RaSRNet: An end-to-end Relation-aware Semantic Reasoning Network for Change Detection in Optical Remote Sensing Images

Yi Liang (梁漪), Chengkun Zhang (张成坤), Min Han (韩敏), *Senior Member, IEEE*

***Abstract*—Optical Remote sensing images (RSIs) are used in surface observation, and one of the most interesting research topics is change detection (CD). The internal problem of RSIs, including multi-scales changed objects and cluttered background, still deserve attention. Existing methods make great efforts to solve this problem but inevitably miss detection, which affects the model performance. To address this dilemma, this article proposes a Relation-aware Semantic Reasoning Network (RaSRNet) in an end-to-end manner to pop-out change objects in RSIs, where the key point is to perceive contextual semantic information. The relation-aware module in RaSRNet combats the lack of contextual information caused by the limited receptive field of the general convolutional layer, which facilitates all-around changed object detection. The multi-level semantic reasoning encoder-decoder backbone in RaSRNet extracts and reconstructs pixel semantic information, alleviates the interference of background noise and improves the integrity recognition of changed objects. In addition, the decoder backend undertakes two semantic segmentation branches, and introduces a semantic reasoning loss between the two branches to infer pixel semantic categories, which provides more accurate semantic features for the CD. Extensive experiments are conducted on the three public RSIs CD datasets, and the results demonstrate that the proposed RaSRNet can accurately locate changed objects, which consistently outperforms the state-of-the-art CD competitors.**

***Index Terms*—Relation-aware, Semantic Reasoning, Optical remote sensing images, Change detection.**

I. Introduction

CHANGE detection (CD) aims to identify the change areas in remote sensing images (RSIs) that cover the same surface. CD can more effectively understand surface changes in the real world, so it has been widely applied in many tasks, such as urban planning [1], land use detection [2], and vegetation change detection [3], etc.

In recent years, many works have made great efforts to treat change detection as a technical problem of computer vision [4], [5]. According to the differences in computer vision tasks, CD methods can be roughly summarized as 1) metric learning-based CD method (MLCD), 2) classification-based CD method (CCD) and 3) segmentation-based CD method (SCD). The MLCD aims to mine the semantic similarity between paired RSIs and generates distance map. Euclidean distance [6], Mahalanobis distance [7], and cosine similarity [8] are often used to measure the similarity. However, a margin split is required to generate change map from distance map. The selection of the margin is closely related to the scene of RSIs, so MLCD is not flexible enough. The CCD aims to fit one or more classifiers to decide whether pixel changes or not. According to the time-ordered of performing classification, CCDs are divided into direct classification (d-CCD) and post-classification (post-CCD). The d-CCD is difficult to solve the semantic confusion between the changed samples since it is not limited by semantic information. The interference information tends to generate more false detections (e.g., FC-Siam-Diff is difficult to judge shadow disturbance when performing building CD in the 2nd row in Fig.1). The post-CCD requires the input RSIs are strictly registered, and they inevitably accumulate errors due to multi-classifier branches. The SCD is an advanced classification method, it allows to represent more complex semantic information and provides a more ingenious technical framework. Many recent studies focus on the semantic segmentation theory, which provides a new idea for RSIs CD [9]–[11].

The implementation of the SCD is to 1) transfer and replicate the semantic segmentation networks applied to single natural scene images (NSIs); 2) build semantic reasoning siamese branches which each outputs the semantic reasoning

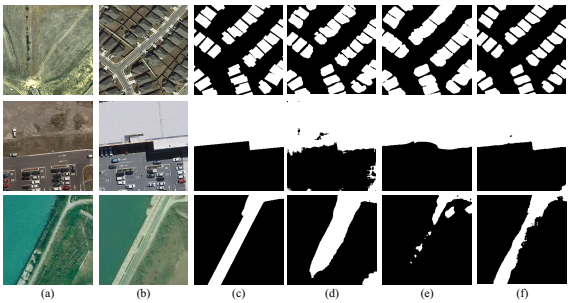

Fig. 1. Visualization of some examples: (a) Time t_1 RSIs. (b) Time t_2 RSIs. (c) Ground Truth. (d) FC-Siam-Diff. (e) DSAMNet. (f) proposed RaSRNet.

Yi Liang is with Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116000, China (e-mail:liangyi@mail.dlut.edu.cn).

Chengkun Zhang is with Department of Computer Technology and Application, Qinghai University, Xining 810000, China (e-mail:zhangchengkundon@163.com).

Min Han is with the Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Liaoning 116024, China (e-mail:minhan@dlut.edu.cn).

map corresponding to the input image; 3) generate change map by comparing semantic reasoning map. Although NSIs match optical RSIs from RGB in color patterns. Transferring the state-of-the-art semantic segmentation network from NSIs to RSIs CD still poses significant uncertainties due to the different imaging environments and imaging heights of RSI. To be precise, influenced by the high-altitude shooting and large-area coverage, the gaps of ground objects in the scale and quantity make it harder to accurately reason out semantic information (e.g., dense small-scale building change objects in the 1st row in Fig.1 and sparse large-scale road augments change objects in the 3rd row).

Existing articles use multi-scale convolutional structures to expand the receptive field (RF), and develop context information to explicitly alleviate the above problems [12]. Such as multi-scale feature fusion (MSFF) [10], pyramid pooling [13], atrous spatial pyramid pooling (ASPP) [14], feature pyramid [15], etc. However, the RF is still limited and the pooling operation is prone to cause local information loss. Graph-based relational awareness is a promising strategy, and it has been proven effective in capturing long-range contextual relations and leveraging global semantic information [16]–[18]. Therefore, in this paper, we propose to embed spatial and channel relation-aware modules into the encoder-decoder backbone. Its role is to model the semantic relationship of different objects on RSI and to model the distribution between feature channels.

On the other hand, deep features facilitate the learning of pixel semantic categories, and the shallow features with detailed boundary information help to achieve detail recovery. Many articles are based on the encoder-decoder networks (ED) that effectively utilize the deep and shallow features to enhance the feature discriminative power and noise immunity [19]–[21]. Depending on the network branch, the ED can be divided into the single-stream ED (SSED) and double-stream ED (DSED) [5]. The SSED directly learns the difference between RSIs (as shown in Fig.2(a)). The DSED learns the differences between RSIs after encoding. Further, DSED is refined into "dual-encoder+single-decoder (DSED-e)" and "dual encoder-decoder (DSED-d)" structures (as shown in Fig.2(b),(c)). In SSED and DSED-e, the direct difference of rough shallow features will increase the noise interference. Therefore, we design a multi-level semantic reasoning ED following DSED-d to learn differential features at the decoder backend. Subsequently, we concatenate two semantic segmentation heads and a change detection head at the siamese decoder backend. A hybrid supervised training mechanism is proposed to guide the network to generate more representative and discriminative features.

In this paper, we devote to fully exploiting an end-to-end semantic reasoning network with relation-aware for CD in optical RSIs. The main contributions are given as follows.

1) An end-to-end relation-aware semantic seasoning network (RaSRNet) is proposed to achieve CD in optical RSIs, equipped with a multi-level semantic reasoning encoder-decoder and relation-aware (Ra) modules that is separable from the backbone.
2) We establish a semantic reasoning encoder-decoder backbone (SRNet), in which a batch of feature fusion blocks in the decoder can integrate multi-scale information. The back-end of the decoder connects the semantic segmentation head and the change detection head, which can explicitly reason semantic properties and provide the necessary feature constraints.
3) We propose the Ra module to learn global contextual features in spatial and channel dimensions through graph embedding, while comprehensively capturing ground objects. This is the first attempt to introduce relational awareness in the CD framework of optical RSIs.
4) We develop a hybrid loss function to constrain the training of RaSRNet. Specifically, (1) the deeply supervised strategy acts on the shallow features extracted by the encoder to constrain the boundary representation of changed objects; (2) a semantic reasoning loss acts on the semantic segmentation maps to constrain the semantic properties of unchanged area.

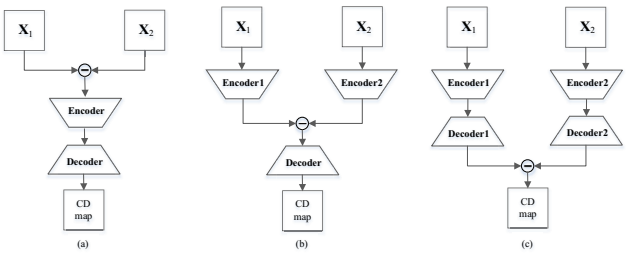

Fig. 2. The structures of ED. (a) SSED. (b) DSED-e. (c) DSED-d.

II. Related Works

A. Semantic Segmentation Transfer in CD Models

The CD model based on semantic segmentation transfer assigns specific changed/unchanged semantic labels to pixels by semantic information. Currently, the most commonly used segmentation framework on CD is the ED structure. This structure merges deep features into shallow features. Since multi-level features have different resolutions, fusing them can capture multi-scale semantic information and provide high-resolution output. Fully Convolutional Network (FCN) is a network with an ED structure [22]. The decoder of FCN uses deconvolution layers for upsampling, which can restore the feature map to the size of the input image. U-Net is an efficient and symmetric ED network, whose skip connections allow encoded features to be retained and reused [23]. U-Net++ adopts dense connections to replace skip connections in U-Net to enrich the fusion of multi-level features [24], [25]. In [26]–[28], multi-scale convolution is used to capture richer comprehensive semantic content.

B. Global relation awareness in CD Models

Deep learning-based CD models are inseparable from the basic structure of "convolutional layers", which are stacked to learn more non-local information. The large number of

the stacked convolutional layers makes it more difficult to learn a suitable model. The latest efforts have focused on the development of attention mechanisms in spatial dimension [29]–[31] and channel dimension [10], [26], [32] for learning non-local details. However, spatial attention will gain an attention map with twice of the feature size. Channel attention obtains a high-dimensional weight vector over deep features. This means the attention mechanism consumes a huge computational cost. The graph structure is an effective way to simplify the attention mechanism. Using graph convolutional topology to learn the relationship between graph nodes, and then mapping the graph nodes on multiple regions of features to form region-region relationship learning. Currently, graph-based relation awareness is applied in object detection [16], semantic segmentation [18], and image classification [33]. It is still a frontier of exploration. Therefore, inducing directed graph to efficiently perceive inter-domain relationships between spatial objects is critical for enhancing the discriminative feature.

C. Graph Convolution

Thomas *et al*. proposed Graph Convolutional Networks (GCN) in 2017 [34]. For a graph representation $G=(V,E)$, where V is the set of nodes and E is the set of edges linking the nodes. A symmetric normalized Laplacian matrix $L=I-D^{-1/2}AD^{-1/2}$ represents the relationship between nodes. I is the identity matrix, D is the degree matrix (diagonal matrix), and A is adjacency matrix with 0 and 1. A_{mn} is 1 when node m and node n are connected by an edge. L has N (the number of nodes) eigenvectors (denoted as U) that can be decomposed as $U\Lambda U^{\mathrm{T}}$. Extending the Fourier transform to the graph structure, there is $\hat{x}=U^{\mathrm{T}}x$, where x is the input signal, the spectral domain signal $\hat{x}$ is obtained through the orthogonal bases U.

By analogizing the above convolution theorem to graph topology, the graph convolution can be defined by spectral filter Λ and spectral domain signal $\mathbf{s}$, i.e., $g(\Lambda)*\mathbf{s}=Ug(\Lambda)U^{\mathrm{T}}\mathbf{s}$. However, U is very computationally intensive. In order to simplify the graph convolution operation, later studies proposed to use K-th order Chebyshev polynomials T_k to approximate $Ug(\Lambda)U$ [35]. Therefore, the convolution of the graph signal can be expressed as follows:

$$g(\Lambda)*\mathbf{s}\approx\sum_{k=0}^{K}\theta_k T_k(L)\mathbf{s} \tag{1}$$

where θ_k is Chebyshev coefficients vector. Then, restricting K to 2 and limiting the largest eigenvalue of L to 2. Formula (1) is can be further simplified as follows:

$$g(\Lambda)*\mathbf{s}\approx\theta(I+D^{-1/2}AD^{-1/2})\mathbf{s} \tag{2}$$

$$\theta(I+D^{-1/2}AD^{-1/2})\mathbf{s}=\theta(\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2})\mathbf{s} \tag{3}$$

where θ is the Chebyshev coefficient. $I+D^{-1/2}AD^{-1/2}$ is renormalized to $\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}$. $\widetilde{A}=I+A$, $\widetilde{D}_{ii}=\sum_{i=1}\widetilde{A}_{ij}$.

In summary, for the multi-channel input signal $\mathbf{X}\in\mathbb{R}^{HW\times C}$ (each node has a feature vector with C-dimension), its graph convolution is shown in formula (4):

$$Z=\sigma(\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}\mathbf{X}\Theta) \tag{4}$$

where Z is updated node feature, Θ is a weight matrix, and σ is activation function.

III. Proposed Model

A. Architecture Overview

The architecture of the proposed RaSRNet is shown in Fig.3, which contains 1) a multi-level semantic reasoning encoder-decoder backbone network (SRNet), 2) three relation-aware modules, 3) two deeply-supervised modules (DSs), 4) two semantic segmentation heads (Seg Heads), and 5) a change detection head (CD Head). Firstly, the input RSI ($\mathbf{X}^{(t_1)}$ and $\mathbf{X}^{(t_2)}$) are passed into the encoder (ResNet18 without fully connected layer) to extract the multi-scale encoded features $\{\mathbf{F}_{e,k}\}_{k=1}^{4}$. Then, $\mathbf{F}_{e,1}$, $\mathbf{F}_{e,2}$ and $\mathbf{F}_{e,3}$ are sent to the three Ra modules. Concretely, Ra tries to model the semantic relation between objects in RSI and outputs the enhanced features ($\mathbf{F}_{ra,1}$, $\mathbf{F}_{ra,2}$ and $\mathbf{F}_{ra,3}$). The cascaded FFBs in decoder are used to fuse multi-scale features and generate the last decoded feature $\mathbf{F}_{d,1}$. Finally, seg heads act on the $\mathbf{F}_{d,1}$ to explicitly display segmentation maps, and a CD head generates a predicted CD map. In addition, we extract more useful information by constraining shallow encoded features ($\mathbf{F}_{e,1}$ and $\mathbf{F}_{e,2}$) with deeply supervised (DS) modules. DS is formed by two stacked transposed convolution layers (transconv) with 3×3 kernel, where the first transconv is followed by batch normalization and Relu activation. During training, the CD map is used to calculate the change loss with labels. The paired segmentation maps constrain each other to enhance the detail of unchanged areas.

B. Semantic Reasoning Encoder-Decoder Backbone

Our backbone is based on an encoder-decoder structure and is named semantic reasoning network (SRNet). The encoder of SRNet is based on the ResNet18, and its decoder consists of a series of FFBs (as shown in the right part of Fig.3). So far, the decoder output feature $\mathbf{F}_{d,1}$ covers both coarse-grained low-level semantic information and fine-grained high-level semantic information. This is important for detail recovery of objects of different scales. Finally, the semantic segmentation map of paired images and CD map are output through three prediction branches (two seg heads and a CD head).

C. Relation-aware module

In order to obtain a more comprehensive internal modeling relationship, we perform graph-based relation-aware in both spatial and channel dimensions. Ra module is an improved work based on the global reasoning unit proposed by Chen *et al* [17]. Specifically, convolutional features $\mathbf{F}_{e,k}\in\mathbb{R}^{C\times H\times W}$ ($k=1,2,3$) are transported to parallel spatial relation-aware (SRa) and channel relation-aware (CRa) modules to generate enhanced features $\mathbf{F}_{ra,k}\in\mathbb{R}^{C\times H\times W}$. The relation-aware process consists of three important steps, i.e., 1) domain

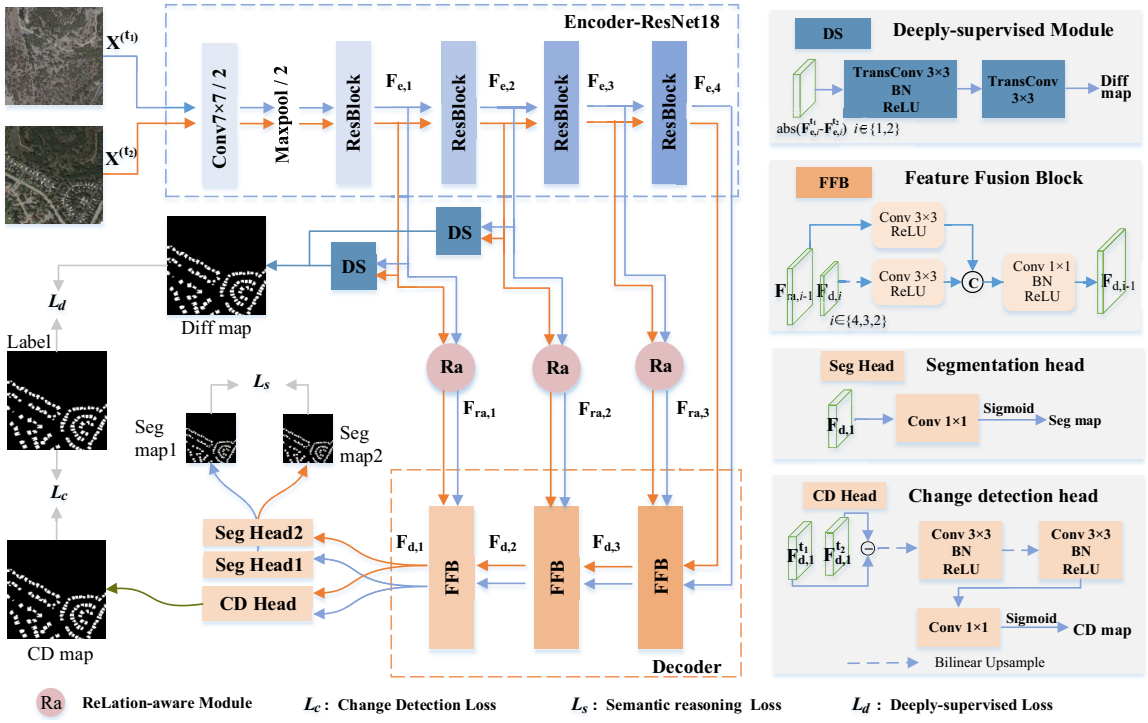

Fig. 3. Architecture of the proposed RaSRNet. The blue and yellow solid lines represent the processing flow of two RSIs. The dotted box constitutes the baseline model (i.e., SRNet).

projection and graph representation, 2) graph node relation-aware, and 3) graph re-projection. In CRa, we permute the dimension of $\mathbf{F}_{e,k}$ and decompose it in the channel dimension (i.e., $\mathbf{F}^c_{e,k} \in \mathbb{R}^{HW\times(C/M)\times M}$, where (C/M) and M are the reshaped height and width), and then do the same processing as SRa. Finally, the configuration of combing SRa and CRa is two convolutional layers (the kernel sizes are 3×3 and 1×1) with batch normalization. The detailed structure is shown in Fig.4.

Step1 : Domain projection and graph representation.

The purpose of this step is to convert CNN features $\mathbf{F}_{e,k} \in \mathbb{R}^{C\times H\times W}$ into graph-structured data $G_{pro,k} \in \mathbb{R}^{N\times C}$, where C, H and W are the number of channels, height and width of the initial feature, respectively. N is the number of nodes in the graph space. One convolutional layer with 1×1 kernel (out_channel = N_s) is used to obtain the graph representation as suggested in [17]. That is, $G^s_{pro,k} = \phi(\mathbf{F}_{e,k}) = W_\phi \times \mathbf{F}_{e,k}$ in SRa, where $G^s_{pro,k} \in \mathbb{R}^{N_s\times HW}$. In addition, high-dimensional channel of feature consume excessive computational resources. Therefore, using another convolutional layer with 1×1 kernel to reduce feature dimension and reshape $\mathbf{F}_{e,k}$ to 2D tensor. As $\mathbf{F}^s_{e-,k} = (\varphi(\mathbf{F}_{e,k}))^{\mathrm{T}} = (W^-_\varphi \times \mathbf{F}_{e,k})^{\mathrm{T}}$, where $\mathbf{F}^s_{e-,k} \in \mathbb{R}^{HW\times C_1}$, and $N_s < C_1 < C$. So far, according to the form of formula (1), we can obtain the spatial graph representation of the feature $\mathbf{F}_{e,k}$ as follows:

$$G^s = G^s_{pro,k} * \mathbf{F}^s_{e-,k} = \phi(\mathbf{F}_{e,k}) \cdot (\varphi(\mathbf{F}_{e,k}))^{\mathrm{T}} \tag{5}$$

where $G^s \in \mathbb{R}^{N_s\times C_1}$, G^s with N_s vertexes, and each vertex is represented by the corresponding channel feature of $\mathbb{R}^{C_1\times 1}$. Similarly, the channel graph representation is as follows:

$$G^c = G^c_{pro,k} * \mathbf{F}^c_{e-,k} = \phi(\mathbf{F}^c_{e,k}) \cdot (\varphi(\mathbf{F}^c_{e,k}))^{\mathrm{T}} \tag{6}$$

where G^c with N_c vertexes, and each vertex is represented by the corresponding spatial feature of $\mathbb{R}^{HW\times 1}$.

Step2 : Graph node relation−aware.

Graph convolution is used to learn the graph relation in step1 and to aware the dependencies between nodes. According to the description of formula (4), the relation-aware of graph nodes in SRa is defined as:

$$G^s_{ra} = \sigma(L^s G^s W^s_g) \tag{7}$$

where G^s and $G^s_{ra} \in \mathbb{R}^{N_s\times C_1}$ are the input and output node state of graph convolution. $\sigma(\cdot)$ is the ReLU activation. $W^s_g \in \mathbb{R}^{C_1\times C_1}$ is a trainable weight matrix. $L^s \in \mathbb{R}^{N_s\times N_s}$ is the Laplacian matrix, here we choose the symmetric normalized form described in formula (8) to construct it:

$$\widetilde{L}^s = I - \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2} \tag{8}$$

where $I \in \mathbb{R}^{N_s\times N_s}$. $\widetilde{A} \in \mathbb{R}^{N_s\times N_s}$. $\widetilde{D} = diag(d_1, d_2, \cdots, d_{N_s}) \in \mathbb{R}^{N_s\times N_s}$. $d_m = \sum_n \widetilde{A}_{mn}, \{m,n\} \in \mathbb{R}^{N_s}$. Then, we need to build the adjacency matrix $\widetilde{A}$ to accomplish the above goal.

$\widetilde{A}$ indicate the similarity between N_s vertexes in G^s. According to existing experience [16], [36], euclidean distance is used to build $\widetilde{A}$. The similarity between node m and n is represented as follows:

$$\widetilde{A}_{mn} = \rho(G^s)_m \widetilde{\Lambda}(G^s) \rho(G^s)^{\mathrm{T}}_n \tag{9}$$

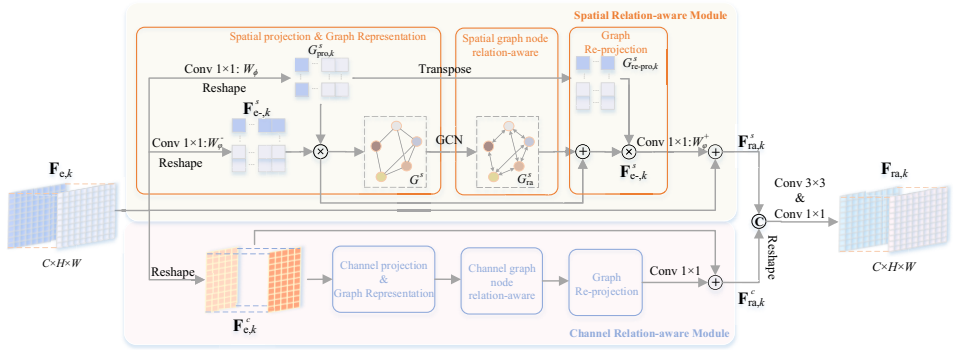

Fig. 4. The structure of Relation-aware module.

where $\rho(\cdot)$ is a fully connected layer with Relu activation. $\widetilde{\Lambda}(G^s)$ is a diagonal matrix that saves inner product of G^s:

$$\widetilde{\Lambda}(G^s) = diag(conv(G^s)) \tag{10}$$

where $conv(\cdot)$ is a 1D convolutional with Relu activation, $diag(\cdot)$ expands a vector to a diagonal matrix. Similarly, channel graph node relation-aware is defined as follows.

$$G^c_{ra} = \sigma(L^c G^c W^c_g) \tag{11}$$

Step3 : Graph re−projection.

The goal of re-projection is to convert graph structure features into CNN features. To comprehensively represent graph node states, we use skip connections to bridge the initial graph G and updated graph G_{ra}. Re-projection is symmetric with the projection process in the step1. For simplicity, we perform the re-projection using the mapping matrix W_ϕ from step1,

$$\mathbf{F}^s_{e+,k} = W_\phi^{\mathrm{T}} \times (G^s + G^s_{ra}) \tag{12}$$

where $\mathbf{F}^s_{e+,k} \in \mathbb{R}^{HW\times C_1}$. $W_\phi^{\mathrm{T}} \in \mathbb{R}^{HW\times N_s}$ is the transpose of W_ϕ. Then, we reshape $\mathbf{F}^s_{e+,k}$ and increase its dimensions to be consistent with $\mathbf{F}_{e,k}$. Combining the features before and after relation-aware is denoted as $\mathbf{F}^s_{ra,k}$,

$$\mathbf{F}^s_{ra,k} = W^+_\phi \times \mathbf{F}^s_{e+,k} + \mathbf{F}_{e,k} \tag{13}$$

where $W^+_\phi \in \mathbb{R}^{C_1\times C}$ is the weight matrix of the dimension-raising. The output feature of CRa is denoted as $\mathbf{F}^c_{ra,k}$.

D. Loss Function

The loss function consists of three components, including the binary change detection loss L_c, the semantic reasoning loss L_s and the deeply-supervised loss L_d. L_c measures the difference between the predicted CD map $\hat{Y}$ and the reference ground truth Y. We use a combination of Dice loss and binary cross-entropy (BCE) loss to calculate the L_c on each pixel, which is described as follows,

$$L_c = f(\hat{Y}, Y) = f_{dice}(\hat{Y}, Y) + f_{bce}(\hat{Y}, Y) \tag{14}$$

$$f_{dice}(\hat{Y}, Y) = 1 - \sum_{i=1,j=1}^{H_0\times W_0} \frac{2\hat{Y}_{ij}Y_{ij} + \varepsilon}{\hat{Y}_{ij} + Y_{ij} + \varepsilon} \tag{15}$$

$$f_{bce}(\hat{Y}, Y) = \sum_{i=1,j=1}^{H_0\times W_0} \frac{Y_{ij} log\hat{Y}_{ij} + (1-Y_{ij}) log(1-\hat{Y}_{ij})}{H_0 \times W_0} \tag{16}$$

where H_0 and W_0 are the height and width of the input RSIs. i and j are the position indexes of the pixels. ε is a smoothing constant set to 0.0001. $Y_{ij} = 1$ ($Y_{ij} = 0$) indicates the changed (unchanged) pixel in the ground truth.

In Chen *et al* [11], L_s is used to measure the difference between semantic segmentation maps $S^{(t_1)}$ and $S^{(t_2)}$. Inspired by it, we aim to minimize the difference between $\{S^{(t_1)}|R_u\}$ and $\{S^{(t_2)}|R_u\}$, where R_u denotes the set of unchanged pixels in the ground truth. L_s is calculated as follows.

$$L_s = f_{R_u}(S^{(t_1)}, S^{(t_2)}) = f_{dice|R_u}(S^{(t_1)}, S^{(t_2)}) + f_{bce|R_u}(S^{(t_1)}, S^{(t_2)}) \tag{17}$$

$$R_u = \{(i,j)|Y_{ij} = 0\} \tag{18}$$

L_d calculates the difference between the underlying difference encoded features and the ground truth Y. We use it to limit the extraction of more useful underlying features. Concretely, the underlying difference feature of dual RSI ($|\mathbf{F}^{(t_1)}_{e,k} - \mathbf{F}^{(t_2)}_{e,k}|$, $\{k = 1,2\}$) is input of the DS module, which outputs a difference map D_k with the same size as Y. Therefore, L_d can be formulated as follows.

$$L_d = f(D_k, Y) = f_{dice}(D_k, Y) + f_{bce}(D_k, Y) \tag{19}$$

Finally, the total loss of RaSRNet is defined as,

$$L = L_c + \alpha L_s + \beta L_d \tag{20}$$

where α and β are hyperparameters, which we discuss in section IV.

IV. Experiments

A. Datasets and Implementation Details

The simulation experiments are carried on three public datasets, including 1) LEVIR-CD [29], 2) WHU-CD [37] and 3) SYSU-CD [30]. They are described as follows.

LEVIR−CD. The dataset includes 637 pairs of optical RSIs with the size of 1024×1024. These RSIs with a resolution of 0.5m and record the changes in various buildings.

We use the training, validation and testing RSIs provided by the authors and crop them to 256×256.

WHU−CD. The dataset consists of two scenes aerial images with the sizes of 15354×21243 and 15354×11265. The resolution of two period RSIs is 0.3m. We crop the RSIs to 256×256 slices with an overlap on the right and bottom. Then, we randomly divide these slices to 5460, 779, 1516 pairs for training, validation and testing, respectively.

SYSU−CD. The dataset includes 20,000 pairs of optical RSIs with the size of 256×256. These RSIs have a resolution of 0.5m. We use the original RSIs provided by the authors, where 12000, 4000, 4000 pairs of RSIs are used for training, validation and testing, respectively.

We implement the RaSRNet using PyTorch on a PC with an Intel Core i7-8700 3.20-GHz CPU, 16-GB RAM, and an NVIDIA GTX 1070Ti GPU. In the domain projection and graph representation of SRa, we set the number of graph node to be $\frac{1}{4}$ of the number of channels of $\mathbf{F}_{e,k}$, and the output channel of the reduced dimensional convolution layer is set to $\frac{1}{2}$ of the number of channels of $\mathbf{F}_{e,k}$, i.e., $N_s = \frac{1}{2}C_1 = \frac{1}{4}C$. The same settings are used in CRa. During the training, the encoder is initialized with pretrained ResNet18 on ImageNet, and the remaining parameters are randomly initialized by Kaiming initialization. Besides, the batch size is set to 16, and the optimizer is Adam. The epoch is set to 50 and the initial learning rate (lr) is set to 0.001. The lr remains constant in the first 25 epochs and decays linearly to e^{-7} in the follow 25 epochs. The settings of hyperparameters α and β are shown in Table.I. We quantitatively evaluate the performance of model using F1-measure (F1), Kappa Coefficient (KC) and mean intersection-over-union (mIoU). Validation is transacted after each training epoch, and the best model on the validation set evaluates the performance of model on the test set.

TABLE I
THE SETTING OF HYPERPARAMETERS α AND β

	LEVIR	WHU	SYSU
α | β	0.2 | 0.3	0.1 | 0.2	0.1 | 0.3

B. Comparison with State-of-the-art Methods

We compare the proposed RaSRNet with five CD methods, including three classification-based methods [19] (i.e., FC-EF, FC-Siam-Conc and FC-Siam-Diff) and two metric learning methods (i.e., STANet [29] and DSAMNet [30]). To be fair, we train the five state-of-the-art CD models following the experimental setup set in this study.

Quantitative Evaluation. Table.II reports the quantitative comparison results. It shows that RaSRNet outperforms other methods on the three datasets. For example, the F1/KC/mIOU of RaSRNet is 8.74%, 9.01% and 6.96% higher than that of the second-ranked STANet on the LEVIR. The gain gap between RaSRNet and FC-Siam-Conc (second-ranked) is more obvious on WHU, the F1/KC/mIOU is higher than 15.23%, 15.92% and 11.64%. Besides, the F1/KC/mIOU is 1.87%, 3.22% and 2.44% higher than STANet on the SYSU. Note that our baseline model (i.e., SRNet) still has clear advantages compared with the Unet in [19]. This illustrates our "DSED-d" backbone is better than the SSED (FC-EF) and DSED-e (FC-Siam-Conc, FC-Siam-Diff). In addition, RaSRNet uses Ra modules to obtain the global context information, which achieves superior performance than the attention mechanism used in DSAMNet and STANet.

TABLE II
QUANTITATIVE COMPARATIVE STUDIES OF DIFFERENCE CD MODELS.

Model	LEVIR			WHU			SYSU		
	F1	KC	mIOU	F1	KC	mIOU	F1	KC	mIOU
FC-EF	.6295	.6023	.6995	.6714	.6565	.7367	.6370	.4796	.5827
FC-Siam-Conc	.7439	.7267	.7781	.7413	.7308	.7838	.6630	.5151	.6036
FC-Siam-Diff	.7783	.7641	.8040	.7183	.7068	.7685	.7256	.6290	.6944
DSAMNet	.7788	.7649	.8049	.5937	.5776	.6807	.7486	.6635	.7202
STANet	.8096	.7988	.8294	.6900	.6739	.7463	.7645	.6878	.7380
SRNet	.8567	.8498	.8677	.8106	.8032	.8333	.7390	.6695	.7271
RaSRNet	**.8943**	**.8889**	**.8990**	**.8936**	**.8900**	**.9002**	**.7823**	**.7200**	**.7624**

Qualitative Evaluation. Fig.5 provides the qualitative comparison results. As seen from this figure, RaSRNet has the fewest false detection (red and blue) in most cases and renders clean visuals. For paired scenes with various scales of changed objects, such as the 5th and 15th rows, the results show that RaSRNet is able to detect changed objects more comprehensively. For paired scenes with various numbers of changed objects, such as the 1st and 2nd rows, the boundaries of the changed object obtained by RaSRNet are all basically the same as those in ground truth. For paired scenes with large-scale changed objects across the RSI, such as the 8th and 13th rows, our model also achieves relatively fine-grained detection.

The brightness difference between the dual RSIs is more significant in LEVIR dataset. Some of our results are affected, such as the 4th row. This may be because the encoded layers of 4th stage in encoder prefers to the upper half of the building in time t_2 RIS, and the encoded features of 4th stage directly guide the information recovery in decoder, resulting in misjudgment. However, our model is confident to overcome this natural factor compared to other models, such as the 1st-3rd rows. In addition, scenarios with building shadows are more common in the LEVIR and WHU datasets, such as the 5th, 7th, 9th rows. Our model is able to distinguish between architectural solids and shadows compared to other methods. Proposed model not only demonstrates excellent performance in building CD, but also has positive effects in capturing urban change. This is reflected in the SYSU dataset. For example, sea construction in 11th and 15th rows, road expansion in 13th row, change of vegetation in 14th row. Our model can produce relatively complete CD maps compared with other models.

C. Ablation Study

In this section, several ablation studies are conducted to demonstrate the effectiveness and rationality of our model, including validation on different network components and different loss function.

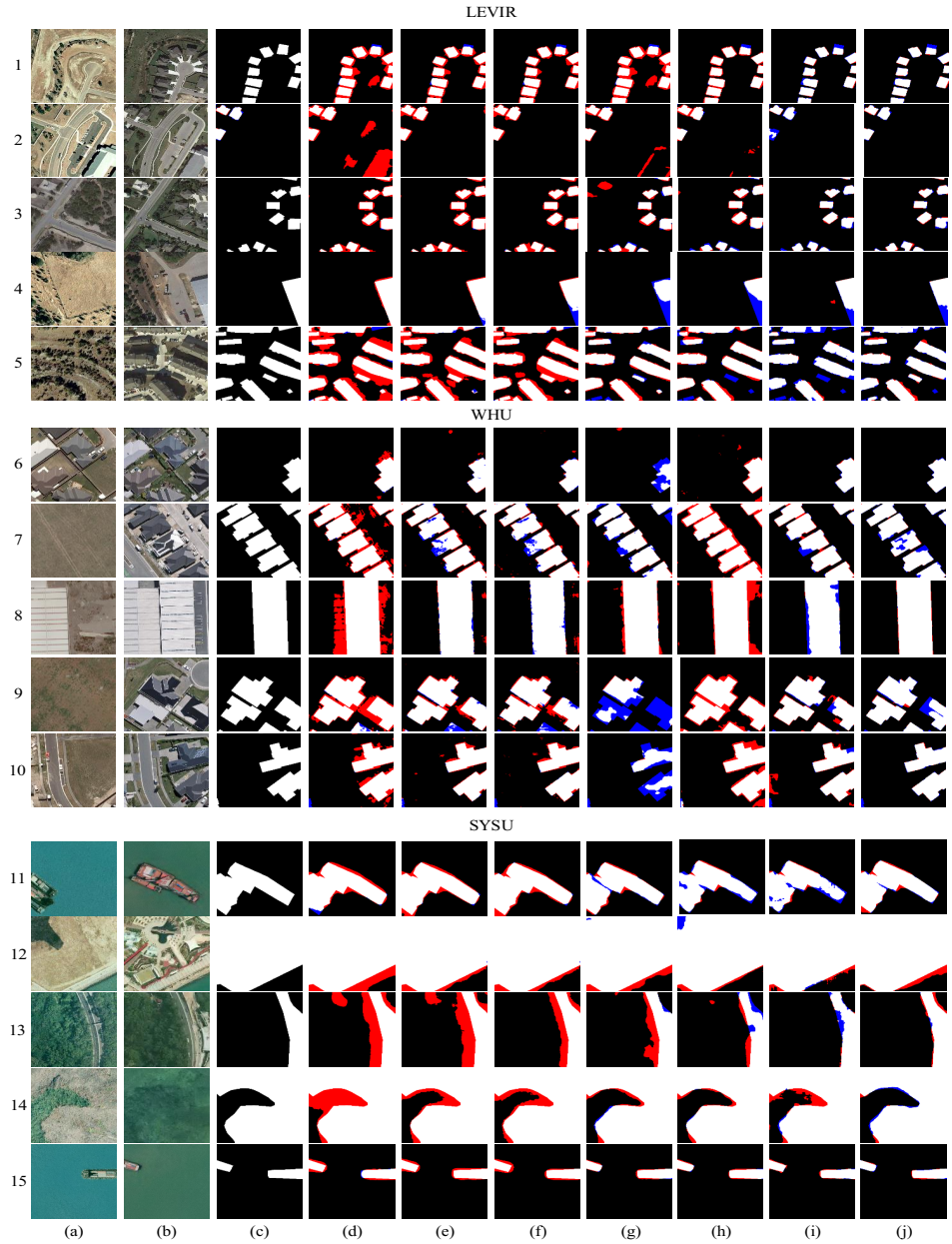

Fig. 5. Visual Comparison of RaSRNet and the state-of-the-art models on three datasets. (a) Time t_1 RSIs. (b) Time t_2 RSIs. (c) Ground Truth. (d) FC-EF. (e) FC-Siam-Conc. (f) FC-Siam-Diff. (g) DSAMNet. (h) STANet. (i) SRNet (ours). (j) RaSRNet (ours). White represents true positive, black represents true negative, red represents false positive, and blue represents false negative.

(1) Validation of different network components.

Different combinations of components, that is SRNet, SRa and CRa in RaSRNet, generate 3 variant networks. Their loss functions are still as shown in formula (20). The first network is SRNet, which has only an encoder-decoder backbone. The second network is called "SRNet+SRa". It is the architecture that the SRNet equipped with SRa, performs only spatial relation-aware on encoded features $\mathbf{F}_{e,k}$ $(k = 1, 2, 3)$. Therefore, the inputs of the decoder in "SRNet+SRa" are $\mathbf{F}^{s}_{ra,k}$ and $\mathbf{F}_{e,4}$. The third network is called "SRNet+CRa". Similarly, the inputs of the decoder "SRNet+CRa" are $\mathbf{F}^{c}_{ra,k}$ and $\mathbf{F}_{e,4}$. RaSRNet can be denoted as "SRNet+SRa+CRa". We conduct a series of experiments, and the quantitative and qualitative results on testing dataset are provided in Table.III and Fig.6, respectively.

From Table.III, it can be found that the addition of different components obviously improves the model performance. SRa is added to SRNet brings the relative gains of F1, KC, mIOU are about 2.24∼5.76%, 2.30∼6.03% and 1.63∼4.56% on three datasets. And then, CRa is independently added to SRNet to obtain relative gains of F1, KC, mIOU are about 1.94∼3.21%, 1.60∼3.33% and 1.04∼2.65%. Further, SRa and CRa are simultaneously introduced into SRNet will obtain the best performance. From Fig.6, it can be seen that

SRNet is more disturbed and does not have the ability to learn global information. In contrast, SRa captures more detailed information and obtains more complete changed objects (e.g., the building on the right side of the 2nd row and the bare ground in the 3rd row). CRa also has a more obvious contribution to SRNet, but other interferences are falsely detected as changed objects in the example in the 2nd row. In comparison, SRNet equipped with SRa and CRa (i.e., RaSRNet) can accurately locate the objects and fully suppress the backgrounds. Changed objects output by RaSRNet are endowed with clear boundaries. This proves that the combination of SRa and CRa is beneficial for learning of objects in complex scenes.

TABLE III
ABLATION STUDIES ABOUT THE COMPONENTS OF RASRNET.

Model	LEVIR			WHU			SYSU		
	F1	KC	mIOU	F1	KC	mIOU	F1	KC	mIOU
SRNet	.8567	.8498	.8677	.8106	.8032	.8333	.7390	.6695	.7271
SRNet+SRa	.8791	.8728	.8859	.8682	.8635	.8789	.7619	.6935	.7434
SRNet+CRa	.8888	.8831	.8942	.8363	.8302	.8532	.7584	.6855	.7375
SRNet+SRa+CRa	**.8943**	**.8889**	**.8990**	**.8936**	**.8900**	**.9002**	**.7832**	**.7200**	**.7624**

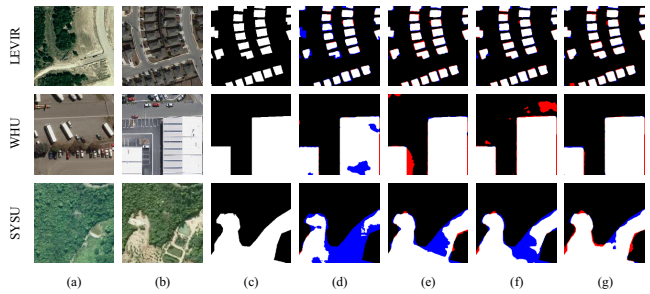

Fig. 6. Visual Comparison of the components of RaSRNet. (a) Time t_1 RSIs. (b) Time t_2 RSIs. (c) Ground Truth. (d) SRNet. (e) SRNet+SRa. (f) SRNet+CRa. (g) SRNet+SRa+CRa.

(**2**) **Validation of different loss components**.

We organize ablation studies that training RaSRNet with different loss functions. Specifically, 1) Only CD loss. 2) The combination CD loss and semantic reasoning loss, i.e., $L_c + \alpha L_s$. 3) The combination CD loss and deeply-supervised loss, i.e., $L_c + \beta L_d$. 4) The total loss is the sum of the three terms, as described by formula (20). The hyperparameters α and β are still set as described in Table.I. The quantitative results on the three datasets are shown in Table.IV. We can clearly see that the sum-of-three total loss trained RaSRNet performs the best on three datasets. Compared with RaSRNet trained by L_c, the introduction of L_s or L_d enable significantly improve the model performance. This effect is more obvious on the SYSU. Therefore, the results in Table.IV can firmly prove the rationality of the design of our loss function. It also confirms that imposing semantic constraints on unchanged pixels and constraints on underlying encoded features play an important role in model performance.

D. Sensitivity Analysis of Hyperparameters

In RaSRNet, the hyperparameters α and β represent the contribution of semantic reasoning loss and deeply-supervised loss to the total loss. Considering that the settings

TABLE IV
ABLATION STUDIES ABOUT THE LOSS COMPONENTS OF RASRNET.

Components			LEVIR			WHU			SYSU		
L_c	L_s	L_d	F1	KC	mIOU	F1	KC	mIOU	F1	KC	mIOU
✓			.8751	.8686	.8825	.8662	.8614	.8772	.7096	.6328	.7026
✓	✓		.8877	.8820	.8933	.8861	.8824	.8940	.7619	.6935	.7434
✓		✓	.8864	.8806	.8922	.8720	.8674	.8820	.7719	.7082	.7540
✓	✓	✓	**.8943**	**.8889**	**.8990**	**.8936**	**.8900**	**.9002**	**.7823**	**.7200**	**.7624**

of α and β will produce influence on the experimental results, we conducted experiments with different parameters to analyze how they affect the performance of RaSRNet. After many attempts, we set the range of α and β to [0.1, 0.2, 0.3]. Therefore, we conducted 9 sets of crossover experiments for each dataset. Fig.7 shows the performance of RaSRNet under different α and β. RaSRNet has a tendency to obtain the best performance when α=0.2 and β=0.3 on LEVIR dataset. Only when α=0.1 and β=0.3, the model performance is significantly different from other settings. In the WHU dataset, there are significant fluctuations when different hyperparameters are acted to RaSRNet. The model performs best only when α=0.1 and β=0.2. In the SYSU dataset, RaSRNet shows a significant outperformance over the other cases when α=0.1 and β=0.3. Based on the aforementioned analysis, α and β settings for each dataset are summarized in Table.I.

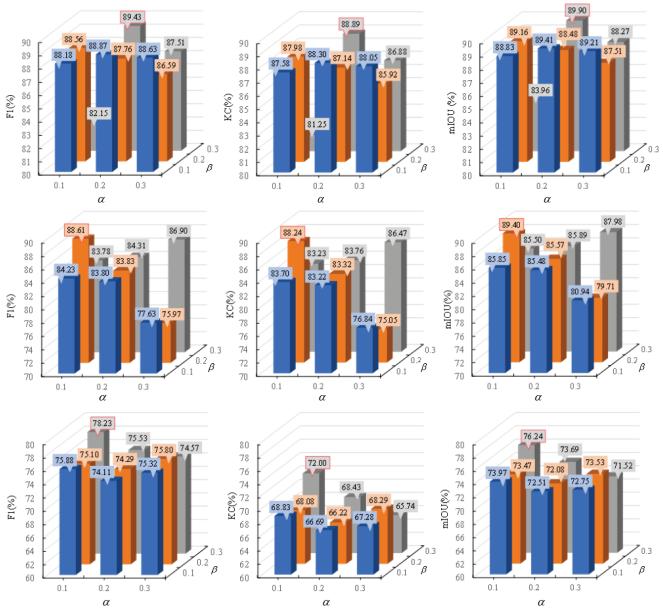

Fig. 7. Comparison of quantitative results about the performance of RaSRNet under different hyperparameters α and β.

V. DISCUSSION

A. Ra Module Feature visualization

In order to better observe the effect of the Ra module on global information collection, we use the Grad-Cam method to visualize the input and output features of the Ra module, and reshape them as 256×256. Fig.8 shows some visualization examples, including low-level single-scale encoded feature maps (i.e., input feature of Ra), SRa-enhanced feature maps and Ra-enhanced feature maps. Since

the CRa module works in the channel dimension, we do not provided CRa-enhanced feature maps. From Fig.8, we can see that Ra can extract salient objects in RSI and suppress background information that is irrelevant to changing semantics. For example, the left cases of Fig.8(a) and Fig.8(b), their ground truths both focus on changed building, and the Ra module obviously suppresses the "road" and "vehicle" in RSIs. Comparing the SRa feature maps and that of Ra, we find that changed objects extracted by Ra have high internal consistency (such as Fig.8(b) and the left case in Fig.8(c)). Taken together, it is shown that our Ra module can improve the overall recognition of changed objects.

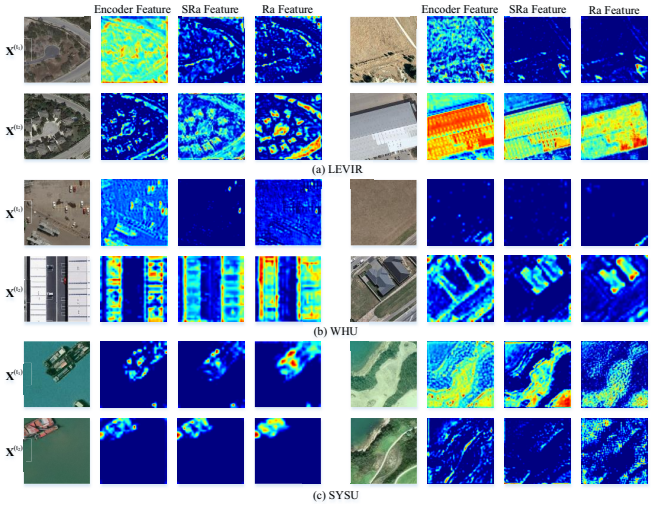

Fig. 8. Visualization feature maps of the Ra module.

B. Model intermediate Feature visualization

We provide the feature activation maps of RaSRNet in each period using the test sample of LEVIR as an example. They are shown in Fig.9. In order to highlight the changed objects, we mask the ground-truth over the dual optical RSI, as the areas surrounded by the red solid lines of $\mathbf{X}^{(t_1)}$ and $\mathbf{X}^{(t_2)}$. Siamese encoder generates 4-scales feature maps $\{\mathbf{F}^{*}_{e,k}\}^{4}_{k=1}$ $(* \in \{t_1, t_2\})$. Then the Ra modules act on $\mathbf{F}^{*}_{e,1}$, $\mathbf{F}^{*}_{e,2}$ and $\mathbf{F}^{*}_{e,3}$ respectively to learn rich global context information and generate the refined features $\mathbf{F}^{*}_{ra,1}$, $\mathbf{F}^{*}_{ra,2}$ and $\mathbf{F}^{*}_{ra,3}$. Comparing $\{\mathbf{F}^{(t_1)}_{e,k}\}^{3}_{k=1}$ and $\{\mathbf{F}^{(t_1)}_{ra,k}\}^{3}_{k=1}$, we can see that our model learns more details of changed objects (buildings) in $\mathbf{X}^{(t_2)}$. In addition, the high-level features $\mathbf{F}^{*}_{e,4}$ are preserved, thus the semantic concept associated with unchanged is further highlighted in decoder features. The difference feature is generated by $|\mathbf{F}^{(t_1)}_{d,1} - \mathbf{F}^{(t_2)}_{d,1}|$. Finally, the CD head filters the change-independent noise in difference feature and generates a change probability map.

C. Failure Cases and Future Work

(**1**) **Failure cases**. Fig.10 shows some of these failure cases. For some special and challenging scenarios, proposed method produces less satisfactory CD maps. (1) It is still challenging for scenes where buildings and public ground have low contrast. For example, as shown in the lower

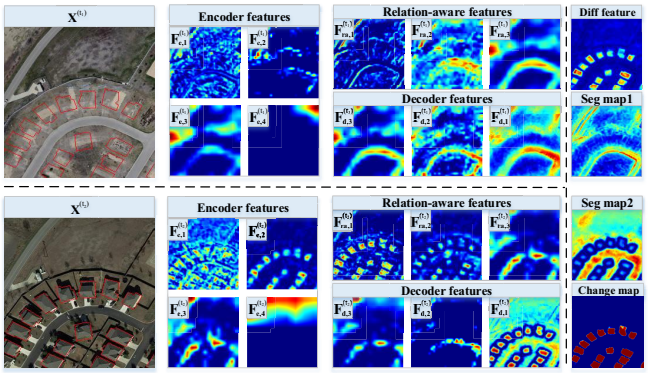

Fig. 9. Visualization feature maps of the RaSRNet. $\mathbf{X}^{(t_1)}$ and $\mathbf{X}^{(t_2)}$ are the input RSIs at time t_1 and t_2.

right of the 1st row in Fig.10, the public ground and the surrounding buildings are mistaken for a single large building. This is because the roof material of the building shows a similar color to the public ground, making it difficult for model to accurately learn the building boundaries. (2) It is still challenging for scenes where shadows are wrapped in changed areas with complex morphological. For example, as shown in the lower left of the 2nd row in Fig.10, the shadows are embedded between the column building and another regular building and are mistaken for changed areas. (3) It is still challenging to suppress the disturbances of natural growth of vegetation. For example, the 3rd row in Fig.10, bare ground change was not accurately detected due to interference from local morphological and color changes in vegetation.

(**2**) **Future work**. In the future, there are two areas where future optimizations can be made. First, our method is data-driven, discriminative features from training samples may not be so effective in a few complex scenes affected by certain factors (e.g., low contrast between changed area and background, biological seasonality). Only a deeply supervised strategy is used to constrain the network to learn underlying features (including edge, textures). Some papers show the better CD maps by introducing an edge prior in building change detection, but it has not been validated in other change types [38]–[40]. Therefore, the rational use of edge information and its role in various scenes is worthy of future study and verification. Second, based on the results in the section IV, proposed method performs better on the LEVIR and WHU datasets than on the SYSU dataset. This may be because the large range of change areas and few change instances in the SYSU dataset. For such scenes, the high-level features (i.e., $\mathbf{F}_{e,4}$) may have a weak or even negative impact on the identification of change categories. While proposed method does not screen ineffective high-level features, which is due to its too high channel dimensionality and requires more computational cost. Therefore, detecting large-range change areas and balancing calculated expenditures would also be of greater significance.

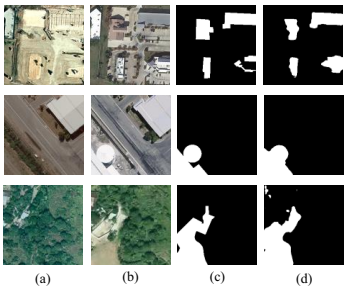

Fig. 10. Visual display of the failure cases. (a) Time t_1 RSIs. (b) Time t_2 RSIs. (c) Ground Truth. (d) CD maps generated by RaSRNet.

VI. Conclusion

In this article, we focus on performing CD in optical RSIs and propose an end-to-end RaSRNet, which can effectively aware global context information and recover the details of changed objects. The SRNet is developed to extract multi-scale features of ground objects in a propensity, and then integrate and reconstruct multi-scale features to identify objects with different semantics, which can effectively filter the background information and complete the accurate detection of target changed objects. The relation-aware module is proposed to model the semantic relationship between objects in the encoded features from the spatial dimension and to model the correlation between feature channels from the channel dimension, to ensure the integrity recognition of objects in RSIs. Following this way, the proposed RaSRNet enables generate high-quality CD map, which can display changed objects more completely and accurately. Multiple experiments were conducted on three public RSIs CD datasets, and the results firmly demonstrate the effectiveness and superiority of RaSRNet.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62173063.

References

[1] G. Liu, Y. Gousseau, and F. Tupin, "A contrario comparison of local descriptors for change detection in very high spatial resolution satellite images of urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3904–3918, 2019.

[2] Y. Hao, Z. Chen, Q. Huang, F. Li, B. Wang, and L. Ma, "Bidirectional segmented detection of land use change based on object-level multivariate time series," *Remote Sens.*, vol. 12, no. 3, 2020.

[3] D. A. Jimenez-Sierra, H. D. Benítez-Restrepo, H. D. Vargas-Cardona, and J. Chanussot, "Graph-based data fusion applied to: Change detection and biomass estimation in rice crops," *Remote Sens.*, vol. 12, no. 17, 2020.

[4] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 7, 2022.

[5] 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.

[6] C. Chen, C. Li, D. Li, Z. Zhao, and J. Hong, "Mechanical assembly monitoring method based on depth image multiview change detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[7] X. Tang, H. Zhang, L. Mou, F. Liu, X. Zhang, X. X. Zhu, and L. Jiao, "An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[8] Q. Xu, K. Chen, G. Zhou, and X. Sun, "Change capsule network for optical remote sensing image change detection," *Remote Sens.*, vol. 13, no. 14, 2021.

[9] 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.

[10] 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.

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

[12] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion ssd for remote sensing object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[13] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, 2020.

[14] Q. Ding, Z. Shao, X. Huang, and O. Altan, "Dsa-net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 105, p. 102591, 2021.

[15] J. Dong, W. Zhao, and S. Wang, "Multiscale context aggregation network for building change detection using high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "Rrnet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[17] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 433–442.

[18] Y. Su, J. Cheng, W. Wang, H. Bai, and H. Liu, "Semantic segmentation for high-resolution remote-sensing images via dynamic graph context reasoning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

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

[20] F. Huo, X. Zhu, Q. Zhang, Z. Liu, and W. Yu, "Real-time one-stream semantic-guided refinement network for rgb-thermal salient object detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[21] Q. Li, R. Zhong, X. Du, and Y. Du, "Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[22] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *2015 IEEE/CVF Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[24] 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.

[25] X. Zhang, Y. Yue, W. Gao, S. Yun, Q. Su, H. Yin, and Y. Zhang, "Difunet++: A satellite images change detection network based on unet++ and differential pyramid," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[26] 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 Obs. Geoinf.*, vol. 103, p. 102465, 2021.

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

[28] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[29] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020.
[30] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[31] M. Gong, F. Jiang, A. K. Qin, T. Liu, T. Zhan, D. Lu, H. Zheng, and M. Zhang, "A spectral and spatial attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[32] Y. Yang, J. Qu, S. Xiao, W. Dong, Y. Li, and Q. Du, "A deep multiscale pyramid network enhanced with spatial-spectral residual attention for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[33] Y. Ding, Y. Guo, Y. Chong, S. Pan, and J. Feng, "Global consistent graph convolutional network for hyperspectral image classification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–16, 2021.
[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.
[35] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 129–150, 2011.
[36] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8947–8956.
[37] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, 2019.
[38] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[39] B. Yang, Y. Huang, X. Su, and H. Guo, "Maeanet: Multiscale attention and edge-aware siamese network for building change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 19, 2022.
[40] Z. Chen, Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin, "Egde-net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 203–222, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271622001940

Yi Liang (梁漪) received the B.S. degrees in measurement and control technology and instrument from the Liaoning Technical University, LiaoNing, China, in 2016, where she is pursuing the Ph.D. degree in electronics and information with Dalian University of Technology, Dalian, China.

Her research interests include remote sensing image change detection and deep learning.

Chengkun Zhang (张成坤) received the B.S. degrees in automation from the Ocean University of China, Qingdao, China, in 2013, and the Ph.D. degree in control theory and control engineering from Dalian University of Technology, Dalian, China, in 2021.

He is a Lecturer with the Qinghai University, Xining, China. His research interests include feature extraction and classification of hyperspectral images.

Min Han (韩敏) received the B.S. and M.S. degrees from the Department of Electrical Engineering, Dalian University of Technology, Dalian, China, and the M.S. and Ph.D. degrees from Kyushu University, Fukuoka, Japan, in 1982, 1993, 1996, and 1999, respectively. She is currently a Professor with the Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology. She serves as a deputy director of the Chinese society of instrumentation youth work committee, a committee member of the Chinese Society of Artificial Intelligence, and an Organizing Chair of ISNN2013, ICICIP 2014, ICIST2016.

Her current research interests include remote sensing image interpretation, complex system modeling and forecasting method, time series analysis and forecasting, and neural networks.