Cross-Difference Semantic Consistency Network for Semantic Change Detection

Qi Wang, *Senior Member, IEEE*, Wei Jing, Kaichen Chi, and Yuan Yuan, *Senior Member, IEEE*

***Abstract*—The objective of Semantic Change Detection (SCD) is to discern intricate changes in land cover while simultaneously identifying their semantic categories. Prior research has shown that using multiple independent branches for the distinct tasks of change localization and semantic recognition is a reliable approach to solving the SCD problem. Nevertheless, conventional SCD architectures rely heavily on a high degree of consistency within the bi-temporal feature space when modeling difference features, inevitably resulting in false positives or missed alerts within change areas. In this paper, we introduce a SCD framework called the Cross-Differential Semantic Consistency (CdSC) network. CdSC is designed to mine deep discrepancies in bi-temporal instance features while preserving their semantic consistency. Specifically, the 3D-Cross-Difference module, incorporating 3D convolutions, explores the interaction of cross-temporal features, revealing inherent differences among various land features. Simultaneously, deep semantic representations are further utilized to enhance the local correlation of difference information, thereby improving the model's discriminative capabilities within change regions. Incorporating principles from contrastive learning, a Semantic Co-Alignment loss is introduced to increase intra-class consistency and inter-class distinctiveness of dual-temporal semantic features, thereby addressing the challenges posed by semantic disparities. Extensive experiments on two SCD datasets demonstrate that CdSC outperforms other state-of-the-art SCD methods significantly in both qualitative and quantitative evaluations. The code and dataset are available at https://github.com/weiAI1996/CdSC.**

***Index Terms*—Remote sensing image, Semantic change detection, Deep learning, Cross-Difference, Semantic consistency.**

I. Introduction

In recent decades, the field of change detection (CD) in remote sensing imagery has witnessed substantial advancements, bolstered by the rapid expansion of remote sensing big data [1]–[3]. It has greatly facilitated tasks such as disaster assessment [4], agricultural surveys [5], and urban planning [6], gradually assuming a pivotal role in the monitoring of surface environments and human activities. However, in real-world scenarios, people are not only concerned with where changes occur but also with the categories of land cover changes. Motivated by this consideration, researchers have undertaken a series of studies in the field of semantic change detection (SCD) [7]–[11].

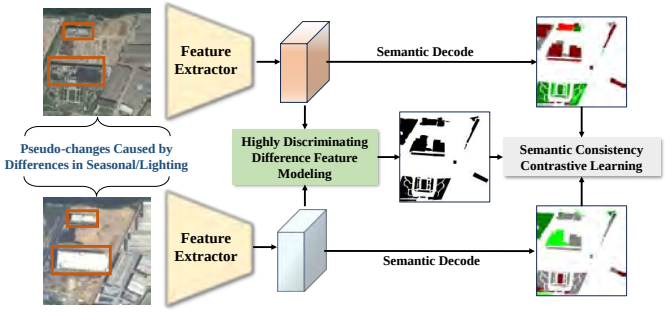

Fig. 1. The schematic illustration of research motivation. The proposed framework alleviates pseudo-changes caused by extrinsic factors by modeling highly discriminative difference features. It enhances the consistency between change map and semantic maps through contrastive learning, ensuring a more robust representation of variations attributable to genuine changes in the observed scene.

Differing from Binary Change Detection (BCD) tasks that solely focus on locating change regions [12]–[14], SCD tasks aim to not only extract information about the location of changes but also map the 'from-to' change directions within those regions [15]–[17]. An intuitive solution to the SCD task is to treat it as a multi-class segmentation task based on BCD, as in HRSCD.str1 designed by Daudt [8]. However, this straightforward approach comes with significant drawbacks: (1) As the number of land cover types increases, the multitude of permutations and combinations for 'from-to' mappings escalates significantly, imposing rigorous demands on the model's parameters and the quantity of training data required to construct a versatile multi-class segmentation network. (2) In the absence of guided prior knowledge, deep models often learn fixed patterns from the training data rather than the deep semantics of dual-temporal images, making them susceptible to local optima and resulting in severe overfitting.

Constrained by the aforementioned problems, the SCD solution, which incorporates BCD before semantic segmentation, has emerged as a viable approach [8]–[11], [15]. These algorithms perceive the SCD task as a multi-task learning process. Initially, they employ deep feature extraction from dual-temporal images to capture difference information and generate binary change maps. Subsequently, through a masking operation, the model is compelled to focus on change regions. In this process, the extraction of discriminative difference features serves as the foundation for achieving high-precision change detection. The strategy of modeling through

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902. (*Corresponding author: Qi Wang.*)

Qi Wang, Kaichen Chi and Yuan Yuan are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com, y.yuan1.ieee@gmail.com).

Wei Jing is with the National Elite Institute of Engineering and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: wei_adam@mail.nwpu.edu.cn).

convolution layers after point-to-point differencing and cascading has been widely applied in deep learning-based CD methods [18], [19]. Recent studies have focused on enhancing the extraction of differential features through strategies such as attention mechanisms and multi-scale feature fusion [20], [21]. Despite researchers extracting and enhancing difference features from various perspectives, these methods still do not deviate from feature cascading and implicit modeling through 2D convolution. On the other hand, semantic consistency expression constitutes another critical aspect of the SCD task. Most existing studies concentrate on the semantic representation of land objects without considering the strong correlation between change information and semantic expression [11], [15].

To sum up, existing multi-task semantic change detectors still encounter the following challenges: (1) Lack of channel relationship modeling for bi-temporal features: Researchers typically adopt the 2D convolution to implicitly extract the difference information from dual-temporal spliced features, without the guidance of an explicit modeling strategy. (2) Inadequate semantic-change consistency: In the SCD task, there exists a strict consistency between the semantic information of dual-temporal objects and their change information. This prior knowledge should be incorporated into the model's constraint functions.

To enhance the difference feature extraction capability and semantic representation prowess of the model, this paper introduces a framework for SCD called the Cross-Difference Semantic Consistency (CdSC) network, and the overall architecture is shown in Fig. 1. Specifically, to fully exploit the difference information in bi-temporal features, we design a difference module that integrates cross-fusion operations and cascaded 3D convolutions. The proposed 3D-Cross-Difference enables simultaneous modeling of spatiotemporal variations in bi-temporal land cover features, both in the spatial and channel dimensions. Subsequently, an attention-based Semantic Context Enhancement module is proposed to enhance the consistency between changes and semantic representations. Furthermore, we formulate a Semantic Co-Alignment loss function through contrastive learning to rigorously enforce the alignment of change and semantic information. The key contributions of this paper can be summarized as follows:

1. For robust modeling of difference information, we develop a 3D-Cross-Difference module to investigate the spatiotemporal discrepancies of bi-temporal geospatial data simultaneously from both spatial and channel dimensions.

2. We propose a Semantic Context Enhancement module, designed to bridge the gap between semantic representations and change information, thereby reducing semantic ambivalence within difference features.

3. To mitigate the risk associated with semantic disparities, we formulate a Semantic Co-Alignment loss function, which supervises the consistency of the semantic representation between BCD and SCD maps.

4. In order to explore deep differences in bi-temporal images and maintain change-semantic consistency, we construct the Cross-Difference Semantic Consistency (CdSC) network for SCD. Extensive experiments have demonstrated that CdSC significantly outperforms other SCD methods in improving the extraction results of both BCD and SCD.

II. Related work

A. Deep Learning-Based Binary Change Detection

The emergence of numerous deep learning-based CD models has been facilitated by the availability of a large volume of annotated remote sensing data, thanks to the advancements in remote sensing big data [22]–[28]. Compared to traditional CD algorithms, deep learning possesses a more robust feature representation capability, capable of capturing both low-level details and high-level semantics in images simultaneously, thus adapting to a variety of complex scenarios [29]–[36]. Typically, single-branch and dual-branch structures are common deep learning-based CD frameworks.

Single-branch structures fuse bi-temporal images using strategies such as differencing or concatenation before feature extraction, treating CD as a segmentation problem [37], [38]. Sun et al. integrated an extended ConvLSTM structure into the U-Net framework to achieve end-to-end CD and further utilized Atrous convolution to mine multi-scale spatial information [37]. Lin et al. framed CD as a video understanding task, explicitly considering the spatiotemporal coupling problem in the encoder, and proposed the P2V-CD model to decouple the spatiotemporal dimensions of bitemporal images [38]. Dual-branch structures, on the other hand, represent a late fusion strategy, with siamese networks being a representative model, which perform CD by extracting discriminative features from image pairs respectively [2], [14], [17]. Zhang et al. aimed to improve the boundary coherence and internal cohesion of objects in the resultant change maps by developing a deeply supervised difference discrimination network [14]. This network enhances the change map by fusing deep features from original inputs with difference features of bi-temporal images. Lei et al. addressed the challenge of identifying irrelevant changes. They utilized a siamese network to distinguish between foreground and background, thereby obtaining a discrepancy representation [2]. Additionally, they incorporated remote relationships to enhance the edge coherence and internal cohesion of the changing objects. Recently, transformer models have shown exceptional capabilities in the realm of computer vision, outperforming CNN-based methods in various tasks, including classification and detection [39], [40]. The core self-attention mechanism enables the model to capture the correlation between different regions of the image, thus modeling global dependencies [41]–[44]. Bandara et al. unified the hierarchical Transformer structure in siamese networks, effectively rendering the multi-scale remote details required for precise CD, achieving SOTA performance on multiple CD datasets [16].

Although researchers have made significant efforts and achieved satisfactory results in CD, previous studies have primarily focused on designing intricate data encoding and decoding methods. There has been comparatively less emphasis on the effective extraction of difference features, an aspect critical for enhancing the field.

B. Semantic Change Detection

Binary change detection, which solely identifies changed regions, frequently proves inadequate for meeting the demands of real-world applications. To extract the "from-to" mappings of change regions, Daudt et al. introduced an intuitive SCD method [8]. In this approach, they replace the binary segmentation used in BCD tasks with multi-class segmentation, where each "from-to" change type is treated as a separate category. However, this method places significant demands on model parameters and data volume due to the necessity of accommodating a large classification space.

To tackle this issue, researchers have designed the SCD framework, in which BCD and semantic segmentation collaborate [8]–[11], [15], [45]. The proposed framework features a dual encoder dedicated to semantic feature extraction, alongside an independent Fully Convolutional Network (FCN) [18] for extracting change features. It employs two separate convolutional decoders: one for semantic segmentation and the other for BCD. Within this framework, the classification map generated from semantic segmentation offers directional insights regarding changes, while the binary change map from BCD pinpoints the regions undergoing change. Subsequently, a simple post-processing step is employed to generate the semantic change map. Yang et al. introduced an asymmetric Siamese network designed to localize and identify semantic changes by harnessing features extracted from structurally diverse modules [9]. Zheng et al. incorporated a semantic-aware encoder into their design to model the causal relationships of semantic changes [15]. This encoder is solely used for learning semantic representations. Subsequently, a Transformer module is employed to learn change representations from these semantic features, with the addition of regularization to prevent overfitting. Ding et al. integrated semantic and temporal features into a deep CD unit to enhance both intra-temporal and inter-temporal semantic consistency [10]. Furthermore, they proposed a new loss function specifically designed to improve the semantic coherence of the CD results.

The endeavors mentioned previously have significantly advanced the field of SCD. However, there is still room for improvement in current methods, especially regarding feature extraction and the representation of semantic consistency. Motivated by this, this paper explores the cross-fusion and difference extraction of bi-temporal features and introduces a 3D-Cross-Difference module to extract profound difference in object representations. Meanwhile, we construct Semantic Context Enhancement module and Semantic Co-Alignment loss to promote semantic consistency representation.

C. Contrastive Learning

Contrastive learning, a widely applied learning strategy in self-supervised learning, revolves around the core concept of contrasting positive and negative samples within a feature space [46], [47]. The aim is to learn feature representations of samples by aligning them as closely as possible with positive samples while making them as dissimilar as possible from negative samples. By discerning between similar and dissimilar samples, the model can effectively capture the intrinsic features of the samples.

In recent years, contrastive learning has seen increasingly widespread development and application within the field of computer vision [46], [48], [49]. In the field of image segmentation, Wang et al. introduced pixel-wise contrastive learning, leveraging pixel-to-pixel correspondences across images to learn a feature space [48]. They defined pixels of the same class as positive examples and those of different classes as negative examples, augmenting the cross-entropy loss with the Noise Contrastive Estimation (NCE) loss. Chaitanya et al. proposed a method of local contrastive loss, utilizing pseudo-labels from unlabeled images and limited semantic label information to learn pixel-level features beneficial for segmentation [49]. In the area of object detection, Xiong et al. enhanced target detection accuracy by applying the InfoNCE loss to all local learnings through increasing the depth of the decoder [50]. ContraGAN employed a conditional contrastive loss to examine the relationships between multiple images within the same batch and between data and classes, thereby improving the visual quality of generated images [51]. Moreover, contrastive learning has been widely applied to clustering learning and natural language processing [52], [53]. Zhang et al. applied contrastive learning to clustering, performing contrastive learning clustering through the joint optimization of a top-down clustering loss and a bottom-up instance contrastive loss [52]. Cheng et al. utilized contrastive learning to mitigate bias in generated text representations (e.g., gender bias, racial bias) by comparing original sentences with their antonyms to maximize mutual information between them, and comparing original sentences with biased words to minimize the mutual information [53].

In this paper, the proposed loss function identifies similar and dissimilar samples by analyzing change maps to extract regions of change and no change. It then seeks to minimize the cosine distance between similar samples to bring them as close as possible, while maximizing the cosine distance between dissimilar samples to keep them as far apart as possible.

III. METHODOLOGY

In this section, we will describe the Cross-Difference Semantic Consistency (CdSC) network, which detects semantic changes by utilizing 3D convolution to extract difference information and model semantic alignment.

A. Overall Framework of CdSC

Given a pair of bi-temporal images, I_1 and I_2, the task of SCD is to determine a mapping function F_{scd} that projects the image pair into semantic change maps as follows

$$F_{scd}\left(I_{1_{i,j}}, I_{2_{i,j}}\right) = \begin{cases} (0,0), & c_{1_{i,j}} = c_{2_{i,j}} \\ (c_{1_{i,j}}, c_{2_{i,j}}), & c_{1_{i,j}} \neq c_{2_{i,j}} \end{cases} \quad (1)$$

where ($c_{1_{i,j}}$ and $c_{2_{i,j}}$) denote the land cover classes in the bi-temporal images, respectively.

Fig. 2 illustrates the overall architecture of CdSC, which takes paired bi-temporal images as input and produces a binary change map along with two semantic segmentation

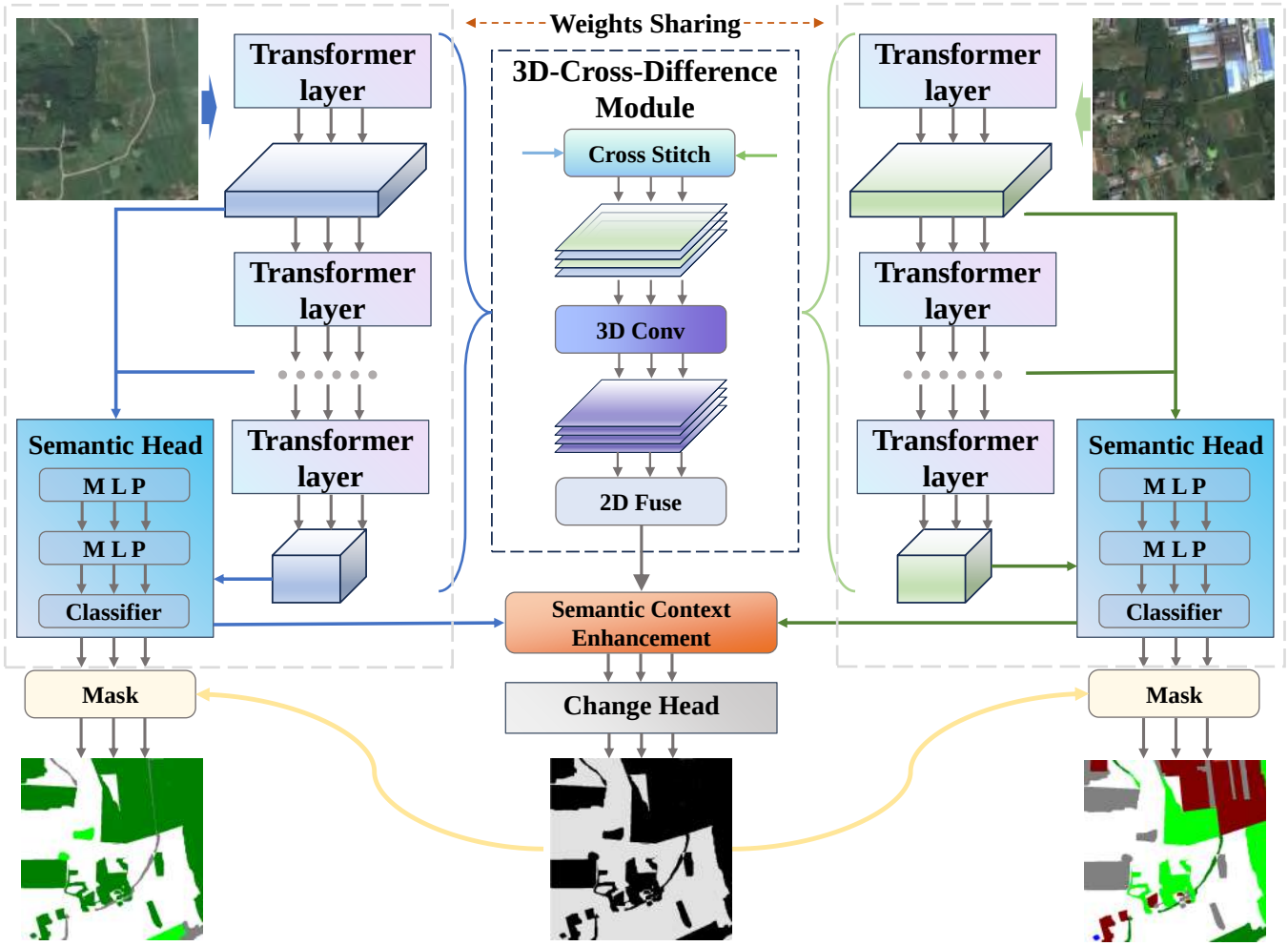

Fig. 2. Architecture of CdSC model. The encoder is the weight-sharing transformer, the 3D-Cross-Difference module is used to extract the bi-temporal difference features, and the Semantic Context Enhancement module is used to unite the semantic representations to further enhance the difference features.

maps delineating the changed regions as output. Specifically, CdSC employs a Siamese Vision Transformer network, which is adept at capturing both shallow details and deep semantic information from bi-temporal images. This network encapsulates the spatiotemporal features encoding the intrinsic information of these images.Using the proposed 3D-Cross-Difference module, we explicitly extract difference information at multiple levels. The computation is as follows:

$$X_{diff}^{i} = X_1^i \Theta X_2^i, \tag{2}$$

where Θ denotes 3D cross differencing, X_1^i represents the features from the first image at the i-th layer of the encoder, i ranges from 1 to 4. Simultaneously, spatiotemporal features from the backbone are fed into the semantic head to generate the initial panoramic segmentation maps. The Semantic Head is primarily constituted by a lightweight decoder composed of MLP (Multi-Layer Perceptron) layers, coupled with a series of upsampling and concatenation operations applied to features from multiple layers. This design circumvents the need for manually crafted and computationally demanding components commonly employed in alternative approaches. The computation is as follows:

$$P_s^t, X_s^t = \mathcal{H}_s\left(\{X^i\}\right), \tag{3}$$

where P_s^t are the panoramic segmentation maps, X_s^t are the semantic features from the semantic head, t represents the index of bi-temporal images, and $\mathcal{H}_s$ indicates the set of operations within the semantic head. In order to maintain semantic consistency, X_s is used to augment the difference information through the Semantic Context Enhancement (SCE) module:

$$\widetilde{X}_{diff} = \mathcal{F}_{sce}\left(X_s^1, X_{diff}, X_s^2\right), \tag{4}$$

Then, the enhanced difference information $\widetilde{X}_{diff}$ is fed to the change head $\mathcal{H}_c$ to generate binary change map as follows

$$P_{bcd} = \mathcal{H}_c\left(\widetilde{X}_{diff}\right), \tag{5}$$

where $\mathcal{H}_c$ is a binary classifier consisting of two layers of convolution as follows

$$\mathcal{H}_c\left(\tilde{X}_{diff}\right) = \mathcal{K}_{1\times1}\left(\delta\left(\mathcal{K}_{3\times3}\left(\tilde{X}_{diff}\right)\right)\right), \tag{6}$$

where $\mathcal{K}_{1\times1}$ denotes a 2D convolution with kernel size 1×1, δ refers to ReLu activation function. Finally, P_{bcd} is utilized to

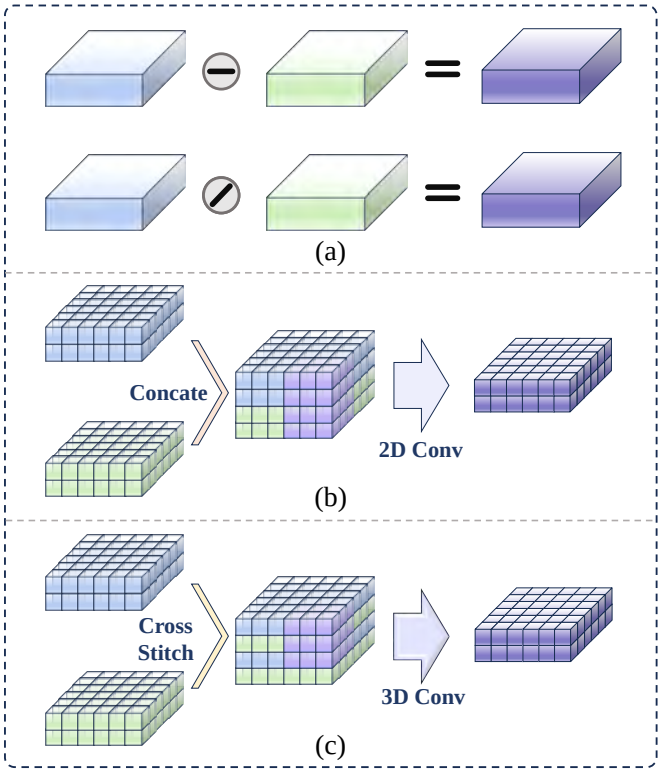

Fig. 3. Structures of (a) Subtracting/Dividing Difference, (b) 2D Convolution Difference and (c) 3D-Cross-Difference.

mask the panoramic segmentation maps P_s^t, thereby generating the semantic change maps as follows

$$P = P_s^t \odot P_{bcd}, \tag{7}$$

where $\odot$ denotes element-wise multiplication.

B. 3D-Cross-Difference Module

Extracting discriminative difference features from paired multi-temporal images is a crucial component of CD tasks [54]–[56]. The most straightforward approach, illustrated in Fig. 3(a), involves directly subtracting or dividing the bi-temporal features to generate difference information. The feature subtraction more directly reflects the absolute changes in the image, while feature ratios focus more on the relative changes, which are expressed as follows:

$$\begin{aligned} X_{diff} &= X_1 - X_2, \\ X_{diff} &= \frac{X_1}{X_2}. \end{aligned} \tag{8}$$

However, due to the prevalent occurrence of phenomena such as "spectral confusion" and "spectral variability" in remote sensing imagery, direct differencing strategies may result in missed detections of objects with similar spectra but different categories, while erroneously detecting objects with similar categories but significant spectral differences. As shown in Fig. 3(b), existing methods usually concatenate bi-temporal features along the channel dimension and depend on a sequence of 2D convolutional layers for extracting difference features. However, capturing the intrinsic differences between objects directly from concatenated features poses a challenge.

In response to the above challenges, we propose a 3D-Cross-Difference module to explicitly extract highly discriminative difference information from multi-temporal features. As shown in Fig. 3(c), the bi-temporal features are no longer directly concatenated but rather fused by channel-wise cross stitch. Specifically, each channel is traversed, and the feature maps corresponding to each channel are interleaved together to create a new feature tensor, represented as follows:

$$X_{cs} = \mathcal{F}_{cs}\left(X_1, X_2\right), \tag{9}$$

where $\mathcal{F}_{cs}$ denotes the operation of cross stitch and $X_{cs}$is the feature after stitching. To explicitly extract difference information, we utilize 3D convolution to process the cross-stitched features. Unlike 2D convolution, which operates directly on the entire channel dimension, 3D convolution extends the operation by applying a three-dimensional kernel that slides across both the spatial dimensions and the channel dimension. This progressive representation of bi-temporal difference information is illustrated as follows:

$$X_{diff} = \mathcal{K}_{3D}^{3\times3\times3}\left(X_{cs}\right), \tag{10}$$

where $\mathcal{K}_{3D}^{3\times3\times3}$ denotes a 3D convolution with kernel size $3 \times 3 \times 3$. To further model the intrinsic properties of difference information, the fusion of spatial-spectral difference features is undertaken to reduce the feature dimensionality, thereby lowering computational complexity. Subsequently, a 2D convolution is applied to X_{diff} as follows:

$$X_{diff} = \mathcal{K}_{3\times3}\left(X_{diff}\right), \tag{11}$$

C. Semantic Context Enhancement Module

To bridge the gap between semantic representations and change information, as well as to diminish semantic ambivalence in difference features, we draw inspiration from spatial attention mechanism [57] and propose the Semantic Context Enhancement module to further refine the difference features. As shown in Fig. 4, initially, a 3D-Cross-Difference module is employed to calculate the semantic difference between the bi-temporal semantic features, X_s^1 and X_s^2, obtained from the semantic head, as follows:

$$X_{sd} = X_s^1 \Theta X_s^2, \tag{12}$$

where X_{sd} is semantic difference feature. Next, we utilize channel-wise pooling and gated convolutional structures to acquire the spatial attention maps for semantic difference features:

$$\begin{aligned} M_{max} &= \mathcal{F}_{maxpool}(X_{sd}), \\ M_{avg} &= \mathcal{F}_{avgpool}(X_{sd}), \\ M_{attention} &= \mathcal{G}(M_{max} \| M_{avg}), \end{aligned} \tag{13}$$

where $\mathcal{G}$ denotes gated convolution operation and is calculated as follows:

$$\mathcal{G}\left(X_{sd}\right) = \sigma\left(\mathcal{K}_{7\times7}\left(X_{sd}\right)\right), \tag{14}$$

where σ refers to sigmoid activation function. Through the aforementioned procedures, semantic change regions will be

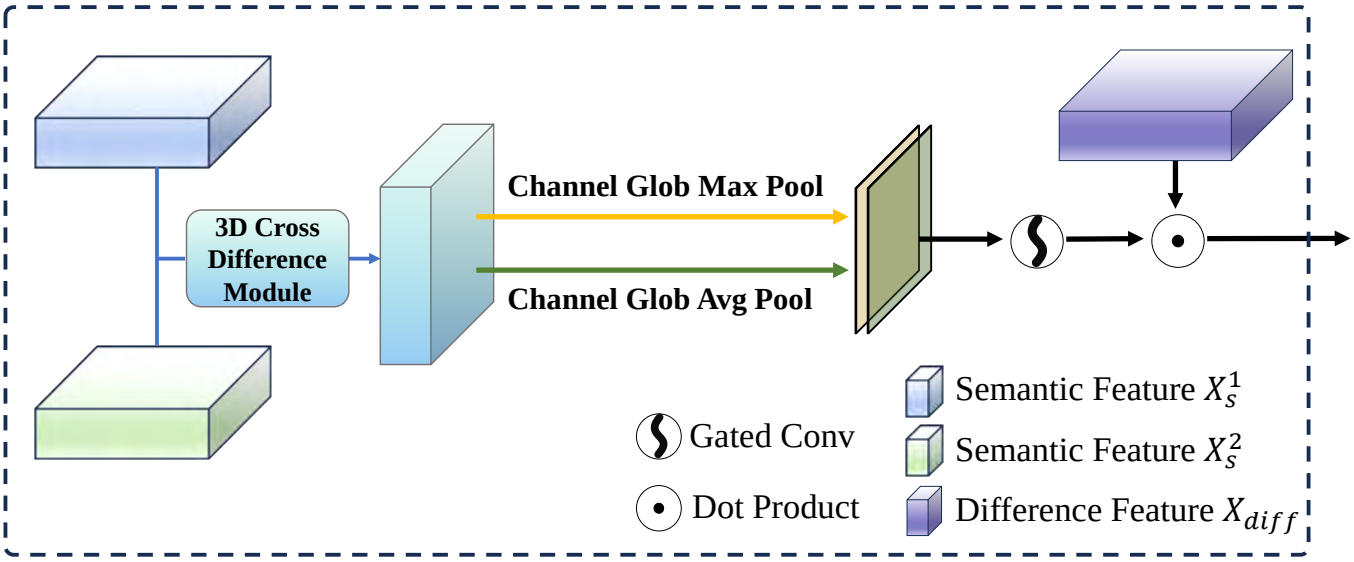

Fig. 4. Structure of the proposed SCE module. Deep difference features are extracted via a 3D cross-differencing module and are augmented by applying the spatial attention mechanism.

close to 1, while similar regions will tend towards 0. Finally, the utilization of the dot product operation in the model plays a crucial role in accentuating the changed regions within the difference feature space, concurrently diminishing the prominence of unchanged areas.

D. Loss Functions

To train the CdSC model, three distinct loss functions were employed: the semantic segmentation loss $\mathcal{L}_s$, the binary change loss $\mathcal{L}_c$, and the introduced Semantic Co-Alignment $\mathcal{L}_{sca}$.

$\mathcal{L}_s$ is the cross-entropy loss between the bi-temporal semantic segmentation maps and the ground truth, expressed as follows:

$$\mathcal{L}_s = -\sum_{i=1}^{M} y_i \log\left(p_i\right), \tag{15}$$

where M denotes the total number of categories, i serves as the index for each category, y_i represents the probability of the true class corresponding to category i, and p_i indicates the probability of the predicted class being categorized as i.

$\mathcal{L}_c$ is the binary cross-entropy loss between the predicted binary change map and the truth change map as follows:

$$\mathcal{L}_c = -y \log\left(p\right) - \left(1-y\right) \log\left(1-p\right), \tag{16}$$

where y represents the truth change map, and p represents the predicted binary change map.

In the context of SCD tasks, it is commonly assumed that unchanged regions will maintain identical semantic meanings, whereas changed regions will exhibit different semantic categories. Building upon this foundation and drawing from the principles of contrastive learning, we introduce the Semantic Co-Alignment (SCA) loss function. This function calculates the logical associations between predicted bi-temporal semantic maps, which can be expressed as

$$\begin{aligned}\mathcal{L}_{sca} = \alpha \cdot \cos\left(p_1 \odot y, p_2 \odot y\right) - \\ \left(1-\alpha\right) \cdot \cos\left(p_1 \odot \left(1-y\right), p_2 \odot \left(1-y\right)\right),\end{aligned} \tag{17}$$

where p_1 and p_2 are the predicted bi-temporal semantic segmentation maps, y represents the truth change map, and α is the balancing factor, which is empirically set to 0.8. To sum up, the overall loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_c + \mathcal{L}_{sca}. \tag{18}$$

IV. EXPERIMENTS

In this section, to assess the efficacy of CdSC, a series of experiments were conducted using the SECOND and JL1 datasets.

A. Dataset Description

1) The SECOND dataset [9], widely used in SCD research, comprises 4,662 pairs of aerial images, with 2,968 pairs in the training set and 1,694 pairs in the test set. Each image has dimensions of 512 × 512 pixels and consists of RGB channels, with spatial resolutions ranging from 0.5 to 3 meters per pixel. The labels annotate seven categories: Nonvegetated Ground (N.v.g. surface for short), Trees, Low vegetation, Water, Building, Playground, and the Non-change class. Notably, 61.2% of the pixels correspond to the Non-change class, 29.9% represent N.v.g. surface, while the remaining categories constitute less than 5% each. Due to the extreme class imbalance, there is an increased demand for models to exhibit strong generalization capabilities.

2) JL1 is a competition dataset used for cropland type change detection. The images are sourced from the Jilin-1 remote sensing satellite and comprise 6,000 images, with

4,000 allocated for training and 2,000 for testing. Each image is 256 × 256 pixels in dimension and consists of RGB channels, with a spatial resolution of 0.75 meters per pixel. In contrast to the annotation approach of the SECOND dataset, the JL1 dataset provides 9 categories of "from-to" labels for the bi-temporal images, indicating the type of change between two time points. These categories include cropland-road, cropland-forest/grassland, cropland-building, cropland-other, road-cropland, forest/grassland-cropland, building-cropland, other-cropland, and unchanged regions. Prior to training, we converted the labels to the same organizational format as the SECOND dataset.

B. Experiment Setup

The experiments were carried out on a server equipped with an NVIDIA GTX 3090 GPU, boasting 24GB of VRAM, and running the Ubuntu 18.04 operating system. To maintain experimental fairness, all models were trained for 50 epochs using publicly available code. The training employed the AdamW optimizer algorithm [58]. The learning rate was dynamically adjusted using a linear decay schedule, starting from an initial rate of 1e-4. No additional training techniques, other than those mentioned, were used during the training process.

C. Evaluation Metrics

The widely adopted metrics for quantitative evaluation of SCD include overall accuracy (OA), mean intersection over union (mIoU) [59], separated kappa (SeK) [9] coefficient, and F1-score (F1).

OA is calculated as the percentage of correctly predicted pixels out of all pixels as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}+\mathrm{TN}}, \tag{19}$$

where TP, TN, FP, and FN refer to the true positive, true negative, false positive, and false negative pixels , respectively.

mIoU is the average of the IoU (Intersection over Union) values for changed and unchanged regions, where IoU is defined as the ratio of the intersection to the union between the prediction and the ground truth. The calculation is as follows:

$$\begin{aligned} \mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}}, \\ \mathrm{mIoU} &= (\mathrm{IoU}_0+\mathrm{IoU}_1)\times 0.5, \end{aligned} \tag{20}$$

where IoU_0 and IoU_1 refer to the IoU of unchanged and changed regions, respectively.

SeK [9] is a metric capable of mitigating the impact of label imbalance by adaptively correcting output bias, calculated as follows:

$$\mathrm{SeK} = e^{(\mathrm{IOU}_1-1)}\cdot(\hat{\rho}-\hat{\eta})/(1-\hat{\eta}), \tag{21}$$

with

$$\begin{aligned} \hat{\rho} &= \textstyle\sum_{i=2}^{C} q_{ii} / \left(\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij} - q_{00}\right) \\ \hat{\eta} &= \textstyle\sum_{j=1}^{C} (\hat{q}_{j+}\cdot\hat{q}_{+j}) / \left(\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij} - q_{00}\right)^2, \end{aligned} \tag{22}$$

TABLE I
QUANTITATIVE RESULTS ON THE SECOND. THE BEST RESULTS ARE MARKED IN BOLD, THE 2ND-BEST IS MARKED WITH UNDERLINE.

Method	Evaluation Metrics			
	OA (%)	mIoU(%)	SeK(%)	F1(%)
HRSCD.str3	84.64	63.34	8.27	48.54
HRSCD.str4	85.30	69.47	15.48	55.61
SSESN	87.68	71.26	20.45	61.58
Bi-SRNet	<u>89.37</u>	<u>72.65</u>	<u>21.91</u>	<u>62.01</u>
CdSC	**90.07**	**73.32**	**23.52**	**63.70**

TABLE II
QUANTITATIVE RESULTS ON THE JL1. THE BEST RESULTS ARE MARKED IN BOLD, THE 2ND-BEST IS MARKED WITH UNDERLINE.

Method	Evaluation Metrics			
	OA (%)	mIoU(%)	SeK(%)	F1(%)
HRSCD.str3	77.19	64.86	10.54	51.74
HRSCD.str4	77.08	63.87	10.60	51.94
SSESN	83.50	67.92	20.30	62.67
Bi-SRNet	<u>90.31</u>	<u>81.03</u>	<u>45.39</u>	<u>81.39</u>
CdSC	**93.05**	**85.10**	**55.89**	**86.98**

where q_{ij} indicates the number of pixels that are identified as the i th change type and actually belong to the j th change type, and $\hat{q}_{j+}$ and $\hat{q}_{+j}$ represent the row sum and column sum of the confusion matrix without q_{00}.

F1 is the harmonic mean of precision and recall, which provides a comprehensive measure of overall performance:

$$\begin{aligned} \mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{TP}}, \\ \mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{F1} &= 2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}. \end{aligned} \tag{23}$$

D. Comparison with Advanced Methods

To evaluate the superiority of CdSC, we compare it with 4 SOTA algorithms. The brief descriptions of the comparative algorithms are as follows:

1) HRSCD.str3 [8]: The initial attempt to separate SCD into two tasks, change detection, and semantic prediction, involves employing three networks to predict the semantic maps and change maps for the bi-temporal images.

2) HRSCD.str4 [8]: Building upon HRSCD.str3, this approach incorporates information sharing by transmitting semantic features from the semantic prediction branch to the CD branch through skip connections.

3) SSESN [11]: An end-to-end spatial and semantic enhancement siamese network that aggregates abundant spatial and semantic information within images through a specially designed spatial and semantic feature fusion module.

4) Bi-SRNet [10]: An SCD architecture that considers semantic temporal features, inferring both single-temporal and cross-temporal semantic correlations through two semantic reasoning blocks and utilizing cosine loss to enhance the semantic consistency of CD results.

Comparison Experiments on the SECOND: Fig. 5 shows the qualitative results of various algorithms on the SECOND dataset, demonstrating the superior detection performance of

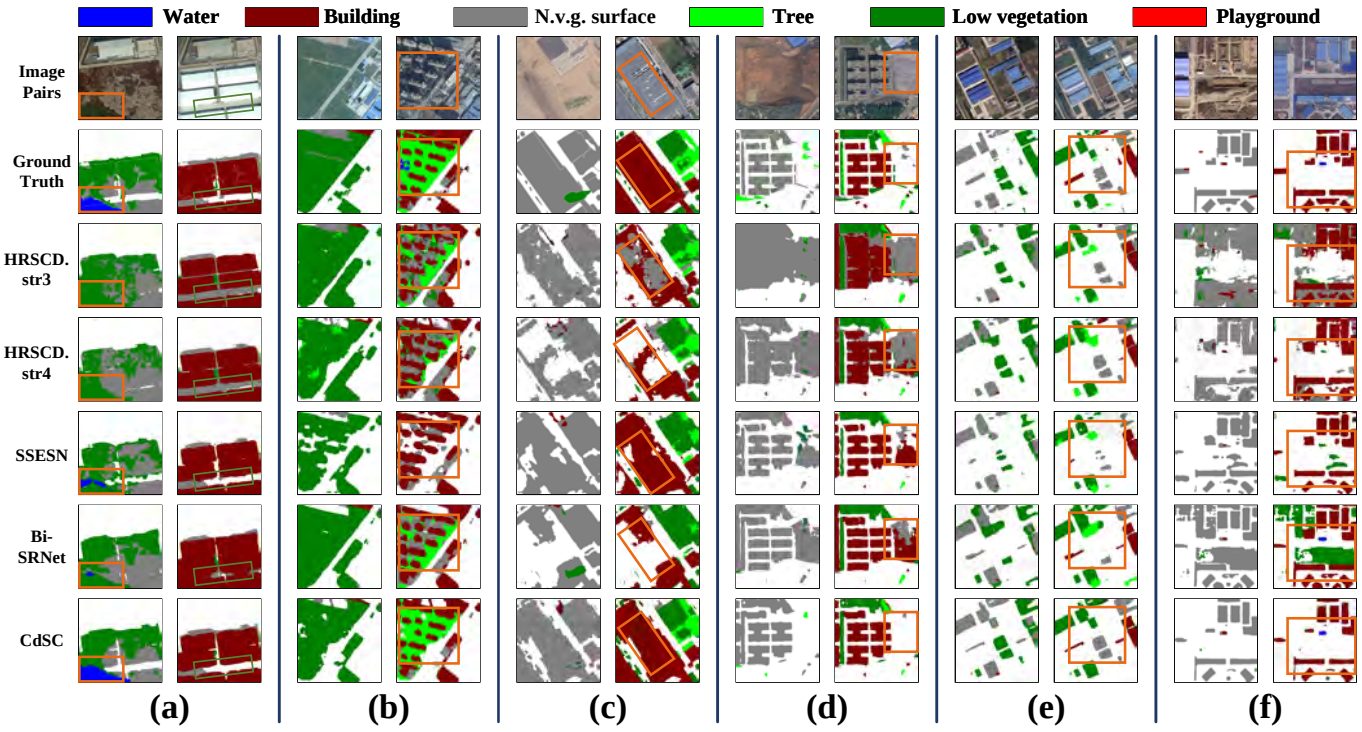

Fig. 5. The qualitative comparison of different methods on the SECOND dataset, please zoom-in for the best view.

CdSC. In Fig. 5(a), compared to other competitors, CdSC uniquely recognizes the water body in the lower left corner of the previous temporal image. At the same time, other methods are prone to false change detection within the rectangular area of the post temporal image. Fig. 5(b) illustrates a scenario where bare land changes into a residential area, most competitors failed to identify trees within the residential area. As shown in Fig. 5(c), due to the spectral similarity between the top of the buildings and nonvegetated ground surface, only CdSC effectively extracted the entire changed building. On the right side of Fig. 5(d), there is a clear spectral change occurring between the two time phases. The superior difference information extraction of CdSC prevents it from false detections. Furthermore, within the building area, CdSC also restores finer details in contrast to other methods. The scene shown in Fig. 5(e) contains scattered changes, CdSC excels in extracting changes comprehensively and distinctly when compared to other methods. In Fig. 5(f), the image pair from two different time phases exhibits significant spectral differences, resulting in a high rate of false positives in other methods, whereas CdSC maintains good performance even in cases of spectral variability.

Table I displays the quantitative results, demonstrating the exceptional performance of CdSC across all evaluation metrics. Our method achieves a mIoU of 73.32% and a SeK of 23.52%, outperforming the second-best method, Bi-SRNet, by 0.67 and 1.61 percentage points, respectively.

Comparison Experiments on the JL1: Fig. 6 presents the visual comparison of different methods. As shown in Fig. 6(a), due to the spectral similarity between bare soil regions and cropland during the autumn and winter seasons, other compared methods failed to detect the transition from cropland to bare soil, with only CdSC successfully identifying this change. Fig. 6(b) shows a scenario where the change occurs from a forested area to cropland. In the detection results of SSESN, there is a significant discrepancy between the change areas and the semantic map. Specifically, the areas where changes occur in the bi-temporal semantic maps are incorrectly categorized as the same category. A minor number of similar errors also appeared in Bi-SRNet. Supervised by the SCA loss, CdSC effectively circumvents the aforementioned issues. In the post image of Fig. 6(c), road features are not distinct, and the cropland exhibits significant spectral differences due to seasonal variations. The detection results of CdSC are in substantial agreement with the actual changes. In the right-side area of Fig. 6(d), the transformation from cropland to industrial

TABLE III
ABLATION EXPERIMENT ON THE SECOND DATASET. THE BEST RESULTS ARE MARKED IN BOLD.

3D-CD	SCE	SCA-Loss	OA (%)	mIoU(%)	SeK(%)	F1(%)
			89.41	71.77	20.81	61.24
✔			89.68	72.70	22.35	62.77
	✔		89.72	72.63	22.03	61.98
		✔	89.78	72.37	21.61	61.75
✔	✔		89.83	73.15	22.92	62.99
	✔	✔	89.40	72.82	22.67	62.95
✔		✔	89.77	73.13	23.09	63.28
✔	✔	✔	**90.07**	**73.32**	**23.52**	**63.70**

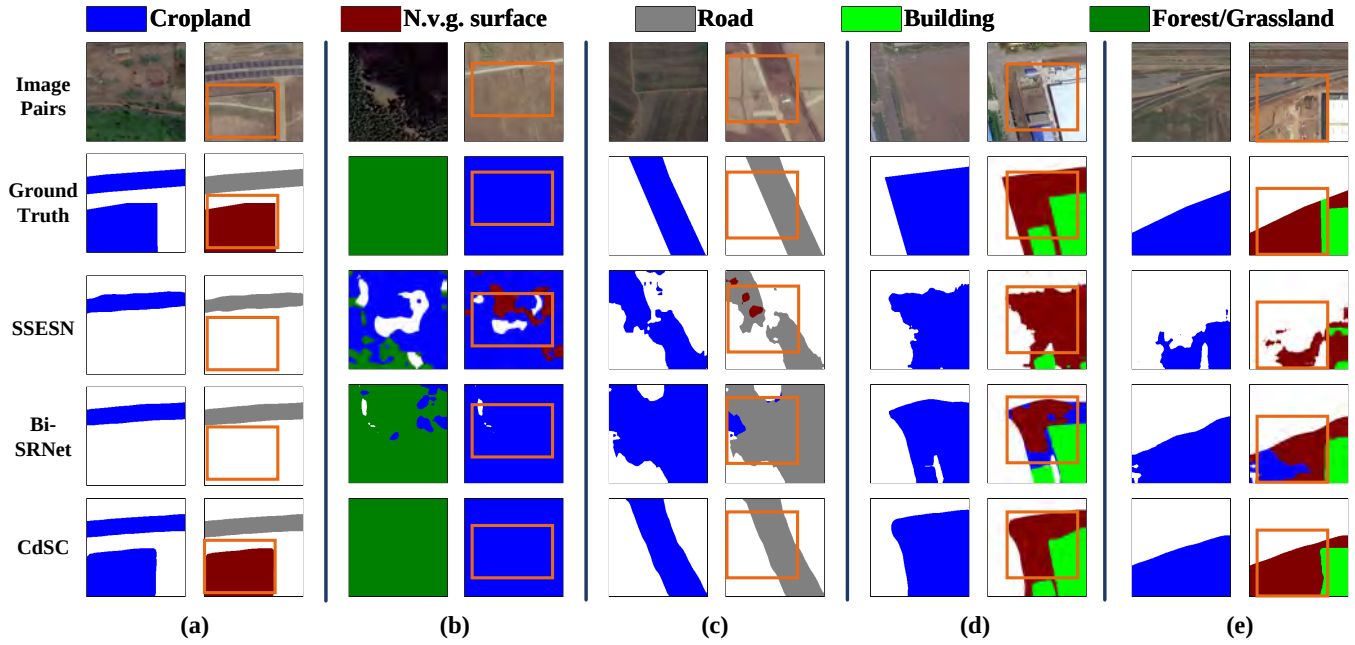

Fig. 6. The qualitative comparison of different methods on the JL1 dataset, please zoom-in for the best view.

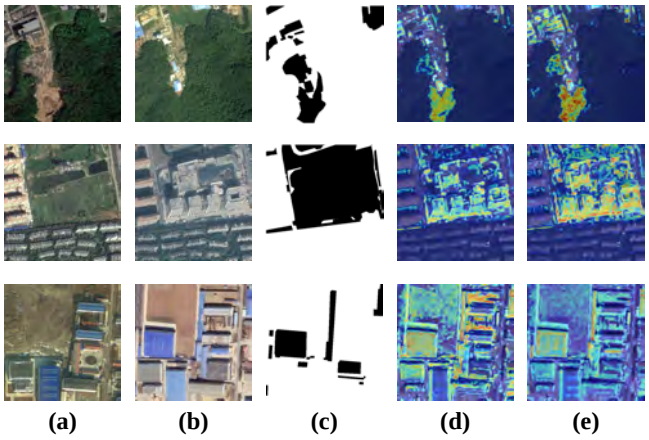

Fig. 7. Example of 3D-Cross-Difference module visualization by Grad-CAM. (a) Pre-temporal image. (b) Post-temporal image. (c) Binary change ground truth. (d) Grad-CAM of the model with 3D-Cross-Difference module. (e) Grad-CAM of the model without 3D-Cross-Difference module.

land occurred, and CdSC detected changes in buildings and non-vegetated ground surfaces with greater precision and detail compared to the competitors.

Table II displays the quantitative results for the JL1 dataset. Apart from Bi-SRNet, the competing methods exhibited subpar performance, primarily attributed to the substantial seasonal spectral variations within the predominant cropland category in the dataset. CdSC maintains a lead of 4.07% and 10.50% in the composite metrics of mIoU and SeK, respectively, compared to the second-ranked method.

E. Ablation Studies

To investigate the impact of the proposed modules and loss function on SCD, comprehensive ablation studies were performed using the SECOND dataset. These studies involved selectively removing certain components for comparative analysis. As indicated in Table III, there is a noticeable improvement in model performance with the incremental addition of components. Furthermore, some results are visually presented in Fig. 8, providing an intuitive demonstration of the effectiveness of the proposed modules and the employed loss function for CdSC.

1) Baseline: The baseline network employs the same backbone as CdSC. In the decoder, the backbone features are used to generate the bi-temporal semantic maps and the change map without the extraction of difference features through the 3D-Cross-Difference module, while the SCE module is omitted. The baseline is supervised using the cross-entropy loss function for fair comparison, with all other hyperparameters kept consistent in all comparative experiments. The baseline achieves 71.77% mIoU and 20.81% SeK on the SECOND dataset.

2) Effects of 3D-Cross-Difference Module: To mine the profound differences between bi-temporal images, we develop the 3D-Cross-Difference Module to extract discriminative difference features. As presented in Table III, the proposed difference module notably improves the performance of SCD, as indicated in rows 2, 5, and 7. Particularly, integrating the 3D-Cross-Difference Module into the baseline results in a significant enhancement of the mIoU metric, elevating it from 71.77% to 72.70%. Fig. 8 illustrates the effective reduction of extensive false change detections achieved by the 3D-Cross-Difference Module.

To further demonstrate the role of the 3D-Cross-Difference

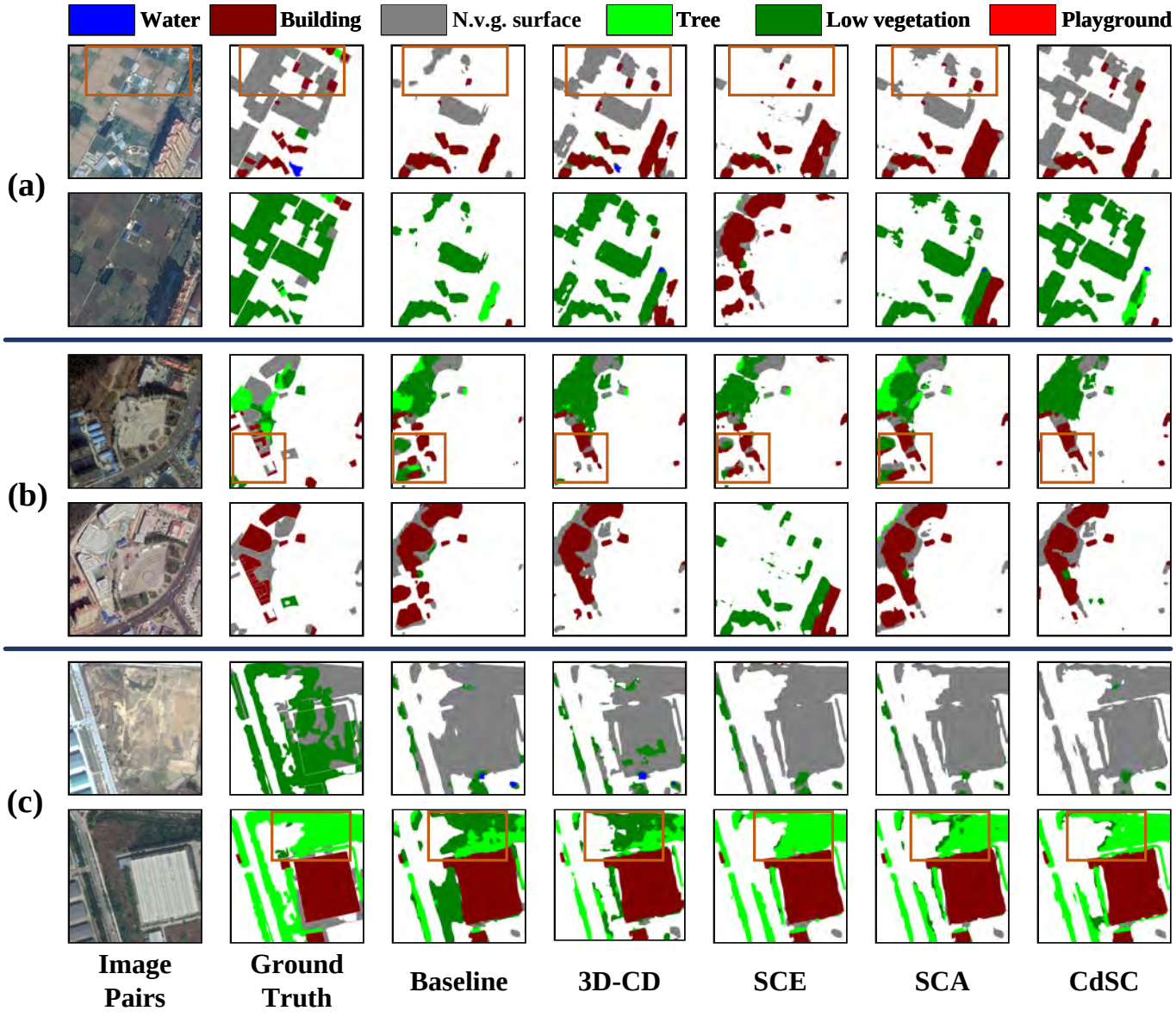

Fig. 8. The qualitative results of ablation study.

module in extracting robust difference features, we visualize the Gradient-weighted class activation maps (Grad-CAM) [60] with or without the 3D-Cross-Difference module. The model without the 3D-Cross-Difference module uses regular convolution for difference computation. As shown in the first two rows of Fig. 7, the 3D-Cross-Difference module extracts change features more comprehensively compared to conventional convolutions. Furthermore, it significantly enhances object change regions, highlighting its advantages in extracting discriminative difference features. In the third row, spectral changes cause traditional convolutions to fail drastically, resulting in numerous false changes, whereas the 3D-Cross-Difference module robustly handles such illumination variations.

3) Effects of SCE Module: To alleviate semantic conflicts within difference features, we introduce the SCE module to refine these discrepancies further. As demonstrated in Table III, the SCE module significantly enhances the SeK metric of the model. In various ablation tests with different combinations of components, the SCE module brings improvements ranging from 0.19 to 1.22 percentage points. In Fig. 8(c), when compared to the baseline network, the addition of SCE results in a noticeable enhancement in the model's ability to detect change regions and improve semantic discrimination.

In addition to the 3D-Cross-Difference and SCE modules, the proposed SCA loss function has also contributed to varying degrees of improvement. When combined with the introduced modules, the SCA loss function offers complementary advantages. Specifically, training the baseline network with SCA loss results in a 0.6% and 0.8% enhancement in mIoU and SeK metrics for the SCD task, respectively. When used in conjunction with the 3D-Cross-Difference module, the model achieves a SeK of 23.09%.

V. Conclusion

In this paper, we propose a SCD framework called Cross-Difference Semantic Consistency (CdSC) network, which exhibits outstanding capabilities in extracting difference features and semantic representations. To explore deep differences within spatiotemporal instance features, we develop the 3D-Cross-Difference module for extracting robust discriminative features. Furthermore, we introduce a Semantic Context Enhancement module to enhance the consistency between difference features and bi-temporal representations. Meanwhile, in response to semantic disparities within the maps, we formulated the Semantic Co-alignment loss function which uses contrastive learning to rigorously supervise the alignment of change and semantic information. Experimental results show that CdSC surpasses SOTA methods in terms of quantitative metrics and qualitative results on two public datasets. However, it is worth noting that the 3D-Cross-Difference module relies on the internal parameter optimization of the network for extracting difference information. An important future improvement would be to develop a dedicated 3D central difference convolution for difference feature extraction, which can further enhance the interpretability of difference features and the robustness of the model to different scenarios.

References

[1] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, 2005.

[2] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference enhancement and spatial–spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[3] L. Ru, B. Du, and C. Wu, "Multi-temporal scene classification and scene change detection with correlation based fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 1382–1394, 2021.

[4] A. A. Abuelgasim, W. Ross, S. Gopal, and C. Woodcock, "Change detection using adaptive fuzzy neural networks: Environmental damage assessment after the gulf war," *Remote Sens. Environ.*, vol. 70, no. 2, pp. 208–223, 1999.

[5] G. Satalino, F. Mattia, A. Balenzano, F. P. Lovergine, M. Rinaldi, A. P. De Santis, S. Ruggieri, D. A. Nafría García, V. P. Gómez, E. Ceschia, M. Planells, T. L. Toan, A. Ruiz, and J. Moreno, "Sentinel-1 & sentinel-2 data for soil tillage change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2018, pp. 6627–6630.

[6] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, 2016.

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

[8] 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, p. 102783, 2019.

[9] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[10] 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, pp. 1–14, 2022.

[11] M. Zhao, Z. Zhao, S. Gong, Y. Liu, J. Yang, X. Xiong, and S. Li, "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, 2022.

[12] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, 2014.

[13] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, 03 2020.

[14] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.

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

[16] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2022, pp. 207–210.

[17] X. Zhang, M. Tian, Y. Xing, Y. Yue, Y. Li, H. Yin, R. Xia, J. Jin, and Y. Zhang, "Adhr-cdnet: Attentive differential high-resolution change detection network for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[18] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 4063–4067.

[19] Y. Zhao, P. Chen, Z. Chen, Y. Bai, Z. Zhao, and X. Yang, "A triple-stream network with cross-stage feature fusion for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[20] R. Huang, R. Wang, Q. Guo, Y. Zhang, and W. Fan, "Idet: Iterative difference-enhanced transformers for high-quality change detection," 2022. [Online]. Available: https://arxiv.org/abs/2207.09240

[21] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

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

[23] B. Xiang, Y. Yue, T. Peters, and K. Schindler, "A review of panoptic segmentation for mobile mapping point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 203, pp. 373–391, 2023.

[24] W. Jing, Y. Yuan, and Q. Wang, "Dual-field-of-view context aggregation and boundary perception for airport runway extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[25] S. Liu, L. Chen, L. Zhang, J. Hu, and Y. Fu, "A large-scale climate-aware satellite image dataset for domain adaptive land-cover semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 205, pp. 98–114, 2023.

[26] Z. Luo, L. Gao, H. Xiang, and J. Li, "Road object detection for hd map: Full-element survey, analysis and perspectives," *ISPRS J. Photogramm. Remote Sens.*, vol. 197, pp. 122–144, 2023.

[27] K. Chi, Y. Yuan, and Q. Wang, "Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[28] W. Lin and A. B. Chan, "A fixed-point approach to unified prompt-based counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 3468–3476.

[29] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[30] U. Stilla and Y. Xu, "Change detection of urban objects using 3d point clouds: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 197, pp. 228–255, 2023.

[31] Y. Yuan, Z. Li, and D. Ma, "Feature-aligned single-stage rotation object detection with continuous boundary," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[32] H. Fang, S. Guo, P. Zhang, W. Zhang, X. Wang, S. Liu, and P. Du, "Scene change detection by differential aggregation network and class probability-based fusion strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.

[33] R. Zhang, H. Zhang, X. Ning, X. Huang, J. Wang, and W. Cui, "Global-aware siamese network for change detection on remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 61–72, 2023.

[34] Y. Yuan, Z. Xiong, and Q. Wang, "Vssa-net: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, 2019.

[35] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "Stadecdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.
[36] I. de Gélis, T. Corpetti, and S. Lefèvre, "Change detection needs change information: Improving deep 3-d point cloud change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–10, 2024.
[37] S. Sun, L. Mu, L. Wang, and P. Liu, "L-unet: An lstm network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[38] 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. Image Process.*, vol. 32, pp. 57–71, 2023.
[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10 012–10 022.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[41] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 599–609, 2023.
[42] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1489–1500, 2023.
[43] F. Zhu, J. Cui, and K. Dou, "Spatio-temporal hierarchical feature transformer for uav object tracking," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 442–452, 2023.
[44] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[45] C. W. H. W. F. Z. Decheng Wang, Feng Zhao and X. Chen, "Y-net: A multiclass change detection network for bi-temporal remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 2, pp. 565–592, 2022.
[46] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.
[47] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2021.
[48] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7303–7313.
[49] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 87, p. 102792, 2023.
[50] Y. Xiong, M. Ren, and R. Urtasun, "Loco: Local contrastive representation learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 11 142–11 153, 2020.
[51] M. Kang and J. Park, "Contragan: Contrastive learning for conditional image generation," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21 357–21 369, 2020.
[52] D. Zhang, F. Nan, X. Wei, S. Li, H. Zhu, K. McKeown, R. Nallapati, A. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," *arXiv preprint arXiv:2103.12953*, 2021.
[53] S. Lee, D. B. Lee, and S. J. Hwang, "Contrastive learning with adversarial perturbations for conditional text generation," in *International Conference on Learning Representations*, 2021.
[54] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, 2021.
[55] Q. Shen, J. Huang, M. Wang, S. Tao, R. Yang, and X. Zhang, "Semantic feature-constrained multitask siamese network for building change detection in high-spatial-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 189, pp. 78–94, 2022.
[56] H. Wang, X. Lv, and S. Li, "A new building change detection method based on cross-temporal stereo matching using satellite stereo imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[57] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham: Springer International Publishing, 2018, pp. 3–19.
[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[59] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, p. 98–136, jan 2015.
[60] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the Uni- versity of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.

Wei Jing received the B.M. degree in e-commerce and the M.S. degree in computer software and theory from Shandong University of Science and Technology, Qingdao, China, in 2019 and 2022 respectively.

He is currently working toward the Ph.D. degree in the National Elite Institute of Engineering and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image processing and deep learning.

Kaichen Chi received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing and deep learning.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.