understanding techniques, this approach models CD from both temporal and spatial dimensions. It proposes a decoupled CD network for time and space.

9) GeSANet: A geographic spatial perception network that utilizes a multi-level adjustment-based geographic spatial location matching mechanism (PMM) and a multi-factor pseudo-change information filtering-based geographic spatial content reasoning mechanism (CRM). It uses ResNet-18 to extract distinct features.

The models were executed in PyTorch. A single NVIDIA RTX 3060 GPU with 12GB of VRAM was used to support the training of the models. The batch sizes for all three datasets were set to 8, and the network parameters were updated using the Adam optimizer [61]. During training, the data were randomly augmented by flipping, shifting, and rotating by 90 degrees. The models were trained for 77,000 iterations with a learning rate (lr) of 0.0004 for WHU-CD, 72,000 iterations with a lr of 0.002 for LEVIR-CD, and 90,000 iterations with a lr of 0.002 for SYSU-CD, respectively. A step decay strategy was employed for regulating the learning rate. The decay rate and step (in terms of epochs) were set to 0.2 and 30 for WHU-CD and LEVIR-CD, and 0.2 and 10 for SYSU-CD. After each epoch, the best model was selected based on the highest F1 score during a validation step. Subsequently, the performance of the selected models on the test set was reported.

D. Performance Comparison with Baseline Methods

1) Experiments on WHU-CD: Table I presents the quantitative comparisons on WHU-CD. Our DS-HyFA-Net achieved the highest performance in F1, IoU, and OA, with values of 93.33%, 87.50%, and 99.43%, respectively. DSIFN achieved the highest precision, indicating that it had fewer FPs. However, this came at the cost of lower recall. DSIFN's recall was

TABLE I
QUANTITATIVE RESULTS OF ALL METHODS ON THE WHU-CD DATASET

Metric(%)	Pre	Rec	F1	IoU	OA
FC-EF	78.58	86.65	82.42	70.10	98.4
FC-Siam-Conc	59.14	85.46	69.90	53.74	96.81
FC-Siam-Diff	66.30	80.11	72.55	56.93	97.37
STANet	73.77	90.53	81.30	68.49	98.19
BIT	89.16	91.21	90.17	82.10	99.14
DSIFN	**95.43**	86.25	90.61	82.83	99.22
SNUNet	81.21	85.81	83.45	71.60	98.52
P2V	89.65	86.95	88.28	79.01	99.00
GeSANet	89.64	**93.04**	91.31	84.00	99.23
DS-HyFA-Net	94.98	91.74	**93.33**	**87.50**	**99.43**

TABLE II
QUANTITATIVE RESULTS OF ALL METHODS ON THE LEVIR-CD DATASET

Metric(%)	Pre	Rec	F1	IoU	OA
FC-EF	90.57	86.84	88.67	79.64	98.87
FC-Siam-Conc	90.94	87.88	89.38	80.81	98.94
FC-Siam-Diff	91.79	87.95	89.83	81.53	98.99
STANet	80.74	84.37	82.52	70.24	98.18
BIT	92.25	86.83	89.46	80.93	98.96
DSIFN	92.29	85.75	88.90	80.02	98.91
SNUNet	92.29	89.17	90.70	82.99	99.07
P2V	92.36	**89.54**	90.93	83.37	99.09
GeSANet	92.07	**89.54**	90.79	83.13	99.07
DS-HyFA-Net	**92.91**	89.12	**90.98**	**83.45**	**99.10**

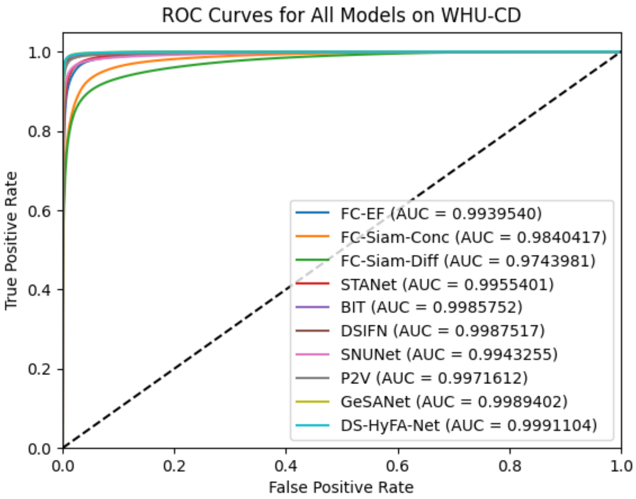


Fig. 7. ROC Curves for All Models on WHU-CD.

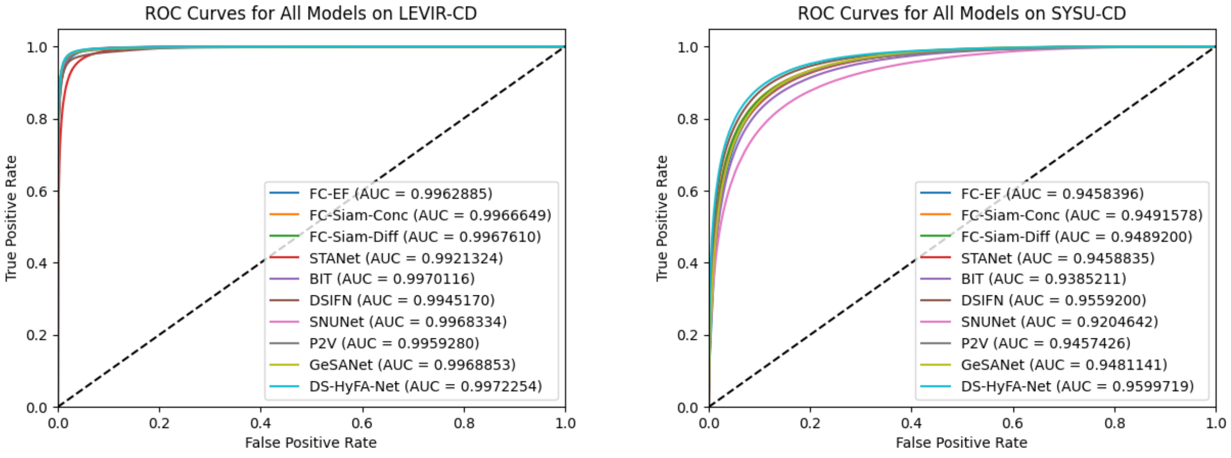
9.18% lower than its precision, resulting in a lower F1 score. The best recall was obtained by GeSANet, demonstrating its capability to reduce FNs. However, its precision was only 89.64%, with a 3.4% difference from the recall, resulting in a lower F1 score. To provide a more comprehensive evaluation of model performance, the ROC curves and AUC values for all models on WHU-CD are presented in Fig. 7. Our proposed model exhibits the highest AUC, approximately 0.99911, demonstrating the superiority of our approach.

The qualitative comparisons on WHU-CD were depicted in Fig. 9. Fig. 9 (a)-(c) demonstrate our method's superiority in detecting minor building changes compared to other CD methods. In Fig. 9 (a), FC-Siam-conc, FC-Siam-diff, and STANet exhibit higher FPs, while SNUNet and GeSANet have higher FNs. In Fig. 9 (c), FC-Siam-conc and P2V have higher FPs. Fig. 9 (d)-(e) demonstrate that our method has superior performance in detecting significant building changes. In Fig. 9, DSIFN and SNUNet exhibit higher FNs. In Fig. 9 (e), GeSANet has higher FPs, while FC-Siam-conc, SNUNet and FC-Siam-diff exhibit more severe FNs. BIT and DSIFN also have higher FPs. Fig. 9 (f)-(g) indicate that our method achieves superior performance in detecting changes in groups of buildings. As shown in Fig. 9, WHU-CD focuses on changes in buildings. However, in reality, there may be other types of changes between the bitemporal images, introducing more complex image information. Using only distinct or common features may not fully exploit the image information, but employing hybrid features can more effectively reveal the image information. Therefore, the proposed DS-HyFA-Net achieves better CD.

2) Experiments on LEVIR-CD: Table II presents the quantitative comparisons on LEVIR-CD. Our DS-HyFA-Net acquires the highest values in Pre (92.91%), F1 (90.98%), IoU (83.45%), and OA (99.10%). The best recall is achieved by P2V and GeSANet. This indicates that they are able to achieve fewer FNs. However, their precision is lower, indicating that more FPs are being generated. As a result, their combined performance metrics of F1, IoU, and OA are lower. Fig. 8(a) presents the ROC curves and AUC values for all models on LEVIR-CD. Our proposed model achieves the highest AUC, approximately 0.99723, which demonstrates the superiority of our approach

Fig. 10 demonstrates the qualitative comparisons on LEVIR-CD. From Fig. 10 (a)-(b), it is observed that our DS-HyFA-Net outperforms when dealing with isolated or sparsely distributed buildings. In comparison, STANet, BIT, DSIFN, SNUNet, P2V, and GeSANet are more prone to generating FPs when encountering objects that resemble buildings in the lower right corner. In Fig. 10 (b), FC-EF and P2V exhibit FNs due to lights and the absence of a building. On the other side of the coin, FC-Siam-conc, FC-Siam-diff, STANet, BIT, DSIFN, SNUNet, and GeSANet exhibit varying levels of FPs. Fig. 10 (c)-(e) demonstrate the performance of all methods for small building clusters. In Fig. 10 (c), when faced with interference, STANet and P2V exhibit higher frequency of FPs. In Fig. 10 (d), when dealing with small building edges, STANet, P2V, and GeSANet clearly exhibit FNs. In Fig. 10 (e), FC-EF and SNUNet have higher FNs, while DSIFN and GeSANet show higher FPs. For large and dense building clusters, as shown in Fig. 10 (f) and (g), STANet, SNUNet, and P2V have higher FPs in Fig. 10 (f). In Fig. 10 (g), FC-EF, FC-Siam-diff, STANet, BIT, DSIFN, SNUNet, and GeSANet exhibit different levels of FNs. As shown in Fig. 10, similar to WHU-CD, LEVIR-CD exists various changes of non-building types. In the presence of complex image information, the proposed DS-HyFA-Net, based on HyFA, achieves improved CD.

3) Experiments on SYSU-CD: Table III presents the quantitative results on the SYSU-CD. It is observed that our DS-HyFA-Net is best in terms of Pre (83.38%), F1 (81.14%), IoU (68.27%), and OA (91.34%). The best recall is achieved by



(a) (b)

Fig. 8. ROC Curves for All Models on (a)LEVIR-CD and (b)SYSU-CD.

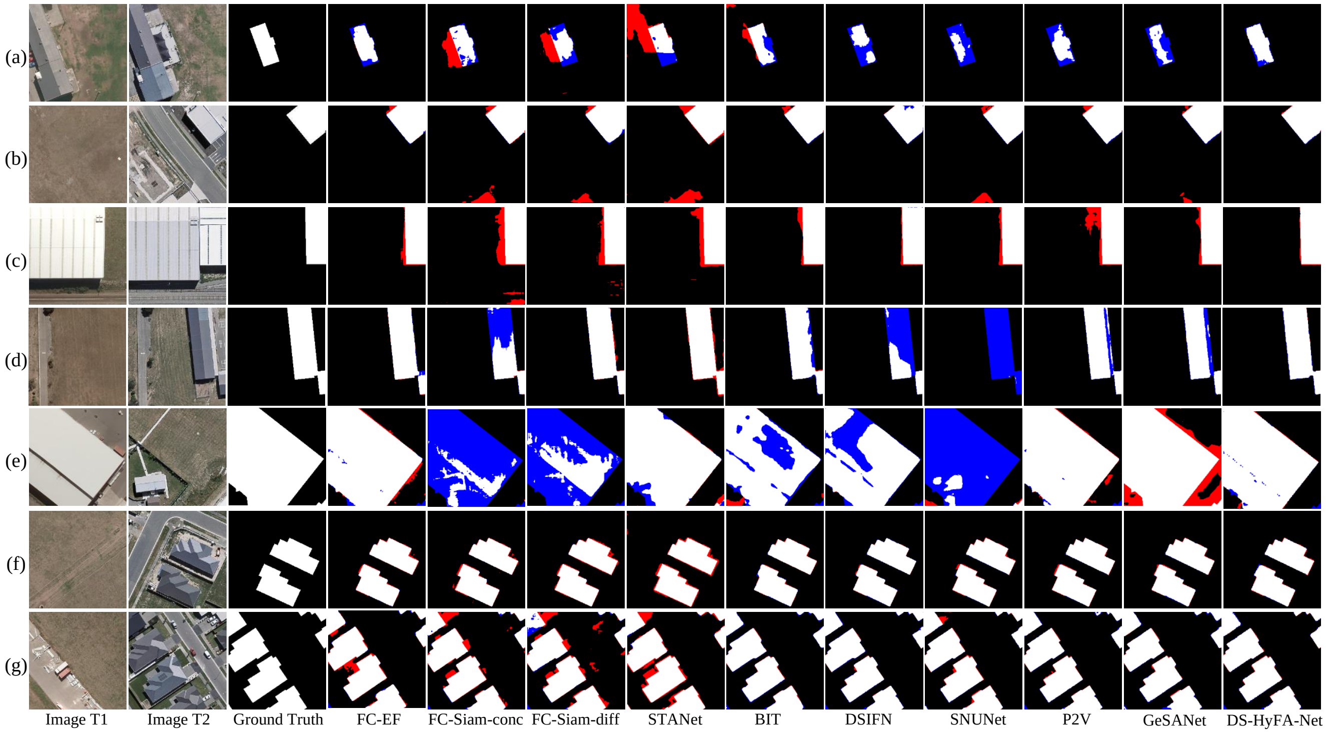


Fig. 9. Qualitative comparisons of all methods on the WHU-CD dataset. (a)–(g) Different image pairs along with their ground truth labels and the predictions of all methods. Images T1 and T2 form a pair of CD samples. Ground Truth represents the actual change regions. The rest are the predictions of different methods. TPs, TNs, FPs, and FNs are represented in white, black, red, and blue, respectively.

DSIFN, indicating that it is able to minimize the number of FNs. However, it has lower precision, resulting in a lower F1 score. The ROC curves and AUC values for all models on SYSU-CD are illustrated in Fig. 8(b), clearly highlighting the superiority of our model. Our proposed model achieves the highest AUC, approximately 0.95997.

The qualitative comparisons on SYSU-CD were presented visually in Fig. 11. Fig. 11 (a)-(c) demonstrate the detection performance of all models for changes in suburban expansion. In Fig. 11 (a), STANet, BIT, and DSIFN show a significant number of FPs, while SNUNet and GeSANet exhibit a substantial number of FNs. In Fig. 11 (b), BIT exhibits a more severe FP, and FC-Siam-conc, STANet, SNUNet, and GeSANet also show a significant number of FPs. In Fig. 11 (c), both FC-Siam-conc and SNUNet exhibit a significant number of FNs. Fig. 11 (d)-(e) showcase the detection performance of all methods for changes in urban development. In Fig. 11 (d), FC-Siam-conc, DSIFN, SNUNet, and GeSANet show significant FPs, while FC-siam-diff, STANet, and P2V also display different degrees of FPs. Additionally, FC-EF, BIT,

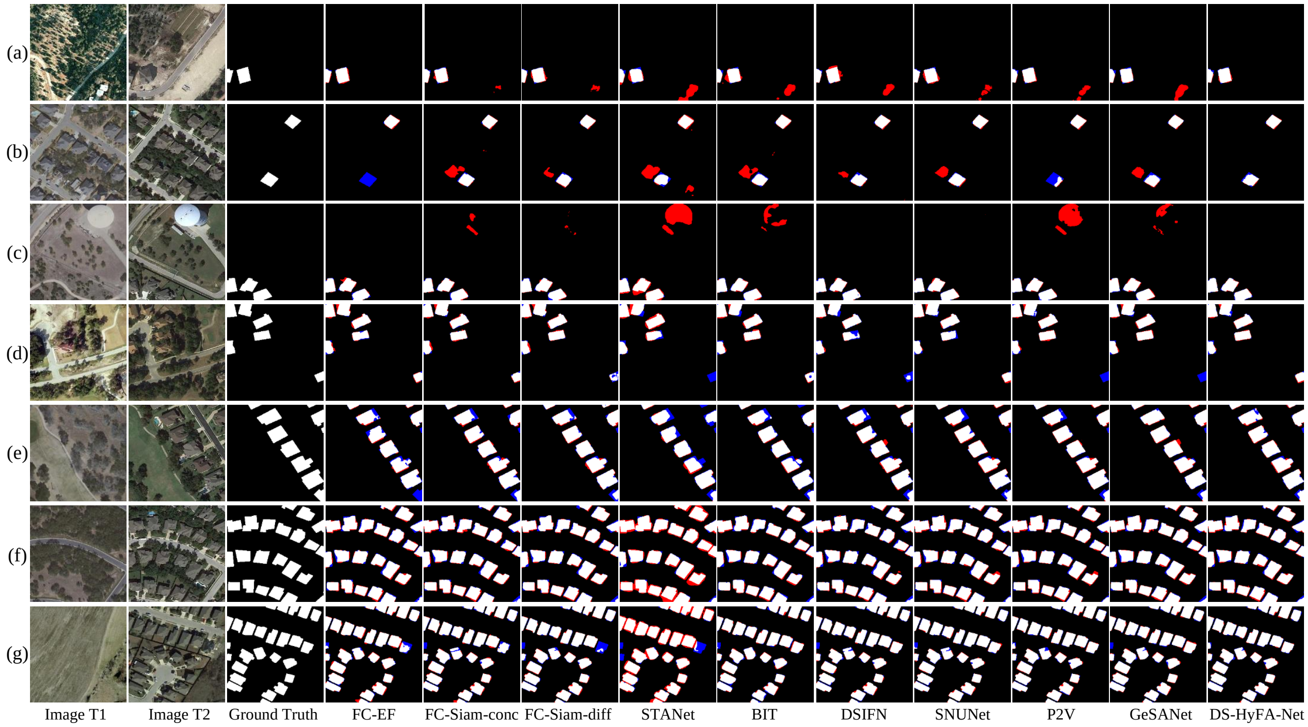


Fig. 10. Qualitative comparisons of all methods on the LEVIR-CD dataset. (a)–(g) Different image pairs along with their ground truth labels and the predictions of all the methods. Images T1 and T2 form a pair of CD samples. Ground Truth represents the actual change regions. The rest are the predictions of different methods. TPs, TNs, FPs, and FNs are represented in white, black, red, and blue, respectively.

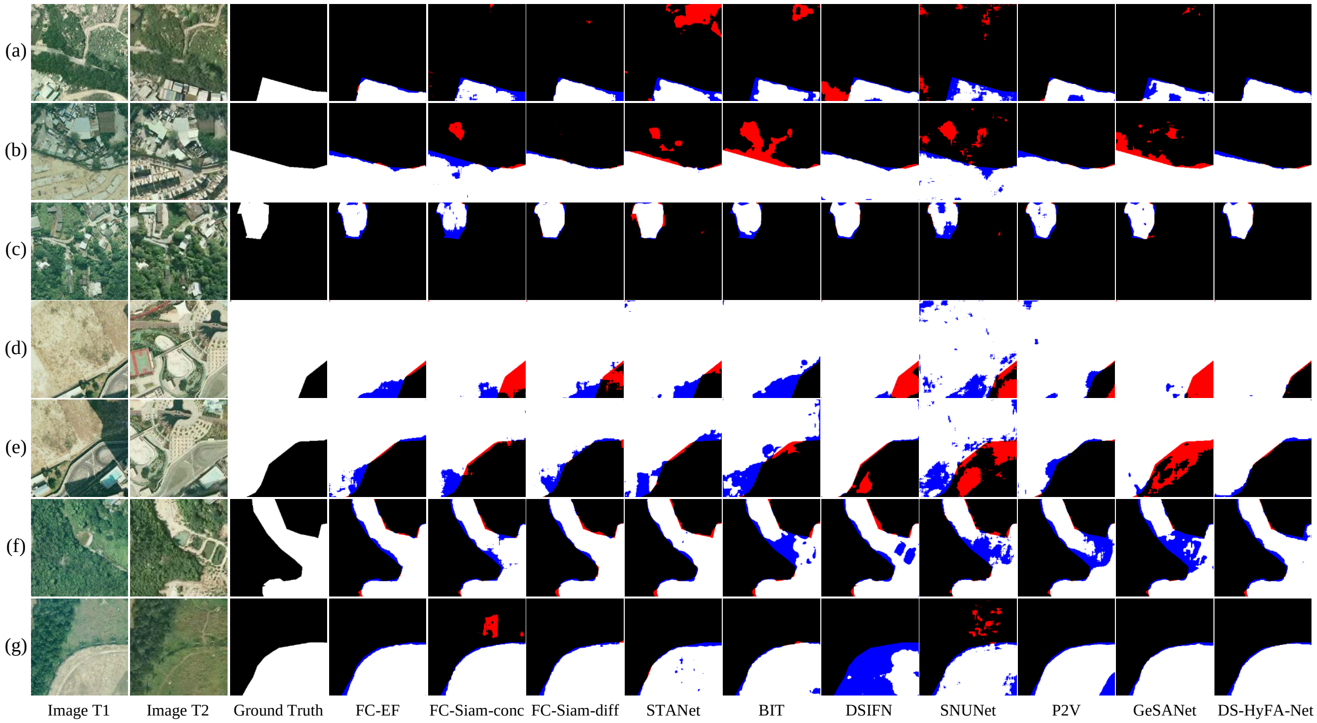


Fig. 11. Qualitative comparisons of all methods on the SYSU-CD dataset. (a)–(g) Different image pairs along with their ground truth labels and the predictions of all methods. Images T1 and T2 form a pair of CD samples. Ground Truth represents the actual change regions. The rest are the predictions of different methods. TPs, TNs, FPs, and FNs are represented in white, black, red, and blue, respectively.

and SNUNet have a significant number of FNs. In Fig. 11 (e), SNUNet and GeSANet have a significant number of FPs, while FC-Siam-conc and BIT exhibit a substantial number of FNs. Fig. 11 (f)-(g) demonstrate the performances in detecting vegetation changes. In Fig. 11 (f), BIT, SNUNet, P2V, and GeSANet exhibit a significant number of FNs. In Fig. 11

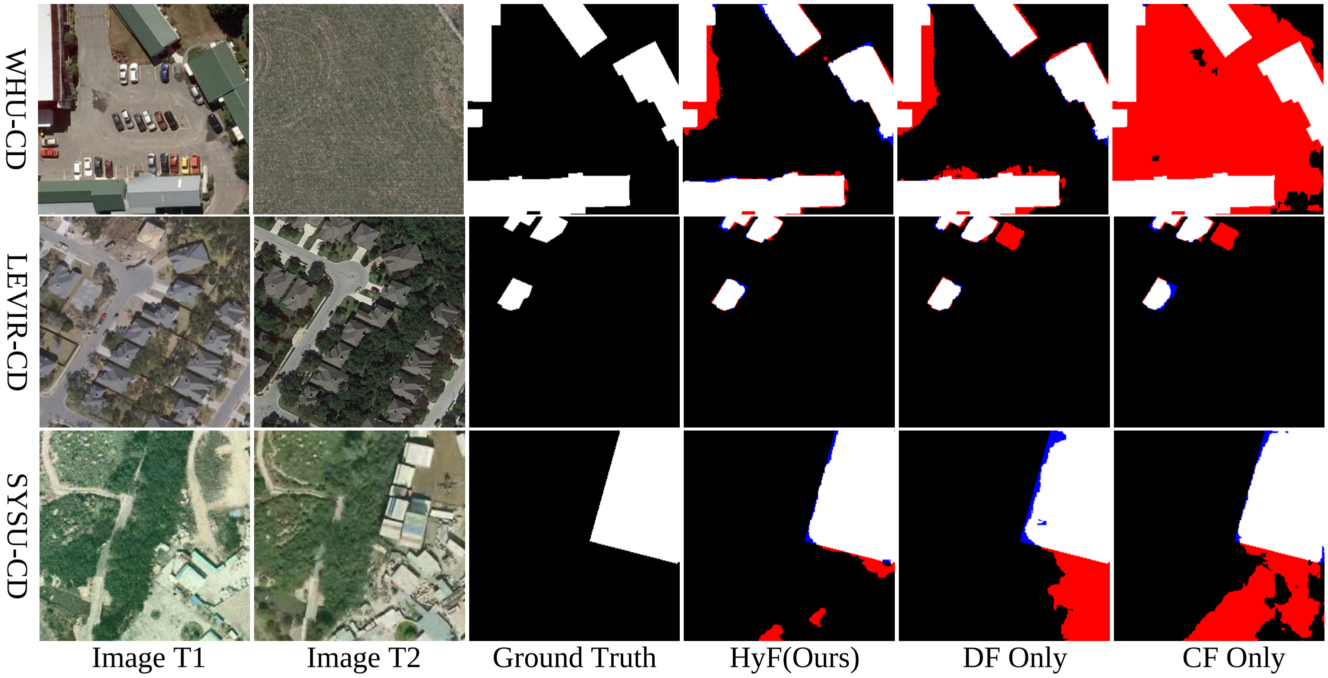


Fig. 12. Ablation study of the proposed hybrid features on WHU-CD, LEVIR-CD, and SYSU-CD datasets.

TABLE III
QUANTITATIVE RESULTS OF ALL METHODS ON THE SYSU-CD DATASET

Metric(%)	Pre	Rec	F1	IoU	OA
FC-EF	79.26	76.40	77.80	63.67	89.72
FC-Siam-Conc	81.80	75.76	78.66	64.83	90.31
FC-Siam-Diff	80.40	77.36	78.85	65.08	90.21
STANet	77.19	79.24	78.20	64.20	89.58
BIT	77.20	76.39	76.79	62.23	89.11
DSIFN	80.05	**80.78**	80.41	67.25	90.72
SNUNet	76.45	69.56	72.84	57.28	87.77
P2V	81.00	74.51	77.62	63.43	89.87
GeSANet	79.44	76.89	78.14	64.12	89.86
DS-HyFA-Net	**83.38**	79.02	**81.14**	**68.27**	**91.34**

TABLE IV
ABLATION STUDY OF THE PROPOSED HYBRID FEATURE ON WHU-CD, LEVIR-CD AND SYSU-CD DATASETS

Dataset	Model	Pre	Rec	F1	IoU	OA
WHU-CD	DF Only	88.64	**92.56**	90.56	82.74	99.16
	CF Only	84.57	83.39	83.98	72.38	98.62
	HyF(Ours)	**94.48**	91.74	**93.33**	**87.50**	**99.43**
LEVIR-CD	DF Only	90.87	**90.08**	90.47	82.60	99.03
	CF Only	91.66	88.75	90.18	82.12	99.02
	HyF(Ours)	**92.91**	89.12	**90.98**	**83.45**	**99.10**
SYSU-CD	DF Only	**83.95**	75.93	79.74	66.38	90.90
	CF Only	81.37	77.34	79.30	65.70	90.48
	HyF(Ours)	83.38	**79.02**	**81.14**	**68.27**	**91.34**

(g),FC-Siam-diff and SNUNet result in a significant quantity of FPs, whereas DSIFN and P2V exhibit a considerable number of FNs. SYSU-CD focuses not merely on building changes but also on road and land changes. It is influenced by other types of changes, especially sun light changes and seasonal changes, making the underlying image information more complex. Therefore, the proposed DS-HyFA-Net utilizes hybrid features to better explore the image information, achieving superior CD, as presented in Fig. 11 .

E. Ablation Studies

Ablation studies were carried out to further confirm the contributions of various modules within DS-HyFA-Net:

*1) **Integrating Distinct and Common Features:*** In this study, the DS-HyFA-Net utilizes multi-encoders to model CD by integrating distinct and common features. The subsequent experiments were undertaken to showcase the efficacy of the integration:

a) “Distinct Features (DF) Only”: In this experiment, Encoder II, responsible for extracting common features, was removed, while keeping everything else unchanged.

b) “Common Features (CF) Only”: In this experiment, Encoder I, responsible for extracting distinct features, was removed, while keeping everything else unchanged.

The quantitative experimental results in Table IV make clear that utilizing hybrid features simultaneously results in a superior performance compared to “Distinct Features (DF) Only” and “Common Features (CF) Only”. This is because hybrid features contain more information from the bitemporal RSIs. Specifically, the three well-recognized datasets using hybrid features resulted in improvements in the main evaluation metrics F1/IoU, when compared to “Distinct Features (DF) Only”: 2.77%/4.76% (WHU-CD), 0.51%/4.76% (LEVIR-CD), 1.4%/1.89% (SYSU-CD), respectively. Compared to “Common Features (CF) Only”, the DS-HyFA-Net has better F1/IoU for WHU-CD (9.35%/15.12%), LEVIR-CD (0.8%/1.33%), SYSU-CD (1.84%/2.57%), respectively.

Fig. 12 presents qualitative experimental results, demon-

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3471075

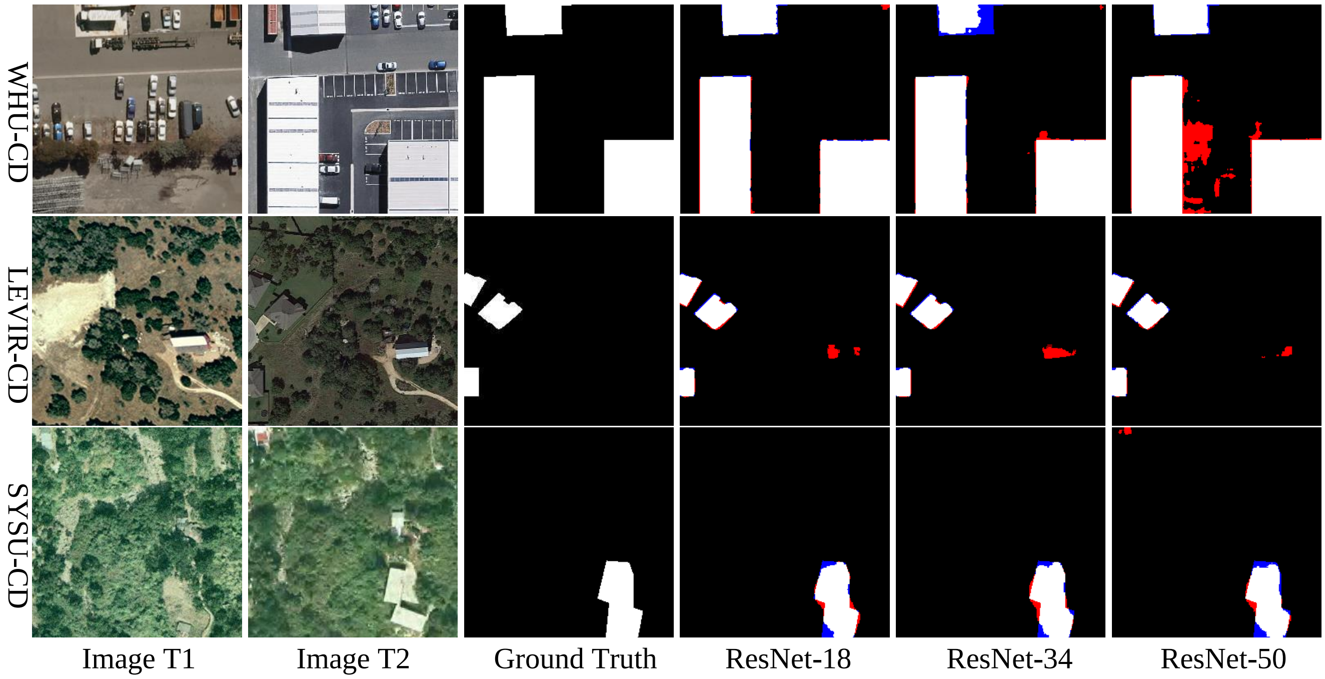


Fig. 13. Ablation study of Encoder I on WHU-CD, LEVIR-CD, and SYSU-CD datasets. The Encoder I of the DS-HyFA-Net is based on ResNet-18.

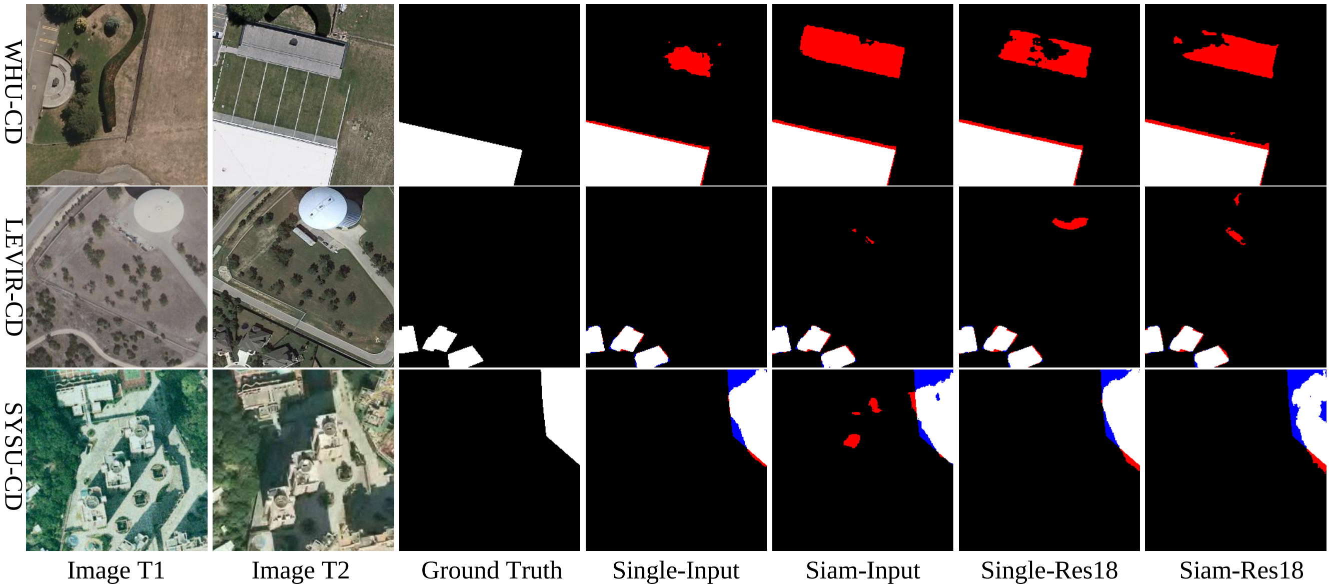


Fig. 14. Ablation study of Encoder II on WHU-CD, LEVIR-CD, and SYSU-CD datasets. The Encoder II of the DS-HyFA-Net is based on Single-Input.

strating that the use of hybrid features results in a reduction in both FPs and FNs. This is because relying solely on a single type of feature, whether distinct or common, poses a challenge in fully exploiting the information within bitemporal images. Existing models struggles to accurately detect changes with limited information. However, utilizing hybrid features allows for a more comprehensive exploration of image information, thus enabling the CD models to capture a more thorough understanding of the image context.

*2) **Encoder I**:* To validate the impact of Encoder I for the model, we performed ablation experiments using ResNet-18, ResNet-34, and ResNet-50. Table V presents the quantitative comparisons. ResNet-18 achieved the best primary evaluation metrics, F1 and IoU, on WHU-CD, with the fewest parameters and FLOPs. ResNet-18 achieved the highest IoU, which is the primary evaluation metric, on SYSU-CD. Both ResNet-18 and ResNet-34 achieved the best F1. However, ResNet-34 has a larger number of parameters and higher computational requirements, making ResNet-18 the preferable option overall. ResNet-34 achieved the best primary evaluation metrics, F1 and IoU, on LEVIR-CD. Nevertheless, it came with more parameters and higher FLOPs. Fig. 13 represents the quantitative experimental results, which demonstrate that ResNet-18, ResNet-34, and ResNet-50 are capable of accurately detecting

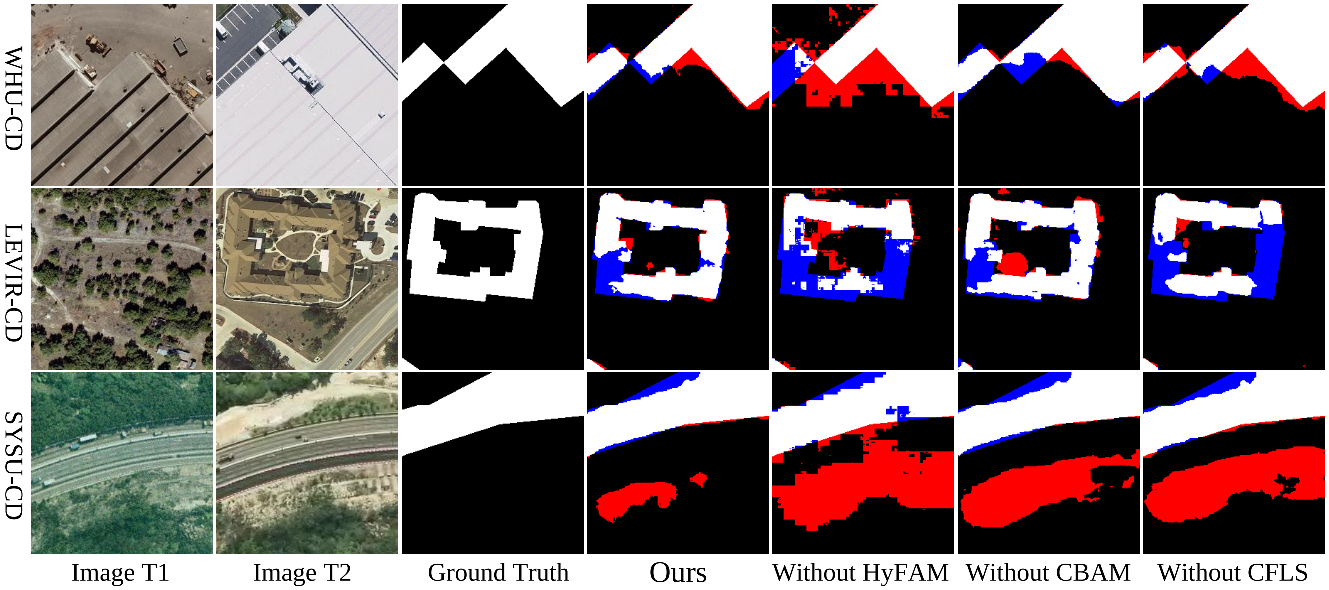


Fig. 15. Ablation study of the HyFA Module, CBAM and CFLS on WHU-CD, LEVIR-CD, and SYSU-CD datasets.

TABLE V
COMPARISON OF RESNET-18, RESNET-34 AND RESNET-50 ON WHU-CD, LEVIR-CD AND SYSU-CD DATASETS

Dataset	Backbone	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)	Parm(M)	FLOPs(G)
WHU-CD	ResNet-18	**94.48**	91.74	**93.33**	**87.50**	**99.43**	**10.58**	**28.54**
	ResNet-34	92.92	**92.94**	92.93	86.79	99.39	15.96	30.65
	ResNet-50	92.76	91.55	92.15	85.44	99.32	60.58	44.38
LEVIR-CD	ResNet-18	**92.91**	89.12	90.98	83.45	99.10	**10.58**	**28.54**
	ResNet-34	92.61	**89.91**	**91.24**	**83.89**	**99.12**	15.96	30.65
	ResNet-50	92.37	89.76	91.05	83.56	99.10	60.58	44.38
SYSU-CD	ResNet-18	**83.38**	79.02	**81.14**	**68.27**	**91.34**	**10.58**	**28.54**
	ResNet-34	81.52	**80.76**	**81.14**	68.26	91.15	15.96	30.65
	ResNet-50	82.73	78.96	80.80	67.78	91.15	60.58	44.38

TABLE VI
ABLATION STUDY OF THE ENCODER II ON WHU-CD, LEVIR-CD AND SYSU-CD DATASETS

Dataset	Model	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)	Parm(M)
WHU-CD	Siam-Input	89.60	89.38	89.49	80.98	99.09	15.00
	Single-Res18	91.43	**92.80**	92.11	85.38	99.31	12.19
	Siam-Res18	91.74	90.10	90.91	83.34	99.22	16.61
	Single-Input(Ours)	**94.48**	91.74	**93.33**	**87.50**	**99.43**	**10.58**
LEVIR-CD	Siam-Input	92.10	89.69	90.88	83.28	90.08	15.00
	Single-Res18	91.96	89.62	90.78	83.11	99.07	12.19
	Siam-Res18	91.74	**89.86**	90.79	83.13	99.07	16.61
	Single-Input(Ours)	**92.91**	89.12	**90.98**	**83.45**	**99.10**	**10.58**
SYSU-CD	Siam-Input	83.30	76.42	79.94	66.58	90.95	15.00
	Single-Res18	**84.30**	78.20	**81.14**	68.26	**91.42**	12.19
	Siam-Res18	82.20	78.77	80.45	67.29	90.97	16.61
	Single-Input(Ours)	83.38	**79.02**	**81.14**	**68.27**	91.34	**10.58**

changed regions. Among the three models, ResNet-18 exhibits fewer FPs.

*3) **Encoder II**:* The method put forward in this article introduces Encoder II as a common feature encoder. We believe that concatenating bitemporal images along channels as input, instead of using Siamese inputs, is more effective in extracting the common features of RS images. Additionally, this approach reduces the quantity of model parameters and computational requirements. The following experiments were undertaken to showcase the efficacy of Encoder II.

a) ”Siam-Input”: The input for Encoder II has been modified from concatenated bitemporal images to separate input for each individual bitemporal image.

b) ”Siam-Res18”: Utilized Siamese ResNet-18 as Encoder II.

c) ”Single-Res18”: Utilized single-branch ResNet-18 as Encoder II. The input consisted of the concatenation of bitemporal images from various channels.

Table VI showcases the quantitative comparisons, which indicate that our DS-HyFA-Net achieved the best primary evaluation metrics, namely F1 and IoU, with fewer parameters. Fig. 14 presents qualitative experimental results, demonstrating that our method resulted in fewer FPs and FNs and exhibited superior performance.

*4) **Common Feature Learning Strategy**:* Our model utilizes CFLS to enhance the effectiveness of learning common features. We performed ablation experiments by removing the technique for the purpose of determining its effectiveness. Both the qualitative comparisons in Fig. 15 and the quantitative comparisons in Table VII, VIII, IX demonstrate the positive impact of CFLS on the model.

*5) **Hybrid Feature Aggregation Module**:* We validate the HyFA module and the CBAM module through the following experiments:

TABLE VII
ABLATION STUDY OF THE HYFA MODULE, CBAM AND CFLS ON THE WHU-CD DATASET

HyFAM	CBAM	CFLS	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
×	✓	✓	90.38	85.96	88.11	78.75	98.99
✓	×	✓	93.18	90.78	91.96	85.12	99.31
✓	✓	×	90.82	90.24	90.53	82.69	99.18
✓	✓	✓	**94.48**	**91.74**	**93.33**	**87.50**	**99.43**

TABLE VIII
ABLATION STUDY OF THE HYFA MODULE, CBAM AND CFLS ON THE LEVIR-CD DATASET

HyFAM	CBAM	CFLS	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
×	✓	✓	89.15	86.63	87.87	78.37	98.78
✓	×	✓	92.04	89.36	90.68	82.95	99.06
✓	✓	×	91.85	**89.63**	90.73	83.02	99.07
✓	✓	✓	**92.91**	89.12	**90.98**	**83.45**	**99.10**

a) ”Without HyFA module”: In this experiment, the HyFA module was removed.

b) ”Without CBAM”: In this experiment, CBAM was removed.

The quantitative comparisons in Table VII,VIII,IX indicate that the performance of DS-HyFA-Net exhibited a noticeable decline without the assistance of the HyFA module and CBAM. Fig. 15 presents the qualitative experimental results, demonstrating that the model performs better and can achieve fewer FPs and FNs with the assistance of the HyFA module and CBAM.

F. Feature visualization

To provide additional validation for the efficacy of hybrid features, CFLS, the HyFA module, and CBAM, the Grad-CAM technique described in [62] was employed to visualize the input features for the predictor. As depicted in Fig. 17, the model using only Common Feature or only Distinct Feature, is susceptible to pseudo-changes and may miss real changes. Similarly, without the assistance of CFLS, the HyFA module, or CBAM, the model is difficult to identify real change.

G. Discussion

In this paper, we propose DS-HyFA-Net, a CD network that predicts changes by integrating the distinctness and commonality in bitemporal images. While the effectiveness and superiority of DS-HyFA-Net have been demonstrated, it is beneficial to further analyze and discuss its limitations, as this will provide valuable insights and guidance for future research. As illustrated in Fig. 16, DS-HyFA-Net exhibits certain false detections and missed detections at the edges of change regions. On one hand, this may be attributed to the fact that edges often contain high-frequency information, which traditional CNNs struggle to capture and process effectively, leading to the neglect of crucial edge details. On the other hand, it has been observed that the edges of change regions often contain gradual transitions or mixed pixel values, introducing a level of ambiguity that complicates detection. Therefore, future research should focus on exploring more efficient encoders, such as multi-scale CNNs, introducing a feature enhancement mechanism based on decoupled high-frequency and low-frequency features [63], and investigating feature fusion mechanisms to improve the detection of change region edges.

TABLE IX
ABLATION STUDY OF THE HYFA MODULE, CBAM AND CFLS ON THE SYSU-CD DATASET

HyFAM	CBAM	CFLS	Pre(%)	Rec(%)	F1(%)	IoU(%)	OA(%)
×	✓	✓	82.73	71.60	76.76	62.29	89.78
✓	×	✓	80.67	81.16	80.92	67.95	90.97
✓	✓	×	78.35	**81.18**	79.74	66.31	90.27
✓	✓	✓	**83.38**	79.02	**81.14**	**68.27**	**91.34**

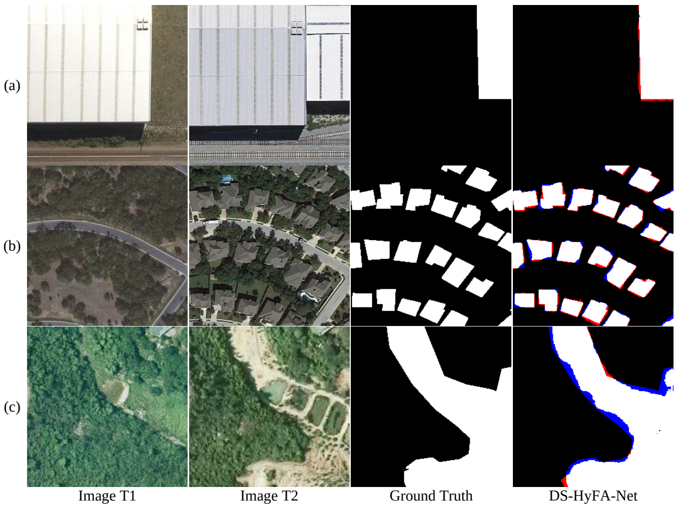


Fig. 16. Examples of Failure Cases in Detecting Challenging Areas.

V. CONCLUSION

In this work, an innovative method, DS-HyFA-Net, is proposed to enhance CD by thoroughly exploiting and utilizing image information. Different from previous methods that solely rely on the distinct features of bitemporal images, the DS-HyFA-Net focuses on both distinct and common features, or hybrid features. It addresses the limitations of existing CD models extracting distinct features. To extract hybrid features, the DS-HyFA-Net implements multiple encoders: Encoder I, to extract distinct features, and Encoder II, to extract common features. In addition, the DS-HyFA-Net incorporates CFLS to augment the common feature learning capability of Encoder II. Furthermore, the HyFA module in the DS-HyFA-Net effectively aggregates hybrid features. Compared with the existing methods, the DS-HyFA-Net has been validated to be superior and effective. Some persisting challenges include the detection of edges in change regions, the relatively large number of parameters, and significant computational complexity. In future studies, we plan to further explore image analysis and improve the efficiency of information utilization. Specifically, we will focus on the mechanisms of CD in

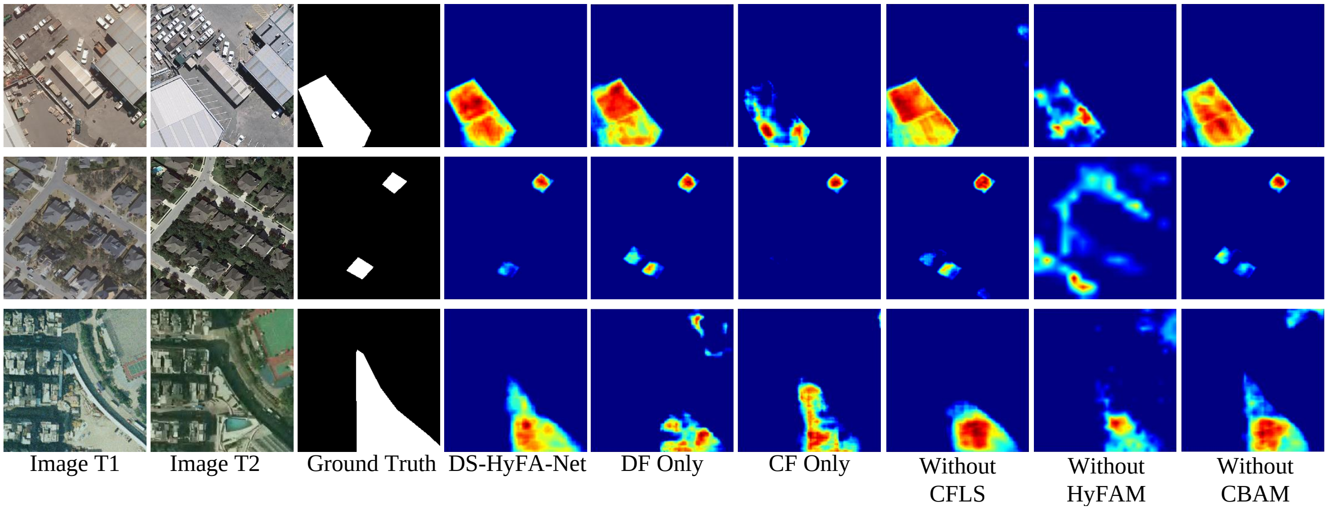


Fig. 17. Visual results of HyFA, CFLS, the HyFA Module and CBAM on the WHU-CD, LEVIR-CD, and SYSU-CD datasets.

bitemporal images, investigate feature extraction theories to develop efficient encoders capable of capturing high-frequency features, explore feature fusion mechanisms to enhance edge awareness in regions of change, and improve feature processing efficiency. Additionally, we will study feature mapping principles to achieve a smooth transition from feature space to image space. Our goal is to enhance CD performance while simultaneously reducing the number of model parameters and computational complexity.

REFERENCES

[1] A. Singh, "Review Article Digital change detection techniques using remotely-sensed data," International Journal of Remote Sensing, vol. 10, pp. 989–1003, 1989.

[2] C. Marin, F. Bovolo, and L. Bruzzone, "Building Change Detection in Multitemporal Very High Resolution SAR Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 5, pp. 2664–2682, 2015.

[3] X. Huang, L. Zhang, and T. Zhu, "Building Change Detection From Multitemporal High-Resolution Remotely Sensed Images Based on a Morphological Building Index," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 1, pp. 105–115, 2014.

[4] X. Li, F. Ling, G. M. Foody, and Y. Du, "A Superresolution Land-Cover Change Detection Method Using Remotely Sensed Images With Different Spatial Resolutions," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 7, pp. 3822–3841, 2016.

[5] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change Detection in Synthetic Aperture Radar Images Based on Deep Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 1, pp. 125–138, 2016.

[6] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric Cross-Attention Hierarchical Network Based on CNN and Transformer for Bitemporal Remote Sensing Images Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.

[7] J. Lei, Y. Gu, W. Xie, Y. Li, and Q. Du, "Boundary Extraction Constrained Siamese Network for Remote Sensing Image Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.

[8] M. Hu, C. Wu, B. Du, and L. Zhang, "Binary Change Guided Hyperspectral Multiclass Change Detection," IEEE Transactions on Image Processing, vol. 32, pp. 791–806, 2023.

[9] H. Chen, Z. Qi, and Z. Shi, "Remote Sensing Image Change Detection With Transformers," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2021.

[10] P. J. Howarth and G. M. Wickware, "Procedures for change detection using Landsat digital data," International Journal of Remote Sensing, vol. 2, pp. 277–291, 1981.

[11] R. D. Jackson, "Spectral indices in N-Space," Remote Sensing of Environment, vol. 13, no. 5, pp. 409–421, 1983.

[12] A. Singh, "Change detection in the tropical forest environment of northeastern India using Landsat," Remote sensing and tropical land management, vol. 44, pp. 273–254, 1986.

[13] J. A. Richards, "Thematic mapping from multitemporal image data using the principal components transformation," Remote Sensing of Environment, vol. 16, no. 1, pp. 35–46, 1984.

[14] S. Jin and S. A. Sader, "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances," Remote Sensing of Environment, vol. 94, no. 3, pp. 364–372, 2005.

[15] O. Ahlqvist, "Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 U.S. National Land Cover Database changes," Remote Sensing of Environment, vol. 112, no. 3, pp. 1226–1241, 2008.

[16] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," Remote Sensing of Environment, vol. 199, pp. 241–255, 2017.

[17] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change Detection Based on Deep Siamese Convolutional Network for Optical Aerial Images," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1845–1849, 2017.

[18] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully Convolutional Siamese Networks for Change Detection," in 2018 25th IEEE International Conference on Image Processing (ICIP), Athens: IEEE, Oct. 2018, pp. 4063–4067.

[19] J. Wang et al., "SSCFNet: A Spatial-Spectral Cross Fusion Network for Remote Sensing Change Detection," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 4000–4012, 2023.

[20] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," Remote Sensing, vol. 12, no. 10, p. 1662, May 2020.

[21] 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 Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–16, 2021.

[22] A. Vaswani et al., "Attention is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[23] K. Zhang, X. Zhao, F. Zhang, L. Ding, J. Sun, and L. Bruzzone, "Relation Changes Matter: Cross-Temporal Difference Transformer for Change Detection in Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.

[24] H. Jiang et al., "A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images," Remote Sensing, vol. 14, no. 7, p. 1552, Mar. 2022.

[25] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical Paired Channel Fusion Network for Street Scene Change Detection," IEEE Trans. on Image Process., vol. 30, pp. 55–67, 2020.

[26] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change Detection Meets Visual Question Answering," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.

[27] X. Zhang et al., "ADHR-CDNet: Attentive Differential High-Resolution Change Detection Network for Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.

[28] T. Lei et al., "Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–14, 2023.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778.

[30] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 38, no. 3, pp. 1171–1182, 2000.

[31] S. Jin and S. A. Sader, "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances," Remote Sensing of Environment, vol. 94, no. 3, pp. 364–372, 2005.

[32] S. Marchesi and L. Bruzzone, "ICA and kernel ICA for change detection in multispectral RS images," in 2009 IEEE International Geoscience and Remote Sensing Symposium, 2009, p. II-980-II–983.

[33] T. Celik, "Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering," IEEE Geoscience and Remote Sensing Letters, vol. 6, no. 4, pp. 772–776, 2009.

[34] T. Habib, J. Inglada, G. Mercier, and J. Chanussot, "Support Vector Reduction in SVM Algorithm for Abrupt Change Detection in Remote Sensing," IEEE Geoscience and Remote Sensing Letters, vol. 6, no. 3, pp. 606–610, 2009.

[35] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal Change Detection in Remote Sensing Images Using an Unsupervised Pixel Pairwise-Based Markov Random Field Model," IEEE Transactions on Image Processing, vol. 29, pp. 757–767, 2019.

[36] T. Liu et al., "Building Change Detection for VHR Remote Sensing Images via Local–Global Pyramid Network and Cross-Task Transfer Learning Strategy," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–17, 2022.

[37] L. Wang, L. Wang, Q. Wang, and P. M. Atkinson, "SSA-SiamNet: Spectral–Spatial-Wise Attention-Based Siamese Network for Hyperspectral Image Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–18, 2022.

[38] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A Hierarchical Attention Network for Change Detection With Bitemporal Very-High-Resolution Remote Sensing Images," IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 16, pp. 3867–3878, 2023.

[39] X. Wang et al., "Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery," International Journal of Applied Earth Observation and Geoinformation, vol. 122, p. 103456, Aug. 2023.

[40] K. Tan et al., "Change detection on multi-sensor imagery using mixed interleaved group convolutional network," Engineering Applications of Artificial Intelligence, vol. 133, p. 108446, Jul. 2024.

[41] M. Wang, X. Li, K. Tan, J. Mango, C. Pan, and D. Zhang, "Position-Aware Graph-CNN Fusion Network: An Integrated Approach Combining Geospatial Information and Graph Attention Network for Multiclass Change Detection," IEEE Trans. Geosci. Remote Sensing, vol. 62, pp. 1–16, 2024.

[42] W. Li, L. Xue, X. Wang, and G. Li, "ConvTransNet: A CNN–Transformer Network for Change Detection With Multiscale Global–Local Representations," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.

[43] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-Transformer Network With Multiscale Context Aggregation for Fine-Grained Cropland Change Detection," IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 15, pp. 4297–4306, 2022.

[44] M. Zhang, Z. Liu, J. Feng, L. Liu, and L. Jiao, "Remote Sensing Image Change Detection Based on Deep Multi-Scale Multi-Attention Siamese Transformer Network," Remote. Sens., vol. 15, no. 3, p. 842, 2023.

[45] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-Shaped Hierarchical Network for Remote-Sensing Image Change Detection," IEEE Trans. Geosci. Remote Sensing, vol. 61, pp. 1–14, 2023.

[46] 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.

[47] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-Resolution-Based Change Detection Network With Stacked Attention Module for Images With Different Resolutions," IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–18, 2022.

[48] 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 Sensing Lett., vol. 19, pp. 1–5, 2021.

[49] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," Artificial intelligence and statistics, vol. 38. 2015, pp. 562–570.

[50] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 166, pp. 183–200, 2020.

[51] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," International Journal of Applied Earth Observation and Geoinformation, vol. 105, p. 102591, 2021.

[52] M. Lin, G. Yang, and H. Zhang, "Transition Is a Process: Pair-to-Video Change Detection Networks for Very High Resolution Remote Sensing Images," IEEE Trans. on Image Process., vol. 32, pp. 57–71, 2022.

[53] Wang D, Haytham A, Pottenburgh J, et al. Hard attention net for automatic retinal vessel segmentation[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(12): 3384-3396.

[54] Zhang Y, Zhang H, Nasrabadi N M, et al. Multi-metric learning for multi-sensor fusion based classification[J]. Information Fusion, 2013, 14(4): 431-440.

[55] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change Detection on Remote Sensing Images Using Dual-Branch Multilevel Intertemporal Network," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.

[56] 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.

[57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International conference on machine learning, pmlr, 2015, pp. 448–456.

[58] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," nature, vol. 323, no. 6088, pp. 533–536, 1986.

[59] S. Ji, S. Wei, and M. Lu, "Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 1, pp. 574–586, 2018.

[60] X. Zhao, K. Zhao, S. Li, and X. Wang, "GeSANet: Geospatial-Awareness Network for VHR Remote Sensing Image Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–14, 2023.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[62] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[63] Hu T, Yan Q, Qi Y, et al. Generating content for hdr deghosting from frequency view[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 25732-25741.

Zilu Ying received the B.S., M.S., and Ph.D. degrees in electrical information engineering from Beihang University, Beijing, China, in 1985, 1988, and 2009, respectively.

He is a Full Professor with Wuyi University, Jiangmen, China. He is an Executive Director of the Guangdong Society of Image and Graphics and a member of the Signal Processing Branch of the Chinese Institute of Electronics. His research interests include biometric extraction and pattern recognition.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3471075

Tingfeng Xian (Student Member, IEEE) received the B.S. degree from Wuyi University, Jiangmen, China, in 2022. He is pursuing the master's degree with the College of Electronics and Information Engineering, Wuyi University, Jiangmen, China.

His research interests include semantic segmentation, change detection and pattern recognition.

Yikui Zhai (Senior Member, IEEE) received his Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013.

Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is a Full Professor now. He is also Associate Dean of the School of Electronics and Information Engineering in Wuyi University, since 2021. He has been a Visiting Scholar with Department of Computer Science, the Universita degli Studi di Milano, Italy, during June 2016 to June 2017, August 2023 and January 2024. His research interests include: Image Processing, Deep Learning, Optical Character Recognition, Object Detection, UAV Change Detection, Self Supervise Learning.

Xudong Jia received the B.S. and M.S. degrees from Beijing Jiaotong University, in 1983 and 1986, respectively, the second M.S. degree from the University of Toronto, Canada, in 1992, and the Ph.D. degree from the Georgia Institute of Technology, in 1996.

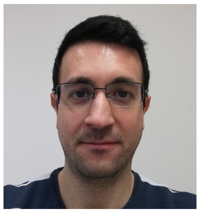
He is a Professor and the Associate Dean with the College of Engineering and Computer Science, CSUN. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality. He is an Associate Editor for the IEEE Intelligent Transportation Systems Society and the IEEE Transaction on Intelligent Transportation Systems.

Hongsheng Zhang (Senior Member, IEEE) received the B.Eng. degree in computer science and technology and the M.Eng. degree in computer applications technology from South China Normal University, Guangzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree in earth system and geoinformation science from The Chinese University of Hong Kong, Hong Kong, China, in 2013.

He is currently an Assistant Professor with the Department of Geography, The University of Hong Kong, Hong Kong, China. His research interests include remote-sensing applications in tropical and subtropical areas, with a focus on the urban environment and coastal sustainability monitoring, using multisource remote sensing data fusion and image pattern recognition techniques.

Jiahao Pan (Student Member, IEEE) received the B.S. degree from Wuyi University, Jiangmen, China, in 2022. He is pursuing the master's degree with the College of Electronics and Information Engineering, Wuyi University, Jiangmen, China.

His research interests include computer vision, change detection and object detection.



Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in Industrial and Information Engineering from the Universita degli Studi della Campania Luigi Vanvitelli, Italy, in 2019.

From 2019 to 2022, he was a post-doctoral researcher at the Universita degli Studi di Padova, Italy. He is currently an Assistant Professor at the Department of Computer Science of the Universita degli Studi di Milano, Italy. He is Co-chair of the Intelligent Measurement Systems Technical Committee (TC-22) of the IEEE Instrumentation and Measurement Society since 2023. His research activities are focused on theoretical, methodological, and applied aspects of computational intelligence for signal and image processing.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Universita degli Studi di Milan, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Universita degli Studi di Milan since 2014.

He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Vincenzo Piuri (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from Politecnico di Milan, Milan, Italy, in 1984 and 1988, respectively.

He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with Politecnico di Milan from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a startup company, Sensuresrl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010) and was active in industrial research projects with several companies. His main research and industrial application interests are intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, signal and image processing, biometrics, intelligent measurement systems, industrial applications, distributed processing systems, Internet-of-Things, cloud computing, fault tolerance, application-specifi digital processing architectures, and arithmetic architectures.

Dr. Piuri is an ACM Fellow.

Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

He was an Assistant Professor with the Department of Information Technologies, Universita degli Studi di Milan, Milan, from 2002 to 2015, where he was an Associate Professor with the Department of Computer Science from 2015 to 2020. He has been a Full Professor with the Universita degli Studi di Milan since 2020. Original results have been published in over 150 papers in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design.

Prof. Scotti is an Associate Editor of the IEEE Transactions on Human–Machine Systems and the IEEE Open Journal of Signal Processing. He is serving as a Book Editor (Area Editor, section Less-Constrained Biometrics) of the Encyclopedia of Cryptography, Security, and Privacy (3rd Edition, Springer). He has been an Associate Editor of the IEEE Transactions on Information Forensics and Security, Soft Computing (Springer) and a Guest Coeditor of the IEEE Transactions on Instrumentation and Measurement.