Dual-Dimension Feature Interaction for Semantic Change Detection in Remote Sensing Images

Biao Wang, *Member, IEEE*, Zhenghao Jiang, Weichun Ma, Xiao Xu, Peng Zhang, Yanlan Wu, and Hui Yang

***Abstract*—Remote sensing semantic change detection (SCD) involves extracting information about changes in land cover/land use (LCLU) within the same area at different times. This issue is of crucial significance in many Earth observation tasks, such as precise urban planning and natural resource management. However, the current methods primarily focus on spatial feature extraction, lacking awareness of temporal features. Consequently, there are challenges in extracting change features, making distinguishing intraclass and interclass differences difficult. This also contributes to pseudochange, posing challenges for SCD tasks. To overcome the limitations of existing methods, we present a dual-dimension feature interaction network (DFINet) for SCD. First, to enhance the assessment and perceptual abilities related to intraclass and interclass differences, we introduce a temporal difference feature enhancement (TDFE) module. This module comprehensively captures features from the temporal dimension. Then, to address the interrelation between multitemporal and multilevel features, we investigate the feature selection interaction (FSIA) and interaction attention modules (IAM), which enable multidimensional deep fusion and interaction of change features. This enhances the capacity for information transfer and integration among the features within multitemporal remote sensing images (RSIs). The experimental results demonstrate that, compared to existing methods, the proposed architecture achieves a significant improvement in accuracy. Additionally, the design enhancements added to DFINet boost the practicality of remote sensing SCD, underscoring its substantial research value.**

***Index Terms*—Dual-dimension, interclass, intraclass, remote sensing images (RSIs), semantic change detection (SCD).**

Manuscript received 14 January 2024; revised 5 March 2024 and 6 April 2024; accepted 24 April 2024. Date of publication 29 April 2024; date of current version 9 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 41901282, Grant 42101381, and Grant 41971311, in part by the International Science and Technology Cooperation Special under Grant 202104b11020022, in part by the Anhui Provincial Natural Science Foundation under Grant 2308085MD126, in part by the Hefei Municipal Natural Science Foundation under Grant 202323 and in part by the Natural Resources Science and Technology Program of Anhui Province under Grant 2023-K-7. *(Corresponding author: Zhenghao Jiang.)*

Biao Wang and Zhenghao Jiang are with the School of Resources and Environmental Engineering, Anhui University, Hefei 230601, China (e-mail: wangbiao-rs@ahu.edu.cn; j-zh@stu.ahu.edu.cn).

Weichun Ma is with the Anhui Province Basic Surveying and Geographic Information Center, Hefei 230031, China (e-mail: gisgirl@126.com).

Xiao Xu is with the Department of Art and Design, Jining Polytechnic, Jining 272007, China (e-mail: jnzyysx@163.com).

Peng Zhang and Yanlan Wu are with the School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: pzhangrs@ahu.edu.cn; wuyanlan@ahu.edu.cn).

Hui Yang is with the Institutes of Physical Science and Information Technology, Anhui University, Hefei 230031, China (e-mail: yanghui@ahu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3394571

I. Introduction

REMOTE sensing change detection (CD) is an active research field with a wide range of applications in Earth observation, aims to quantitatively and qualitatively detect anthropogenic and natural surface changes occurring at the same geographic location over different periods [1], [49], [54]. Existing binary change detection (BCD) methods primarily can show the locations and shapes of changes in land features, but they often cannot provide detailed information about the types of changes. Therefore, the pursuit of a novel change detection method capable of identifying not only "where" the changes occurred but also "what" the changes are has become a focus of current research. This approach is referred to as multiclass change detection (MCD) or semantic change detection (SCD) [2], [3]. Such methods play a critical role in various fields, including urban management [4], environmental monitoring [5], and damage assessment [6].

In recent years, deep learning techniques have been widely applied in CD by utilizing remote sensing images (RSIs). These techniques can identify and extract nonlinear features based on the statistical consistency of multitemporal observation data. Furthermore, deep learning methods demonstrate robust capabilities in perceiving deep-level abstract feature information, enabling them to effectively handle changes in complex scenarios or under various conditions. Examples of such techniques include generative adversarial networks [7], joint sparse representation [8], and spatial structure extraction based on convolutional neural networks (CNNs) [9]. However, existing methods often lack the necessary support for understanding the relationships between change classes. To address such issues, SCD methods have been proposed. Current research directions include using the feature extraction results of the backbone network of the encoding stage as inputs for CD decoding. Additionally, features from different scales in the encoding stage are extracted and integrated as inputs for CD decoding [2], [10], [11], [12], [13], [14]. The introduction of these methods has greatly improved the accuracy of remote sensing SCD results. The proposed SCD frameworks and methods provide reliable technical support for subsequent research work.

Existing SCD methods have shown dominant advantages in various scenarios, but there remain two limitations that need to be addressed. First, as illustrated in Fig. 1, for the SCD task focusing on regions with high intraclass variation but low interclass variance [49], these approaches face challenges in decoupling, refining, and effectively fusing spatial and temporal information. Consequently, this limits their ability to precisely locate change

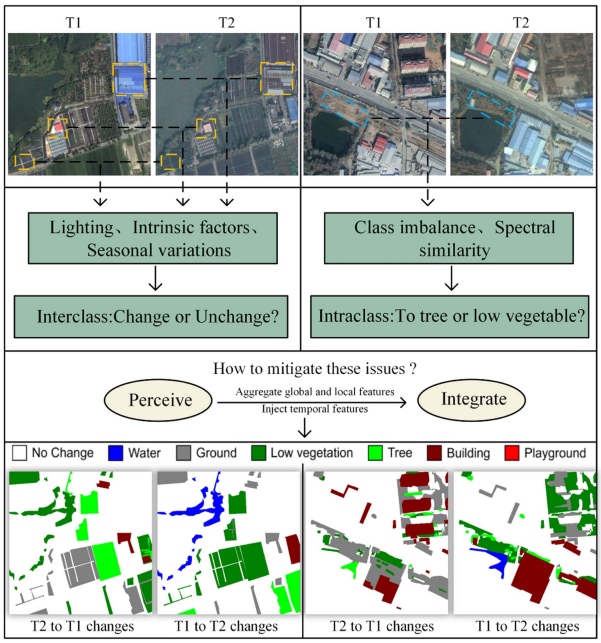

Fig. 1. SCD proves to be a formidable task due to the pronounced intraclass variability and subdued interclass differentiation inherent in RSIs.

areas and suppress false alarms [20]. Within the feature fusion phase of dual-temporal RSIs, the existing works [15], [16], [17], [18], [19] mostly emphasize the manipulation of spatial feature dimensions, neglecting the critical temporal dimension information [20]. Furthermore, shallow-level features encompass rich detail information [44], [45], while deep-level features distill more abstract contextual information [26]. Current methods [37], [46], [47] struggle to bridge the semantic gap efficiently [48] between shallow and deep-level features, potentially subjecting valuable semantic information to interference from redundant data, thereby directly impacting the precision of SCD [26].

These issues persist, leading to the ongoing interference of pseudochange from lighting, seasons, or intrinsic factors in SCD, thereby affecting its accuracy. To remediate these deficiencies, in this article, we introduce the dual-dimension feature interaction network (DFINet) model. The primary contributions of this research are outlined as follows.

1) Considering the problem of underutilized temporal dimension information in multilevel feature fusion between two RSIs, we introduce the temporal difference feature enhancement (TDFE) module. The dual-temporal feature fusion phase transforms discrete feature information into continuous change data.
2) In the multiscale feature fusion stage, we introduce the feature selection interaction (FSIA) module, which is designed to perform cross-fusion and selection of features enhanced through TDFE processing. The objective is to convey valuable change information to the decoder. The interaction attention module (IAM) incorporates cross-temporal interaction attention mechanisms and extends the module's local receptive field, thereby enhancing its capability to extract complex terrain features.

The rest of this article is organized as follows. Section II discusses the literature work on SCD of RSIs. Section III introduces the DFINet network architecture proposed in this study and the relevant experimental datasets. Section IV describes our experimental setup. Section V presents and analyzes the experimental results. Finally, Section VI concludes this article.

II. Related Work

A. Temporal Consistency in Change Detection

CD can be understood as an ongoing task of monitoring temporal changes. Colloquially speaking, it involves identifying continuous changes over time caused by internal and external factors. RSIs exhibit higher intraclass variation, leading to reduced separability between different entities such as roads and buildings. Consequently, when detecting changes and measuring their magnitude between a pair of bitemporal images, various salt-and-pepper noises often appear in the detection map. These noises represent pseudochange [49]. By discovering and identifying slow-changing features among rapidly changing input features, pseudochange can be perceived and distinguished [50]. Based on this, Ye et al. [20] and Lin et al. [51] designed P2V-Net and AFCF3D-Net. By incorporating temporal features to simulate multiple frames of signals in videos, they transformed the task of rapid discontinuity learning into a continuity detection task. Compared to 2-D convolution, 3-D convolution can not only handle high-dimensional features but also provide higher parallelism and address the issue of long-term dependencies [52]. These model architectures achieved state-of-the-art results on BCD datasets such as LEVIR and CDD. Because SCD tasks encounter challenges such as significant intraclass variability and small interclass differences, considering introducing the temporal dimension from BCD into SCD for expansion could be worthwhile.

B. Perception and Interaction of Features

Identifying objects of different sizes is a challenge in computer vision, and a feature pyramid has always been a commonly used structure [53]. By integrating deep abstract semantic information with shallow rich detailed information, more accurate recognition results can be obtained. In the domain of CD in RSIs, high-resolution images encompass a diverse range of land cover types, exhibiting considerable variations in scale. Directly fusing features from different layers in the model might result in information redundancy. Therefore, an essential aspect of CD involves effectively filtering and connecting features to mitigate semantic gaps. Lin et al. [51] used cross-fusion of adjacent features to reduce discrepancies and confusion in the network. However, the high computational complexity of this approach remains a concern. We believe that enhancing feature selection and lightweight modules could address this issue, and this forms a motivation for our work.

C. Semantic Change Detection

SCD entails capturing alterations in land cover/land use (LCLU) within a given area across various periods [11]. As

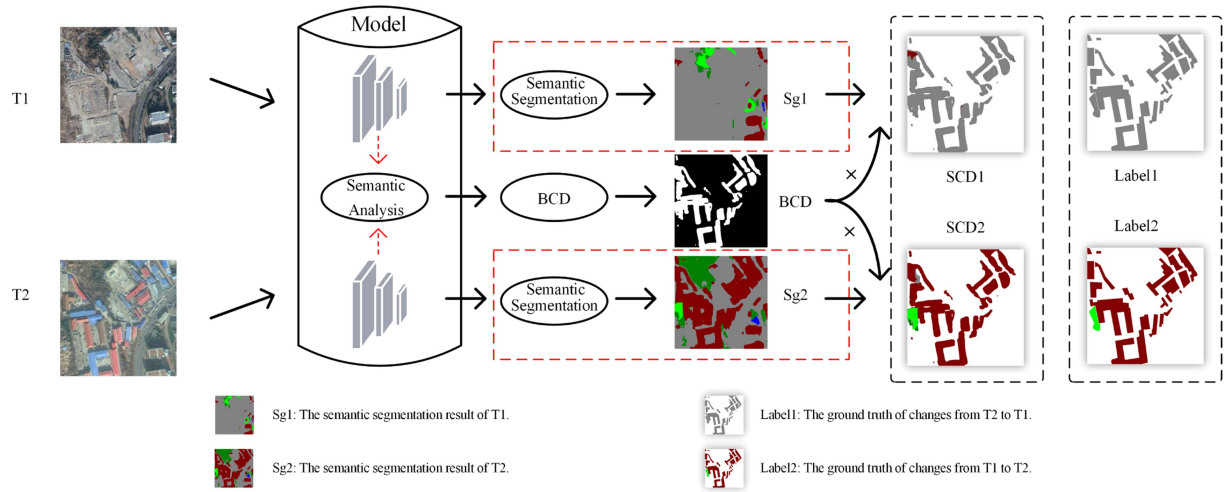

Fig. 2. Implementation mechanism of SCD, where the red box represents the added semantic segmentation task based on BCD and the × represents pixelwise multiplication. The change result is obtained by multiplying the pixel values. It should be noted that SCD reflects vectorized change scenarios, capturing changes from T1 to T2 and from T2 to T1.

illustrated in Fig. 2, SCD specifies what changes occurred. This enables multiclass categorization of land cover changes. In postprocessing, the BCD results are multiplied pixelwise with the multiclass results from T1 and T2 temporal images, yielding the final semantic segmentation results. Specifically, in the BCD result, pixel values are either 0 or 1, whereas in the semantic segmentation result, pixel values range from 0 to n, where n represents the categories of land objects. When the BCD pixel value at a specific point is 0, it indicates that no change has occurred at that location. If the pixel value is 1, the specific change type is determined based on the pixel value from the semantic segmentation result.

Compared to earlier methods that involve preclassification followed by comparison, utilizing CNNs for deep semantic feature extraction enhances the recognition of change features, reducing interference from classification results in change detection. Peng et al. [11] employed an innovative framework comprising two encoders and two decoders within a Siamese U-Net architecture, aiming to refine coarse boundaries and enhance the accuracy of SCD. Chen et al. [39] proposed a feature-constrained change detection network, which imposes constraints on features during bitemporal feature extraction and effectively robustly integrates bitemporal features. Yang et al. [13] designed an asymmetric Siamese network to locate and identify semantic changes through feature pairs obtained from modules of widely different structures. Ding et al. [3] introduced the SSCD-l CNN architecture for SCD and designed BiSRNet, achieving good results on the SECOND dataset. However, this method neglects the utilization of intermediate features in the encoding part, resulting in insufficient capability for pseudochange identification.

III. Methodology

A. Overall Framework

As shown in Fig. 3, a Siamese residual network [22] is used as the backbone to extract semantic information between the two temporal phases, which serves as input for both multiclassification and CD. The simplified calculations are

$$\alpha_1, \alpha_2 = \varepsilon_s(T_1, T_2) \tag{1}$$

$$S_1, S_2 = I^S\left(\alpha_1^l, \alpha_2^l\right) \tag{2}$$

$$C = \theta^\tau\left(\alpha_1^M, \alpha_2^M\right) \tag{3}$$

$$SC_1, SC_2 = C \cdot (S_1, S_2). \tag{4}$$

Specifically, T_1 and T_2 represent the input dual-temporal images, while ε_s represents the encoder part of the model. After feature extraction by ε_s, α_1 and α_2 are obtained, where α_1^l and α_2^l represent the output features at the deepest layer, and α_1^M, α_1^M represent multiscale features. In the semantic segmentation branch, I^S represents processing through the IAM module and the semantic segmentation head, resulting in S_1 and S_2 as the output dual-temporal semantic segmentation results. In the CD branch, θ^τ represents multidimensional feature perception and interactive integration operations on multiscale features, obtaining the change result C. In (4), $\cdot$ denotes pixelwise multiplication, and SC_1 and SC_2 are the final SCD results.

B. TDFE Module

In response to the pseudochange caused by factors such as lighting, intrinsic factors, seasonal variations, and external environments [23], [24], as well as the complexity of multiclass tasks and diverse CD directions in SCD, we propose a feature enhancement module based on spatial-temporal feature extraction at different scales denoted TDFE. This module decouples spatial and temporal information at different scales, processes them separately, and then integrates them effectively for classifying complex land features [20].

As shown in Fig. 4, TF denotes the learned temporal feature using 3-D convolution to capture the temporal information between the two-time phases. This additional time branch helps

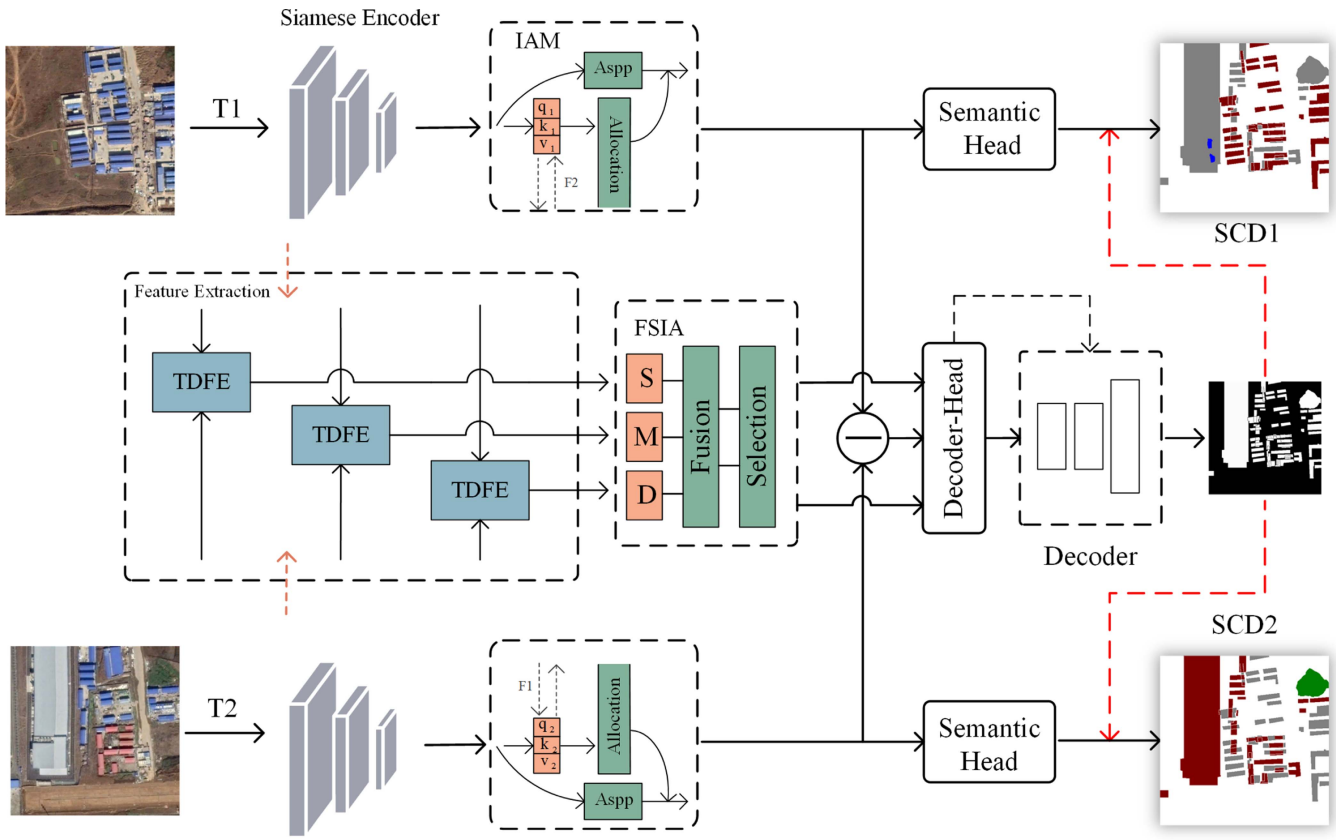

Fig. 3. DFINet model is based on a Siamese encoding–decoding structure, wherein the FSIA module, S, M, and D represent shallow, intermediate, and deep-level feature information.

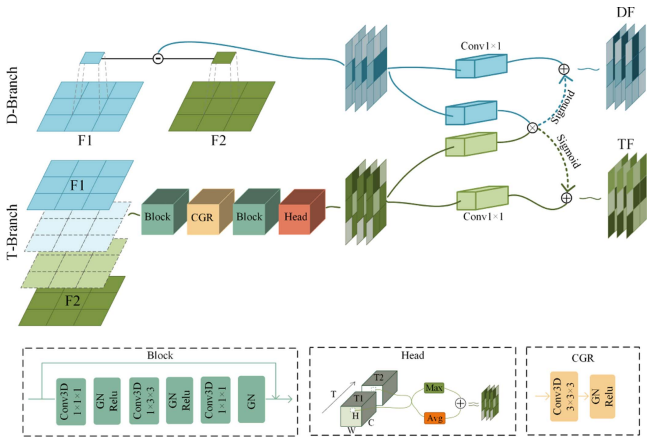

Fig. 4. Architecture of the proposed TDFE module. D-Branch and T-Branch represent the difference branch and the temporal branch, while DF and TF represent the extracted difference features and the extracted temporal features.

enhance the extraction of genuine difference features. The simplified calculations are

$$\text{Int} = C_d^{\tau}\left(F_1, F_2\right). \tag{5}$$

It should be noted that Int represents the input of the T-branch, C_d^{τ} denotes concatenation operations of features along the temporal dimension. Between F_1 and F_2, two temporal phases are further extracted based on their pixel value differences to extend the features. Since 3-D convolutions have higher computational complexity than 2-D convolutions, we only added one $3 \times 3 \times 3$ convolution in the CGR module to reduce the computational load. In Fig. 4, Max and Avg in the head represent taking the maximum and average values of the features in the temporal dimension [20]. Then, 1×1 convolutions are utilized to unify the channel dimensions [25], and they are cross-fused to enhance the extracted features.

C. FSIA Module

As shown in Fig. 5, the module consists of a detail branch and a context branch [26], [27]. F contains richer details and edge information, while M and D contain more accurate contextual information. To fully integrate multiscale feature information [28], [29] and select the enhanced semantic information at the same location but at different stages, we first perform a cross-multiplication between S, M, and D to obtain F1, F2, and F3. Subsequently, the features derived from F1, F2, and F3 are condensed in the channel dimension to obtain W1, W2, and W3, transitioning their tensor structure from $B \times C \times H \times W$ to $B \times 1 \times H \times W$. These three tensors are then concatenated along the channel dimension and synthesized into a unified weight vector fusing a 1×1 convolution operation. CW represents the results obtained from the cross-scale fusion of features through S, M, and D. To control this outcome, the sigmoid activation

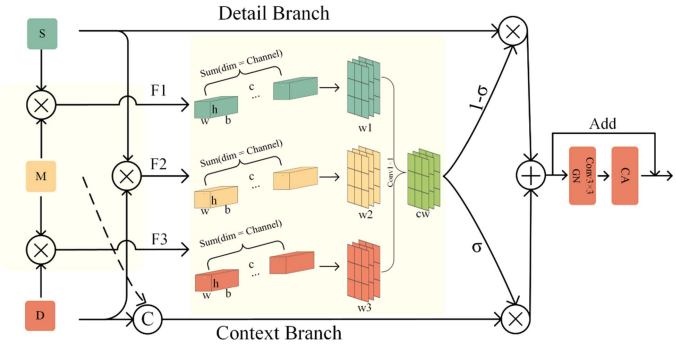

Fig. 5. Architecture of the proposed FSIA module. Here, b, c, h, and w, respectively, represent the batch size, number of channels, height, and width of the input feature maps, while Add represents the elementwise addition of pixel values between feature maps.

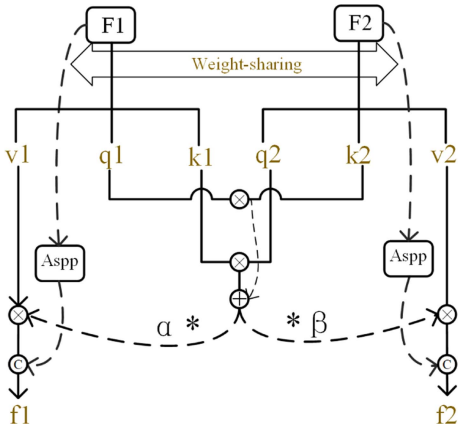

Fig. 6. Architecture of the proposed IAM.

function is applied, as illustrated by the following formula:

$$\sigma = \text{Sigmoid}(CW) \tag{6}$$

$$F_D = S \times (1 - \sigma) \tag{7}$$

$$F_C = C_d^c(M, D) \times \sigma. \tag{8}$$

F_D and F_C, respectively, represent the extracted detail features and context features. C_d^c denotes concatenation operations of features along the channel dimension. When $\sigma > 0.5$, the model's output results are more inclined to the contextual branch. Otherwise, it tends to favor the detailed branch [26]. Finally, a 3×3 convolution operation and a direction-aware attention module (CA) [30] are applied to explore further feature relationships in both the spatial and channel dimensions. Unlike the commonly used squeeze-and-excitation module (SE) and convolutional block attention module (CBAM), the CA module not only focuses on channel attention but also takes into consideration its spatial relationships. It combines channelwise attention with spatial attention, generating a pair of direction-aware and position-sensitive attention maps for the feature maps. These attention maps are applied complementarily to the input feature maps to enhance the representation of the objects of interest.

D. IAM Module

In SCD, it is essential not only to effectively detect regions but also to identify the types of changes that have occurred. This necessitates the model taking into account both the spatial correlations between temporal images and the semantic relevance and consistency across them. Self-attention has been to capture spatial dependencies between any two positions in feature maps, allowing it to obtain long-range contextual information [32]. This makes it widely applicable in CD. Considering the characteristics of SCD tasks, this article introduces a cross-temporal interactive global and local attention module called IAM.

As shown in Fig. 6, the first step involves adding cross-temporal interaction connections to the existing self-attention mechanism [3], [33], [34], [35]. This can be expressed by the following formula:

$$F_G = M_S\left[(k_1 \times q_2), (k_2 \times q_1)\right] \tag{9}$$

$$A_1, A_2 = F_G \times (v_1 \times \alpha, v_2 \times \beta) \tag{10}$$

k_i, v_i, and q_i represent three relationship vectors obtained through a 1×1 convolution for F_i, while M_S represents taking the mean after applying the softmax activation function and then obtaining the global features F_G. By multiplying k and q from different time phases, the correlation features between the two-time phases are enhanced. Then, they are multiplied by v_1 and v_2 to enhance the spatial feature information for each respective time phase. To distinguish the differences between the two temporal phases, we introduced learnable dynamic parameters α and β, which serve as weights for F_1 and F_2. In the second step, we incorporate a Siamese local feature learning branch to compensate for the limited local feature learning in the self-attention module. The branch incorporates the atrous spatial pyramid pooling (ASPP) module, which serves two purposes [21]. First, it aggregates local features. Second, it leverages the large receptive field of the ASPP module to obtain local information at multiple scales. Finally, we concatenate the features from both branches along the channel dimension to obtain the final features f_1 and f_2.

IV. Experiment

A. Datasets Environment

To demonstrate the effectiveness of the model for SCD, we used the SECOND dataset as the basis for our experiments. To ensure the effectiveness of our method and subsequent comparative experiments, we randomly divided the dataset into training, validation, and test sets at a 4:1:1 ratio. Example images from the dataset are shown in Fig. 7.

The SECOND dataset [38] consists of 4662 image pairs, with 2968 pairs available for training. The geographical locations associated with these image pairs are distributed across cities such as Hangzhou, Chengdu, and Shanghai, with resolutions ranging from 0.5 to 3 m and image sizes of 512×512 pixels. The primary change types in this dataset involve transitions between land cover categories, which include water bodies, bare ground, low-lying vegetation, trees, buildings, and sports fields. In each data pair, there are two images corresponding to different periods, and each image is associated with an annotation map

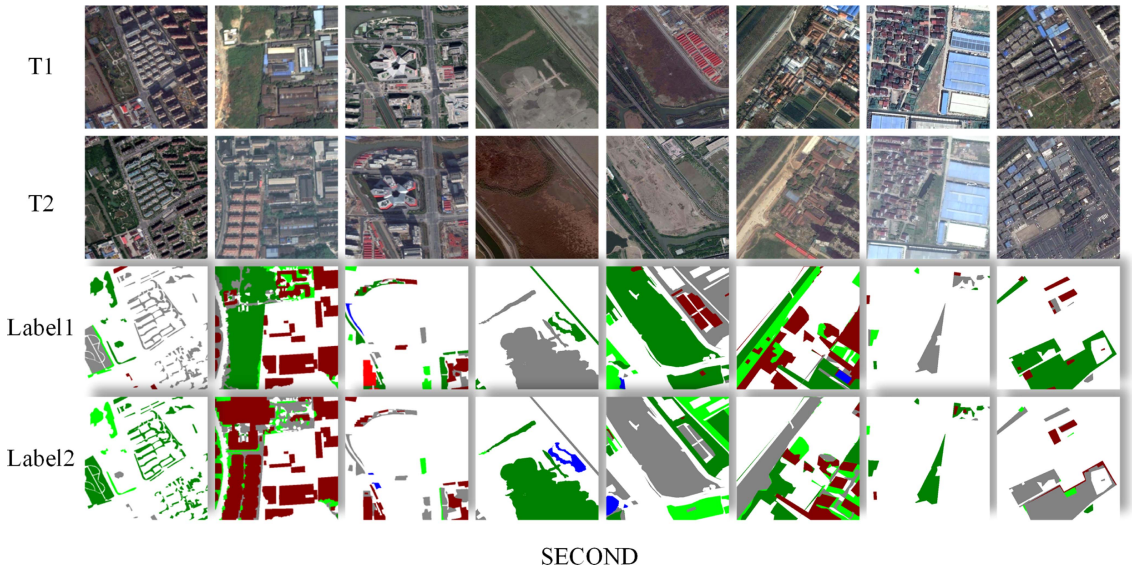

Fig. 7. SECOND dataset is a SCD dataset that covers water bodies, ground, low vegetation, trees, buildings, and sports fields.

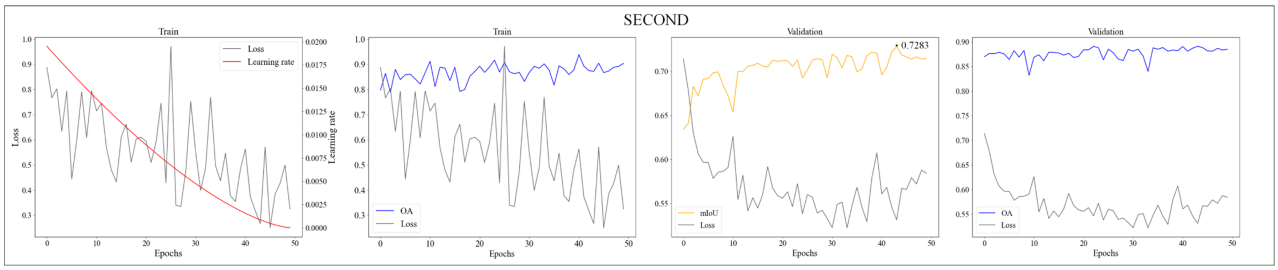

Fig. 8. Performance of the DFINet on the train and validation sets. The red curve in the graph represents the learning rate, the grey curve represents the loss, the orange curve represents mIoU, and the blue curve represents OA. We highlight the best accuracy on the validation set with a black dot on the curves.

that indicates the areas of change and the land cover categories within those changed regions over time. The experiments were conducted on a Windows 10 operating system with an Intel Core i7-9700F CPU and an NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of RAM. The model was implemented using the PyTorch framework. All experiments were conducted using the same experimental parameters, including a batch size of 4, training for 50 epochs, an initial learning rate of 0.02, and a weight decay of 0.0003. The optimization method used stochastic gradient descent with Nesterov momentum [40]. The image augmentation strategies include random flipping and rotation of the image pairs during training.

In Fig. 8, we plot the performance of the DFINet on the training and validation sets. We observe from Fig. 8 that the DFINet can achieve satisfactory performance within 50 epochs. To further validate our models on the test sets, we save the weights corresponding to the highest validation accuracy as checkpoints for testing.

B. Metrics

SECOND is an SCD dataset, so we use the mean intersection over union (mIoU), the separation kappa coefficient (SeK) [3], and overall accuracy (OA) [41] to evaluate SCD results.

Let $Q = \{q_{ij}\}$ be the confusion matrix, where i, j represents the count of pixels belonging to class i and classified as class j. OA measures the proportion of correctly classified samples over all samples [3]. Since the change area in the SECOND dataset only constitutes 19.87%, OA is easily influenced by the unchanged areas. Therefore, further evaluation using mIoU and kappa is needed

$$\mathrm{OA} = \sum_{i=0}^{N} q_{ii} \Big/ \sum_{i=0}^{N}\sum_{j=0}^{N} q_{ij} \tag{11}$$

$$\mathrm{mIoU} = \left(\mathrm{IoU}_{\mathrm{uc}} + \mathrm{IoU}_{c}\right)/2 \tag{12}$$

$$\mathrm{IoU}_{\mathrm{uc}} = q_{00} \Big/ \left(\sum_{i=0}^{N} q_{i0} + \sum_{j=0}^{N} q_{0j} - q_{00}\right) \tag{13}$$

$$\mathrm{IoU}_{c} = \sum_{i=1}^{N}\sum_{j=1}^{N} q_{ij} \Big/ \left(\sum_{i=0}^{N}\sum_{j=0}^{N} q_{ij} - q_{00}\right) \tag{14}$$

where $\mathrm{IoU}_{\mathrm{uc}}$ represents the unchanged areas, and IoU_{c} represents the over for changed areas. SeK is a metric used in combination with OA to provide a better evaluation of the performance of multiclass classification tasks, and it can be

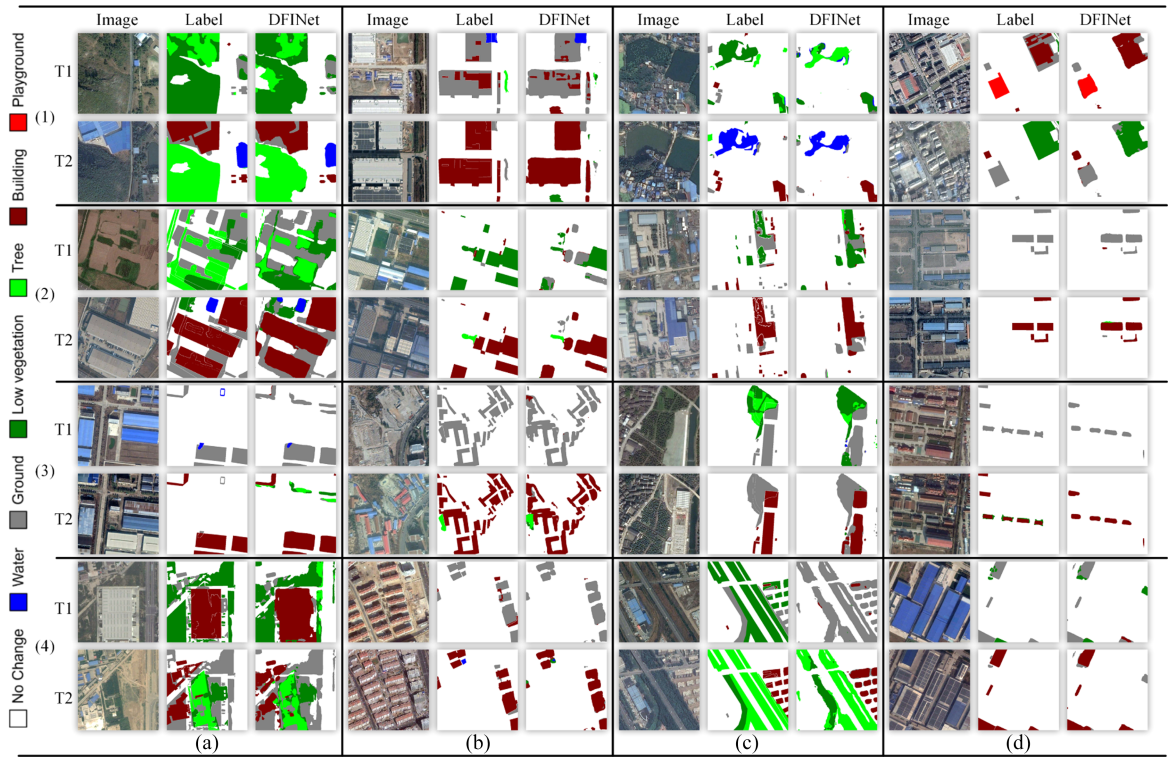

Fig. 9. Results obtained by DFINet on the SECOND dataset, where (a)–(d) represent diverse change types, changes to buildings, changes to vegetation, and others, respectively.

expressed using the following formula:

$$p_0 = \sum_{i=0}^{N} / \sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij} \tag{15}$$

$$p_e = \sum_{i=0}^{N} \left(\sum_{j=0}^{N} q_{ij} \times \sum_{i=0}^{N} q_{ij} \right) / \left(\sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij} \right)^2 \tag{16}$$

$$\text{Kappa} = (p_0 - p_e) / (1 - p_e) \tag{17}$$

$$\text{SeK} = e^{\text{IoU}_c - 1} \times \text{Kappa}. \tag{18}$$

Based on the calculation formula for SeK, it can be observed that the more unbalanced the confusion matrix is, the higher the SeK value will be, which, in turn, results in a lower kappa value. This is a useful characteristic, as it helps penalize models with a strong bias.

C. Loss Function

We use the binary cross-entropy loss function to compute the loss for the BCD results of the CD branch and utilize the categorical cross-entropy loss function to compute the loss for the final SCD results, SCD1 and SCD2. Thereby optimizing the model parameters

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \frac{1}{2} \left(\mathcal{L}_{\text{Seg}_1} + \mathcal{L}_{\text{Seg}_2} \right) \tag{19}$$

$$\mathcal{L}_{\text{BCE}} = \frac{1}{N} \sum_{i} - \left[y_i \times \log\left(p_i\right) + \left(1 - y_i\right) \times \log\left(1 - p_i\right) \right] \tag{20}$$

$$\mathcal{L}_{\text{Seg}} = -\frac{1}{N} \sum_{i} \sum_{c=1}^{M} \log\left(p_{ic}\right) \tag{21}$$

In $\mathcal{L}_{\text{BCE}}$, N represents the number of samples, y_i represents the binary classification label for CD, where the change class is represented as 1 and the unchanged class as 0, and p_i represents the predicted value. In $\mathcal{L}_{\text{Seg}}$, M represents the number of classes, for example, in the SECOND dataset, there are a total of seven classes including the unchanged class, y_{ic} represents the label for SCD, and p_{ic} represents the predicted value for SCD.

D. DFINet Results

Fig. 9 illustrates the qualitative results of the DFINet model based on the SECOND dataset. In a (2) of Fig. 9, even in areas with low interclass differences, rich land cover types, and complex terrain, the model is capable of recognizing changes in low vegetation, trees, and bare soil. In a (4), it can be observed that even when the road appears similar to bare soil due to construction activities, the model did not incorrectly identify this as a false change. Section b demonstrates the accurate identification of changes from other land cover types to buildings. In Section c, the results show the recognition of changes from other land cover types to low vegetation or trees. The identification of trees is notably accurate; however, there are some limitations in

detecting changes in low vegetation within highly intricate areas, occasionally resulting in false positives. Overall, whether for area a (rich in changes), b (changes to buildings), c (changes to vegetation), or other types, DFINet performed well in identifying changes.

V. Discussion and Analysis

A. Comparison

To comprehensively evaluate the performance of the proposed DFINet model, we will compare it with several state-of-the-art methods through SCD experiments. The compared methods include the following.

1) *FC-EF, FC-EF-Cat, FC-EF-Diff [16]:* FC-EF is a single-branch encoder–decoder model based on the U-Net architecture proposed for CD. Cat and Diff are twin-branch CD model architectures evolved from FC-EF.
2) *SNUNet [42]:* SNUNet integrates Siamese networks, U-Net++, channel attention mechanisms, and dense skip connections to reduce the uncertainty of edge pixels in the changing area.
3) *HRSCD4 [41]:* This is a specific model architecture designed for SCD. It utilizes a triple-encoding branch structure that incorporates residual modules and an encoder–decoder architecture. This architecture is tailored to the task of SCD, where the goal is to detect and understand semantic changes in images over time. The use of residual modules and encoder–decoder structures helps improve the model's performance in capturing and analyzing these changes.
4) *FCCDN [39]:* An optimized feature constraint CD network is proposed based on a dual encoder–decoder architecture. It utilizes a nonlocal feature pyramid network to extract and fuse multiscale features while introducing a densely connected feature fusion module to enhance robustness.
5) *SCDNet [11]:* SCDNet is based on a Siamese U-Net architecture, which consists of two encoders and two decoders with shared weights, aiming to address the SCD task of large-scale remote sensing datasets in an end-to-end manner.
6) *SSCDL and BiSRNet [3]:* SSCDL and BiSRNet are two models used for CD tasks. SSCDL employs a Siamese CNN encoder to extract semantic information. These semantic features are then utilized in the CD decoder to capture differences between the images effectively. BiSRNet builds upon the SSCDL model by introducing global self-attention (SR) and cross-temporal self-attention (Cot-SR) modules. These modules enhance the exchange of information between the temporal and CD branches, improving the model's performance in handling temporal and change-related features.
7) *SCanNet [55]:* SCanNet is a semantic change transformer (SCanFormer) specifically engineered to model the "from–to" semantic transitions between bitemporal remote sensing images. It utilizes a variant of the cross-shaped window transformer, optimizing its ability to capture the "semantic change" dependencies within bitemporal remote sensing images.

Among the compared methods mentioned above, 1) and 2) were originally not designed for SCD tasks, so we added semantic segmentation heads to these models at the network's end to meet the experimental requirements [3].

TABLE I
Quantitative Results on the SECOND Dataset

Method	Accuracy			Params(Mb)
	mIoU(%)	OA(%)	SeK(%)	
FC-EF-Cat	67.31	85.71	10.75	2.74
FC-EF-Diff	68.83	86.34	12.89	1.66
Snunet	66.29	85.44	9.31	10.20
HRSCD4	68.33	87.97	13.46	13.71
SCDNet	70.78	87.93	17.23	**39.62**
FCCDN	69.25	85.91	16.02	24.20
SSCDL	70.76	88.25	17.17	23.31
BiSRNet	71.67	88.61	18.67	23.38
SCanNet	72.26	88.58	20.07	27.90
DFINet(Ours)	**72.61**	**89.11**	**20.12**	23.85

The quantitative results are presented in Table I. The highest scores are highlighted in bold. Compared to the latest BiSRNet and SCanNet networks, DFINet achieved a 1.45% and 0.05% increase in SeK values, as well as a 0.94% and 0.35% improvement in mIoU. Specifically, FC-EF-Cat and Snunet concatenate all feature information at the channel level, which can lead to the extraction of valuable features being overwhelmed by redundant information. This approach lacks effective information selection during feature fusion. FC-EF-Diff, by using Sub for feature extraction, effectively captures the differences between dual-temporal features. This method has improved SeK scores by 2.14% compared to FC-EF-Cat and 3.58% compared to Snunet. However, solely using Sub is insufficient to capture change information and suppress false changes. Compared to FC-EF-Diff, DFINet has shown a 3.78% increase in mIoU and a 7.23% improvement in SeK.

As shown in Fig. 10, in T1 (1), dedicated SCD models all achieve good results. They are capable of identifying change areas well even in complex terrains with intertwined land cover types such as forests and low vegetation. However, only DFINet provides more complete and accurate results. In the more challenging central region, influenced by the uncertainty of land cover change properties, all the compared models exhibit the phenomenon of misclassifying trees as low vegetation. However, DFINet, with the TDFE module, effectively suppresses this issue. It makes better use of temporal information, reducing pseudochange interference caused by seasonal factors. In T1 (2), DFINet was not affected by the vegetation near the water body, successfully extracting the change areas. In T2, despite the diverse types of changes, DFINet still effectively identified the changes in the sports field.

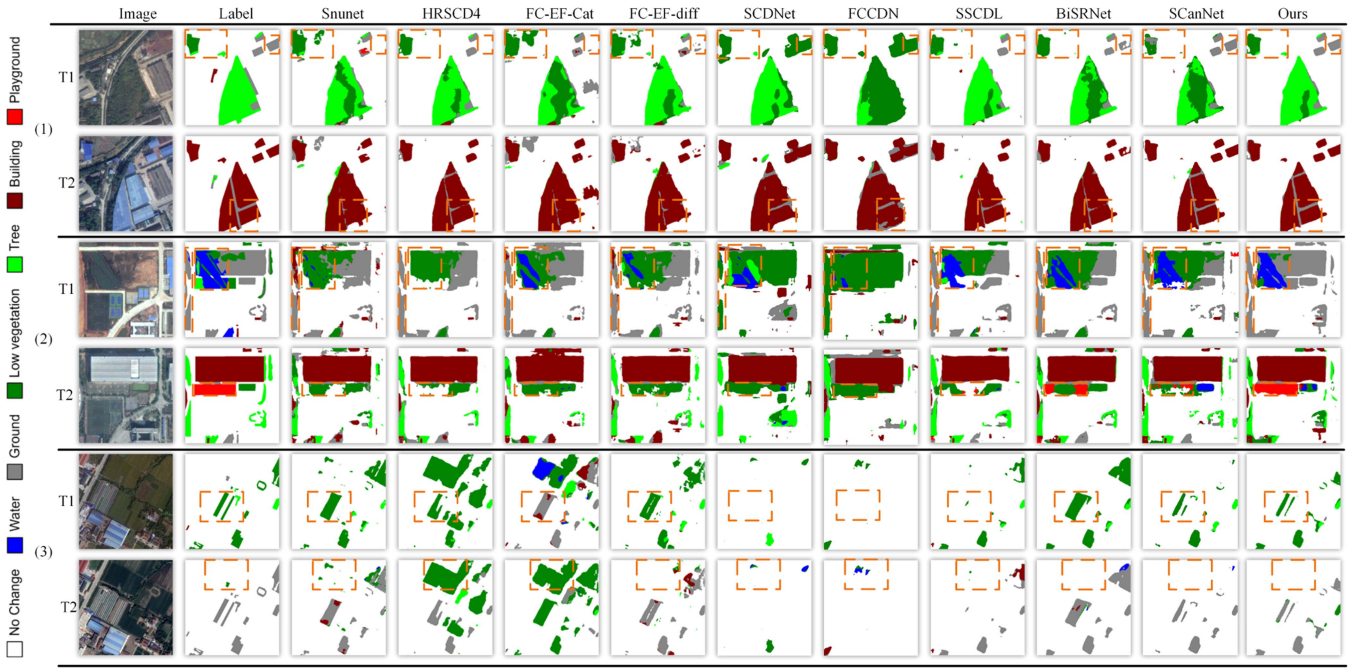

Fig. 10. Comparative experiments provide result examples for different methods. The major differences are highlighted in orange rectangles.

TABLE II
ABLATION RESULTS ON THE SECOND DATASET

Module	mIoU(%)	SeK(%)	OA(%)
Base	71.75	18.65	88.08
Base+FSIA	71.81	18.87	88.50
Base+TDFE+FSIA	72.28	19.45	88.48
Base+TDFE+FSIA+IAM	**72.61**	**20.12**	**89.11**

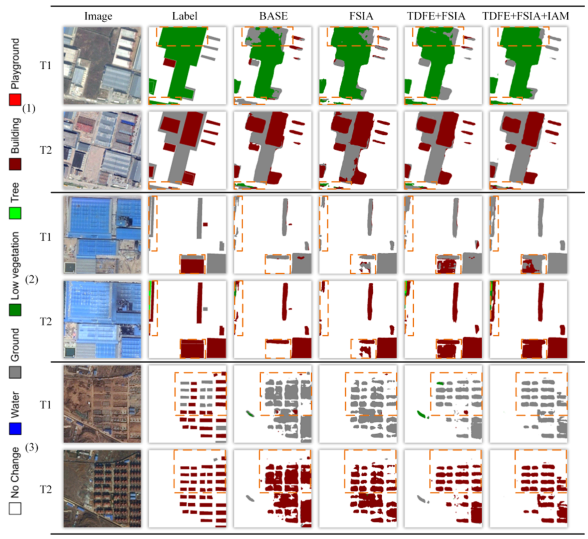

Fig. 11. Qualitative results of the ablation experiments for FSIA, TDFE, and IAM. The major differences are highlighted in orange rectangles.

B. Ablation

To demonstrate the effectiveness of TDFE, FSIA, and IAM, we conducted ablation experiments on the SECOND dataset. As shown in Table II, FSIA, TDFE, and IAM led to improvements of 0.22%, 0.58%, and 0.67% in SeK, respectively. The visual comparison results are shown in Fig. 11, from left to right, showing the results with the FSIA, TDFE, and IAM modules added. From (1) of Fig. 11, it can be observed that with the inclusion of the modules, the model becomes more accurate in detecting changes in low vegetation and ensuring the integrity of the results. For environments with high detection requirements, where it may not be very easy to clearly distinguish between buildings and bare soil, as seen in (3), the inclusion of TDFE helps the model differentiate between different land cover types, and FSIA and IAM reduce the interference of contextual noise in the detection results [26].

To provide further evidence of the effectiveness of the TDFE and FSIA modules we introduced, we conducted separate comparisons between TDFE, FSIA, and traditional feature extraction and fusion methods. The respective quantitative results can be found in Table III. The utilization of the TDFE module outperformed the sole use of Sub in feature extraction, yielding improvements of 0.8% in mIoU and 1.25% in SeK, as demonstrated in the qualitative results in Fig. 12. From Table IV, it can be observed that compared to traditional feature fusion methods, the FSIA module exhibits advantages in both feature extraction results and computational speed. Utilizing the FSIA module for multiscale feature fusion is not only faster but also yields better results than using traditional channel concatenation.

TABLE III
COMPARATIVE RESULTS OF DIFFERENT FEATURE FUSION METHODS ON THE SECOND DATASET

Method	Accuracy		
	mIoU(%)	SeK(%)	OA(%)
Sub	71.81	18.87	88.50
TDFE	**72.61**	**20.12**	**89.11**

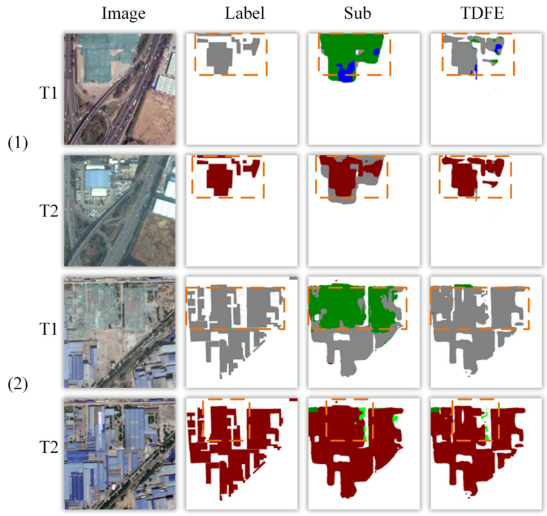

Fig. 12. Results of ablation experiments for the TDFE module.

TABLE IV
COMPARATIVE RESULTS OF DIFFERENT FEATURE FUSION METHODS ON THE SECOND DATASET

Fuse Method	Accuracy		Speed		
	mIoU(%)	SeK(%)	Params(M)	FLOPs(G)	Per-Epoch/min
Cat	72.27	19.78	**29.59**	**232.55**	**16.67**
FSIA	**72.61**	**20.12**	24.24	208.96	10.22

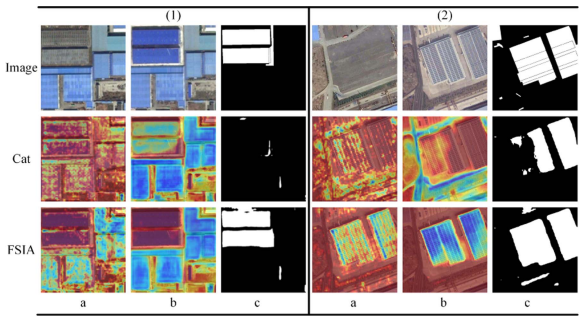

Fig. 13. Visualization results of heatmaps under different fusion methods.

Using gradient-weighted class activation mapping (Grad-Cam) [43], we also generated heatmaps for the model under different feature fusion methods, as shown in Fig. 13. This study's approach exhibits a relatively stable performance when handling features of the same type. The FSIA module, through feature selection and enhancement, can mitigate temporal differences arising from external environmental conditions and inherent imaging features, thus reducing false detections.

VI. CONCLUSION

In this article, we have analyzed the challenges in SCD, including the difficulty in evaluating interclass differences, the challenge of comprehensively perceiving multilevel image features and the complexity of interacting with multitemporal features. Based on these challenges, we have proposed an SCD framework called DFINet.

In this framework, we first introduce temporal information into multiple hierarchical feature layers using the TDFE module to enhance the extraction of multitemporal and multilevel semantic features, thereby improving the evaluation of interclass differences and perceptual capabilities. Second, we proposed the FSIA and IAM modules to interact, filter, and redistribute features from multiple temporal phases and hierarchical levels, thereby strengthening the transfer and integration capabilities among features from different time sequences. Finally, we showed that the experiments conducted on three public datasets yielded promising results, validating the effectiveness of our proposed framework.

This advances the practicality of remote sensing SCD and provides valuable insights for feature perception and fusion in deep learning methods. In future work, a significant improvement direction is to further reduce the computational complexity of DFINet and explore the utilization of unlabeled data to enhance the model's detection performance.

REFERENCES

[1] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111.
[2] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, Jun. 2012.
[3] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620014.
[4] A. Frick and S. Tervooren, "A framework for the long-term monitoring of urban green volume based on multi-temporal and multi-sensoral remote sensing data," *J. Geovisualization Spatial Anal.*, vol. 3, no. 1, pp. 1–11, 2019.
[5] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Review article digital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.
[6] A. A. Abuelgasim, W. D. Ross, S. Gopal, and C. E. Woodcock, "Change detection using adaptive fuzzy neural networks," *Remote Sens. Environ.*, vol. 70, no. 2, pp. 208–223, 1999.
[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.
[8] S. Park, N. Y. Yu, and H. -N. Lee, "An information-theoretic study for joint sparsity pattern recovery with different sensing matrices," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5559–5571, Sep. 2017.
[9] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Inf. Theory*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.
[10] Y. Deng et al., "Feature-guided multitask change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9667–9679, 2022.

[11] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "SCDNET: A novel convolutional network for semantic change detection in high-resolution optical remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102465.

[12] Y. Niu, H. Guo, J. Lu, L. Ding, and D. Yu, "SMNet: Symmetric multi-task network for semantic change detection in remote sensing images based on CNN and transformer," *Remote Sens.*, vol. 15, Feb. 2023, Art. no. 949.

[13] K. Yang et al., "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[14] M. Zhao et al., "Spatially and semantically enhanced Siamese network for semantic change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2563–2573, Jan. 2022.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[16] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[17] 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, 2019, Art. no. 1382.

[18] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support: 4th Int. Workshop*, 2018, pp. 3–11.

[19] J. Xue et al., "Multi-feature enhanced building change detection based on semantic information guidance," *Remote Sens.*, vol. 13, no. 20, 2021, Art. no. 4171.

[20] M. Lin, G. Yang, and H. Zhang, "Transition is a process: Pair-to-video change detection on networks for very high-resolution remote sensing images," *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2023.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4282–4291.

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

[24] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1688.

[25] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[26] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired from PID controller," Jun. 2022.

[27] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[28] M. Jiang, X. Zhang, Y. Sun, W. Feng, Q. Gan, and Y. Ruan, "AFS-Net: Attention-guided full scale feature aggregation network for high-resolution remote sensing image change detection," *GIScience Remote Sens.*, vol. 59, no. 1, pp. 1882–1900, 2022.

[29] X. Xiang, D. Tian, N. Lv, and Q. Yan, "FCDNet: A change detection network based on full-scale skip connections and coordinate attention," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[30] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design.," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[33] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc.IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12416–12425.

[34] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 12, no. 4, Feb. 2020, Art. no. 701.

[35] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.

[36] X. Zhao et al., "M2SNet: Multi-scale in multi-scale subtraction network for medical image segmentation," Mar. 2023, *arXiv:2303.10894*.

[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[38] K. Yang et al., "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

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

[40] N. K. Sinha and M. P. Griscik, "A stochastic approximation method," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-1, no. 4, pp. 338–344, Oct. 1971.

[41] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large scale semantic change detection," *Comput. Vis. Image Understand*, vol. 187, Oct. 2019, Art. no. 102783.

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

[43] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, pp. 336–359, Feb. 2020.

[44] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[45] 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. Photogrammetry Remote Sens.*, vol. 175, pp. 247–267, 2021.

[46] S. Zhu, Y. Song, Y. Zhang, and Y. Zhang, "ECFNet: A Siamese network with fewer FPs and fewer FNs for change detection of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[47] J. Mo, S. Seong, J. Oh, and J. Choi, "SAUNet3 CD: A Siamese-attentive UNet3 for change detection in remote sensing images," *IEEE Access*, vol. 10, pp. 101434–101444, 2022.

[48] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 4229–4238.

[49] Z. Lv, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 44–63, Mar. 2022.

[50] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[51] Y. Ye, M. Wang, L. Zhou, G. Lei, J. Fan, and Y. Qin, "Adjacent-level feature cross-fusion with 3-D CNN for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[52] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12325–12334.

[53] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[54] S. Hayet, S. E. Salah, A. G. Rezgui, B. N. El Islam, and S. Ait-Aoudia, "What is a remote sensing change detection technique? Towards a conceptual framework," *Int. J. Remote Sens.*, vol. 41, pp. 1788–1812, 2019.

[55] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610814.