WSMsFNet: Joint the Whole Supervision and Multiscale Fusion Network for Remote Sensing Image Change Detection

Bin Wang, Xiaohu Jiang, Pinle Qin, and Jianchao Zeng

***Abstract*—Remote sensing image change detection aims to extract high-level semantic feature to identify the changed areas (CAs) between dual-temporal images (DTIs). However, the diversity in the CA shape and size poses certain challenge to the change detection (CD) task. Besides, different illumination conditions in the same scene of the DTI further increase the CD difficulty. In response to these above issues, this article proposes a multiscale feature fusion CD network-WSMsFNet, which fully utilizes the local and global information of multiscale features to achieve comprehensive representation of the change scene. In addition, the network improves the feature extraction ability of each module through the whole process supervision loss function. First, the network hierarchically extracts different scale information of the two temporal RS. Then, special information enhancement and fusion modules are constructed for various feature layers (i.e., the same level, adjacent level, and global features), aiming to enhance the local feature representation ability and contextual information relevance of the deep network. Finally, the whole-process loss function is set to supervise the intermediate layer learning, which can effectively enhance the feature representation ability and guide feature extraction direction of each module. Experiments have shown that the WSMsFNet has achieved significant results in both qualitative and quantitative indicators.**

***Index Terms*—Change detection, convolutional neural network (CNN), multiscale feature (MSF) fusion, the whole process supervision, transformer.**

I. Introduction

REMOTE sensing image change detection (RSCD) aims to understand the differences between images of the same area at different times and mark these changed areas (CAs) [1]. It has been widely used in fields such as urban expansion analysis, disaster assessment, military strikes, and vegetation cover detection [2], [3], [4], [5]. At present, the change detection (CD) field is experiencing the following challenges, as shown in Fig. 1: 1) the shooting conditions in different periods (such as illumination, season, etc.) may cause varying degrees of interference to intelligent models (even humans) in judging whether the scene has changed; and 2) the change range about CA attributes such as the size, shape, and quantity is relatively large. Therefore, this article intends to study the CD network based on multiscale feature (MSF) fusion, aiming to extract high-level semantic feature (HLSF) to bridge the semantic gap caused by noise disturbances such as lighting, and achieve accurate identification for interest change regions.

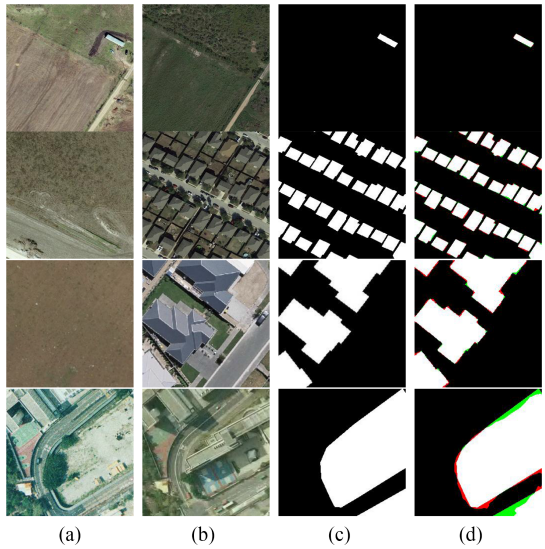


Fig. 1. Multiscale change regions and CD results of DTIs. According to (a) and (b), light disturbance does increase the difficulty of RSCD. Therefore, deep semantic information needs to be extracted to achieve high-precision change identification. According to (c) and (d), the shape and size fluctuation range of RSCD is indeed significant, involving multiple scale features. Besides, it can be seen from lines 2–3 that some CD regions show certain correlation in spatial context, reflecting the importance of global features in CD. (a) T_1. (b) T_2. (c) GT. (d) Ours.

Traditional CD methods can be roughly divided into three categories [6]: 1) image arithmetic-based; 2) image transformation-based; and 3) postclassification methods. Image

Manuscript received 30 May 2024; revised 13 July 2024; accepted 30 July 2024. Date of publication 7 August 2024; date of current version 26 August 2024. This work was supported in part by the State Key Laboratory of Resources and Environmental Information System, in part by the Basic Applied Research Projects of Shanxi Province in China under Grant 20210302124165 and Grant TZLH20230818007, in part by the Ministry of Education Industry-University Cooperative Collaborative Education project under Grant 221002722143739, in part by the Natural Science Foundation of Hainan Province under Grant 422QN350, in part by the National Natural Science Foundation of Youth Science Foundation Project under Grant 42001360, in part by Shanxi Province Graduate Education Innovation Plan under Grant 2024SJ277 and Grant 2024SJ278, and in part by Research Project supported by the Shanxi Scholarship Council of China under Grant 2024-118. *(Corresponding authors: Pinle Qin.)*

Bin Wang and Xiaohu Jiang are with the Department of Computer Science and Technology, North University of China, Taiyuan 030051, China, and with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

Pinle Qin and Jianchao Zeng are with the Department of Computer Science and Technology, North University of China, Taiyuan 030051, China (e-mail: qpl@nuc.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3439991

arithmetic-based methods include image differencing, image ratioing, Change Vector Analysis, etc. They usually first uses relevant arithmetic (e.g., subtraction or division) to obtain feature maps, and then distinguish CA and unchanged area (UA) information through segmentation thresholds. Obviously, segmentation threshold may be the difficulty and key point for such methods. Image transformation-based methods include Principal Component Analysis, Multivariate Alteration Detection, etc. They transform these images into specific feature spaces, and improve the CD accuracy by highlighting the CA and suppressing the UA. The postclassification methods first classify the target objects of DTIs, and then compare and analyze the classification results to generate CA. Obviously, that accuracy depends on the classification accuracy. Therefore, the cumulative error effect gradually becomes more severe.

The development of machine learning has improved the CD effect to a certain extent [7], [8], such as supporting vector machine, random forest, etc. However, the selection of relevant methods and the model generalization ability cause challenges for practical applications when facing different scenarios.

Deep learning (DL) have been widely used in CD tasks and have shown good performance due to their powerful feature extraction and nonlinear representation [9], [10], [11]. Intrinsically, this is attributed to various deep networks that can break through the surface interference of change noise, mine HLSF about the interest region, and achieve intelligent target recognition.

According to the feature map extraction and fusion, the DL-based CD framework can be roughly divided into three categories [12], [13]: 1) early fusion; 2) late fusion; and 3) multilevel feature fusion (i.e., correlation fusion). The early fusion-based CD deep networks cascade the input DTIs at the beginning, and perform feature extraction, encoding, decoding, and other operations to obtain the final CD result. The late-stage methods often adopt a dual-stream structure to extract deep features of the original images, then perform feature fusion and pixel classification. However, the abovementioned methods clearly lack hierarchical interaction and global representation of MSF depth features, which is definitely not conducive to the extraction of HLSF for CD. Recently, more research efforts have been dedicated to extracting deep information through Siamese CNNs (SCNNs) and fusing corresponding layer features to determine CAs [14], [15]. This is known as the multilevel feature fusion-based CD method. Obviously, these nets can effectively integrate shallow features (e.g., target texture, corner, edge, etc.) and deep semantic information (e.g., image content and semantic concept), alleviating or even eliminating the semantic gap between various levels to improve CD performance.

At present, the common multilayer feature fusion can be roughly categorized into two ways: 1) same scale; and 2) multiple different scales synchronous fusion. Same-scale features share the most direct correlation in terms of pixel positions and are the simplest to realize, such as direct addition, subtraction, or cascading [16], [17], [18], [19]. However, this oversimplified fusion model is hardly able to bridge the semantic gap between different features to extract their effective information. For this reason, several studies have been conducted to recognize changes of interest occurring in dual-temporal RS images through CNN-based attention mechanisms (e.g., spatial, channel). To ensure feature consistency across bitemporal images, Peng et al. [20] made some modifications to the channel attention mechanism. Chen et al. [21] used spatial and channel graph convolution networks to effectively explore the dual temporal image features relationship. Although they all may learn the correlation between feature maps in a fixed neighborhood, the limited sampling scope makes it impossible to characterize long-distance information relations. Recently, some studies have achieved contextual modeling in both spatial and temporal scales through the Transformer, which effectively improves the global properties of RSCD. Luo et al. [22] filtered out redundant information through residual Transformer and focused on changing features. Mao et al. [23] fed multiple edge features to the Transformer and constructed a high-frequency cues guidance module. Huang et al. [24] designed dense cross spatial attention to capture long-range dense spatial interactions. For MSF fusion, researchers often realize them based on pyramids or U-Net and its variants [25], [26], [27], [28], [29]. However, the fusion approach that only consists of CNN still leads to its inability to develop global dependencies [30], [31]. In addition, the relevant features in the RSCD are not exactly the same but roughly similar, so the information embedded in the neighboring feature layers needs to be mined deeply. However, this element has been neglected by most research works. Feng et al. [32] combined spatial and channel attention to extract local correlation information from adjacent scale feature maps. In this article, we will explore the potential relationship among features and integrate their valuable information from the same-scale, adjacent-scale, and multiscale, and focus on the global-local association information modeling at each stage.

Furthermore, most DL-based CD networks only impose loss functions between the prediction and label ends, such as cross-entropy, dice loss, contrastive loss, and their simple linear combinations. However, many experiments have shown that multilevel supervision oriented towards the training process is beneficial for improving model convergence and detection performance [33], [34], [35]. Therefore, it is quite meaningful to conduct in-depth research on the entire supervision about the model training.

To address the aforementioned issues, we construct a hierarchical MSF fusion DL architecture for RSCD-WSMsFNet as shown in Fig. 2, aiming at extracting multiscale HLSFs to identify whether the scene has changed. Then, the same-level difference feature enhancement module (SLDFEM) and cross-scale adjacent level feature fusion module (CSALFFM) are proposed to improve contextual relevance of local features. Furthermore, WSMsFNet further extracts global dependencies between DTIs based on cross-attention mechanism by multiscale global feature fusion module (MSGFFM). Finally, the fully supervised loss function is deliberately designed for alleviating gradient vanishing and enhancing the recognition power of module feature extraction.

The main contributions of this article are as follows.

1) Targeting the uncertainty and underlying contextual relations of CAs in bitemporal RS images, this article proposes the CD deep network WSMsFNet based on MSF fusion and Transformer variants.

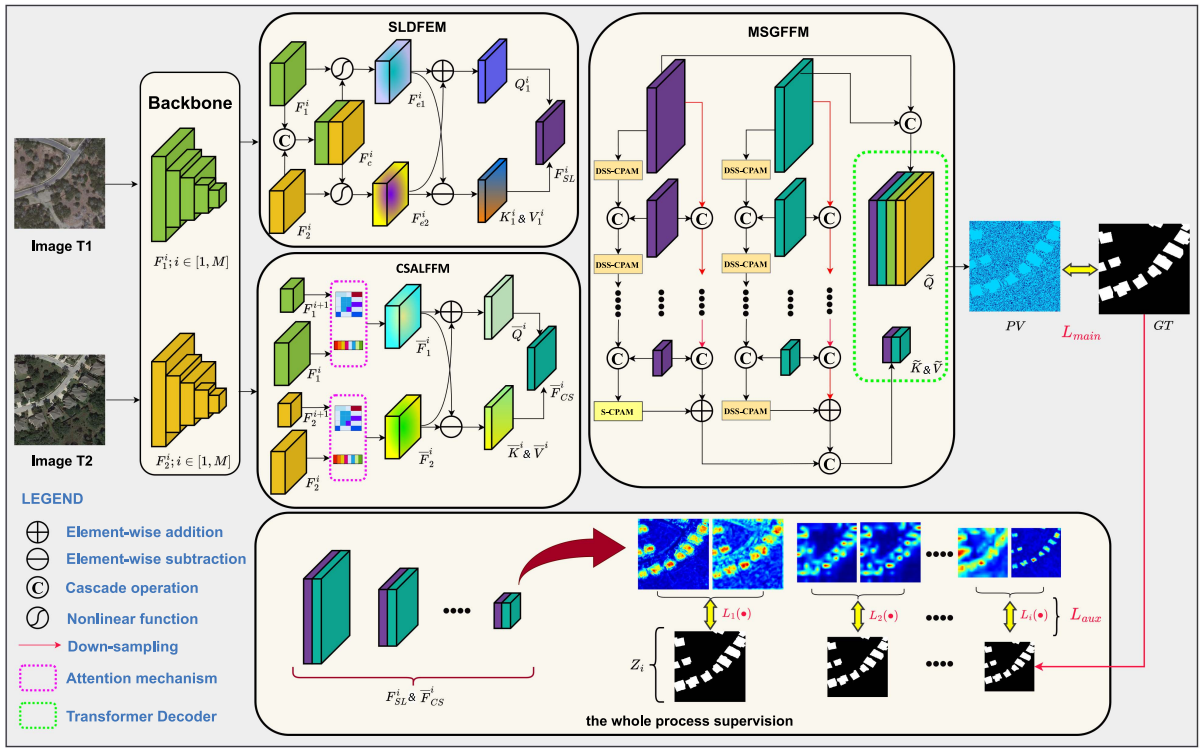


Fig. 2. Overall architecture of the WSMsFNet. The SLDFEM is designed to enhance and fuse effective information of the same-layer features with consistent size. The CSALFFM is proposed to promote the correlation between neighboring feature map. The MSGFFM has effectively achieved long-distance relationship modeling for various depth features through transformer. Those outputs of the above three modules are compared with the labels obtained by downsampling GT to achieve full process supervision.

2) To achieve comprehensive MSF merging, we have developed modules for same scale layer, adjacent variable scale layer, and multiple different scale layers to implement feature integration and strengthening. Moreover, each feature convergence synchronously takes into account local feature extraction and contextual global modeling to guarantee the effectiveness of feature refinement.
3) To ensure the correctness and efficiency of feature learning, the whole-procedure supervision paradigm has been utilized to orient the model main intermediate feature extraction, which is designed to achieve a supervised training process for each multiscale fusion module learning by cropping the original labels.

The rest of this article is organized as follows. Section II briefly describes related work, Section III mainly describes the model architecture and details, Section IV focuses on the experiment, Section V provides a summary, and Section VI discusses the results and reflects on the implications and limitations of the work. Finally, Section VII concludes this article.

II. Related Work

A. CNN-Based CD Models

With the advancement in HRRS, the fine and complex texture of land objects pose new challenges for traditional CD methods. Fortunately, the excellent big-data processing capability of DL provides the possibility to solve this problem, and its most widely used is no different from CNNs.

Numerous researchers have developed CD models based on the Full Convolutional network and Unet, such as FC-EF [36], FC-Siam-conc [36], FC-Siam-diff [36], EUNet-CD [37], MTL-CD [38], etc. In addition, another mainstream of CNN-based CD networks is Siamese networks, such as BESNet [39], ECFNet [40], and MCDnet [41], etc. These models employ the dual-stream architecture to extract deep features from dual-temporal remote sensing image (DTRSI) and strengthen relevant features to obtain CD results.

To explore the temporal correlation of feature tensors, some researchers have proposed CD models incorporating Long Short Term Memory networks, such as EGRCNN [42] and ML-EDAN [43], etc.

Notably, to enable CNNs can accurately focus on important features and suppress anothers, researchers have proposed CD networks based on attention mechanisms, including channel, spatial, and self-attention mechanisms, such as EUNet-CD [37], FHD [44], and ASGF [45], etc. These methods mainly achieve feature reweighting in various ways (e.g., channel, spatial, and correlation) to highlight informative features and ultimately enhance the CD performance.

B. Transformer-Based CD Models

Because of its ability to model global context information excellently, the Transformer has been rapidly introduced into computer vision from natural language processing, including image classification [46], [47], segmentation [48], object

detection [49], [50], super-resolution [51], denoising [52], video detection [53], and tracking [54], [55].

Correspondingly, the Transformer has also attracted extensive research attention about CD Networks (e.g., ChangeFormer [56] and ICIF-Net [32], etc.) utilized Transformer as a backbone for extracting global features from the originals. Song et al. [57] proposed a multiscale Swin transformer supervised network (MSTDSNet) for monitoring urban land changes using DTHRRS. Chen et al. [58] designed a DTI transformer (BIT) to efficiently model contexts within the spatial-temporal domain.

In summary, the complementary advantages of CNN and Transformer help extract more comprehensive and identifiable information [59], [60]. Inspired by this, we will continue to explore the potential of Transformer in RSCD.

C. Loss Functions

Loss function plays a crucial role in continuously improving model by measuring the difference between the prediction and the ground truth (GT). Commonly, the Loss function of CD models includes cross entropy, contrast loss, dice loss, and their linear weighted combination.

Besides the aforementioned Loss functions, researchers also attempted to apply relevant improvements for specific challenges. To encourage category related features, Sun et al. [38] constrained CD task by the auxiliary semantic segmentation loss function. Feng et al. [32] minimized the cross-entropy loss between three prediction heads to get local and global characteristics. Lei et al. [39] used a new boundary extraction loss combined with the contractive loss function to optimize the BESNet. ML-EDAN [43] is trained in an end-to-end manner with a new joint loss function considering both reconstruction error and pixel-wise classification error.

In general, the loss function is placed at the network end. To make these intermediate features more representative, some researchers have explored multilayer supervised learning. SCDTN [61] employed cross-entropy to supervise feature learning of two subnetworks, effectively enhancing the module's feature extraction capability. Bandara and Patel [62] minimized their unsupervised "similarity-dissimilarity loss," where they simultaneously maximized the distance between CA while minimizing the distance between UA in DTIs and deep features.

III. Method

Due to differences in imaging conditions such as season, lighting, sensor, and the complexity of the land surface, objects with the same semantic information in DTRSI often exhibit various features. To bridge such semantic gap, this article proposes a RSCD deep network framework to achieve robust recognition of HLSF about complex scenes. The model aggressively utilizes the local feature extraction of CNNs and the long-distance relationship modeling of Transformer.

A. Overall Architecture

The pipeline of WSMsFNet is shown in Fig. 2. Intuitively, it adopts the SCNN structure, consisting of a feature extraction network and three auxiliary modules (i.e., SLDFEM, CSALFFM, and MSGFFM).

Algorithm 1: Inference of WSMsFNet-based Model for CD.

Input: The DTIs \boldmath $\{T_1, T_2\} \in \mathbb{R}^{H\times W\times 3}$
Output: PV (a prediction change mask)
//step1: Extract MSFs by a SCNN backbone (M-layers)
for $i \in \{1, 2\}$ **do**
  for $j \in \{1, M\}$ **do**
    $F_i^j = \text{SCNN\_Backbone}\{T_i\}$
  end for
end for
//step2: Use SLDFEM to integrate same-level features
for $i \in \{1, M\}$ **do**
  $F_{SL}^i = \text{SLDFEM}\{F_1^i, F_2^i\}$
end for
//step3: Enhance neighborhood relation via CSALFFM
for $i \in \{1, M-1\}$ **do**
  $\overline{F}_{CS}^i = \text{CSALFFM}\{F_1^i, F_1^{i+1}, F_2^i, F_2^{i+1}\}$
end for
//step4: Implement global context information modeling
for $i \in \{1, M\}$ **do**
  $F_o = \text{MSGFFM}\{F_{SL}^i, \overline{F}_{CS}^i, F_1^1, F_2^1\}$
end for
//step5: Obtain change mask by the prediction head
$PV = \text{Prediction\_head}\{F_o\}$

The WSMsFNet frame can be summarized as Algorithm 1.

First, the DTIs $T_1, T_2 \in \mathbb{R}^{H\times W\times 3}$ are fed into the SCNN for feature extraction, parallelly attaining original multiscale local features $\{(F_1^i, F_2^i); i \in [1, M]\}$ through the backbone (e.g., Resnet34), where $\{F_1^i, F_2^i\} \in \mathbb{R}^{H/2^i\times W/2^i\times C_b^i}$.

Then, the paired features in ith stage $\{F_1^i, F_2^i\}$ are, respectively, input into SLDFEM, and their salient and differential features are further extracted as the transformer's Q, K and V, respectively. This operation forms new feature representation $F_{\text{SL}}^i \in \mathbb{R}^{H/2^i\times W/2^i\times C_s^i}$, which could establish global dependency relationship on local features from CNN.

Simultaneously, these features $\{F_1^i, F_1^{i+1}, F_2^i, F_2^{i+1}\}$ are fed into CSALFFM for multiscale aggregating information $\overline{F}_{\text{CS}}^i \in \mathbb{R}^{H/2^i\times W/2^i\times C_c^i}$, which is conducive to develop the context correlation between neighboring features.

Next, the transformer is exploited to refine and decode the aforementioned enhanced MSFs. The output of transformer decoder $F_o \in \mathbb{R}^{H/2^M\times W/2^M\times C_o}$ is the tensor prototype of CD mask.

Finally, the predicted value ($PV \in \mathbb{R}^{H\times W\times 2}$) is generated by the detection head. It is worth noting that the midresults of involved modules are all supervised by multiscale labels, which derives from downsampling to GT.

B. Illustration of the Module SLDFEM

In general, the CD difference map is obtained by directly subtracting the DTIs, and then drawing discernible features through

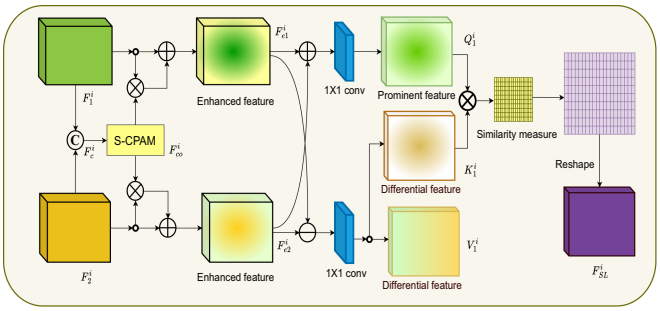


Fig. 3. Overall structure diagram of the module SLDFEM.

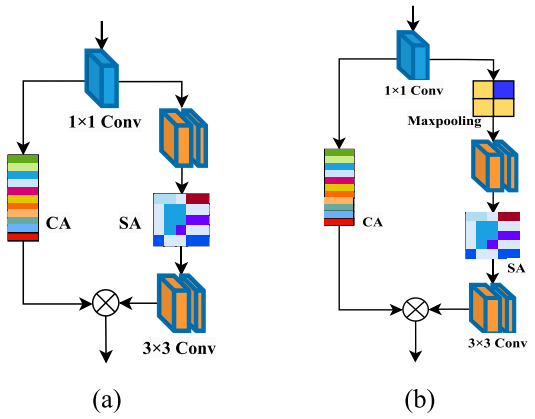


Fig. 4. Illustration of attention modules: (a) S-CPAM. (b) DS-CPAM.

a series of convolution, pooling, or attention mechanisms, etc. However, interferences (e.g., illumination, shadows, and UA noise) may lead to the erroneous accumulation and hinder the discrimination of CD targets. In addition, the SCNN backbone extracts HLSFs from the DTIs by weight-sharing. Therefore, the UAs in the same-level map pairs are expected to show the same semantic information. However, due to noise and other factors, some UAs exhibit significant pixel-level differences and interfere with the final CD.

To address the abovementioned issues, we propose SLDFEM to improve the quality of local differential feature, as shown in Fig. 3. The SLDFEM first cascades the input through spatial and channel perspectives and extracts the effective information from the originals. Then, differential and salient features are obtained through element level addition and subtraction, respectively. At last, the long-distance dependency relationship of peer features is modeled via cross self-attention.

For input features F_1^i and F_2^i, SLDFEM cascades and feeds them into the space-channel parallel attention module (S-CPAM) to select and focus on important deep information. As depicted in Fig. 4(a), the S-CPAM cascades F_1^i and F_2^i into convergent feature F_c^i and performs feature aggregation through 1x1 convolution. Then, the dimension-reduced result undergoes a 3×3 convolution, spatial attention, and another 3×3 convolution to obtain spatially enhanced features F_{sa}^i. Simultaneously, $f_{d=1}^{1*1}(F_c^i)$ is passed via channel attention and multiplied with F_{sa}^i to yield the output F_{co}^i, representing spatial and channel-wise consistent features.

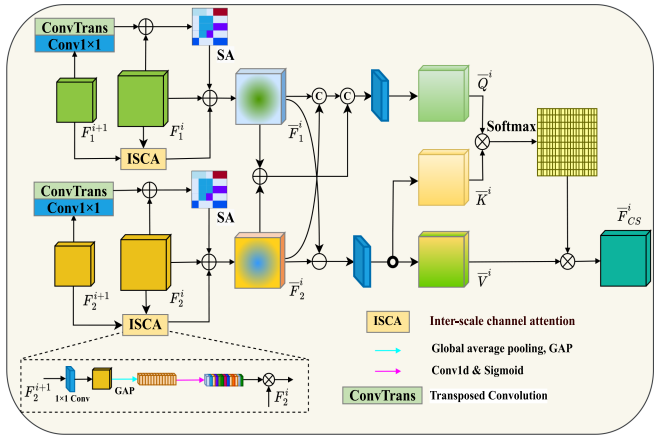


Fig. 5. Overall structure diagram of the module CSALFFM.

The entire S-CPAM can be represented by (1)

$$g_1(\mathcal{X}) = CA\left(f_{d=1}^{1*1}(\mathcal{X})\right) \odot \left\{f_{d=1}^{3*3}\left(SA\left(f_{d=1}^{3*3}(\mathcal{X})\right)\right)\right\} \quad (1)$$

where $\mathcal{X}$ represents a tensor. For $F_{co}^i = g_1(F_1^i \copyright F_2^i)$, multiplication and residual-connections are performed with F_1^i and F_2^i separately. Then, it would result in deeply enhanced related features F_{e1}^i and F_{e2}^i. To explore the global correlation of the same-level features, SLDFEM performs self-attention on the feature-enhanced F_{co}^i. As shown in Fig. 3, element-wise operations (i.e., addition and subtraction) are applied to F_{e1}^i and F_{e2}^i, respectively, to obtain the prominent and differential information about DTIs. It is worth noting that in the design of self-attention, the prominent feature is regarded as as Q, and the differential feature F_{e1}^i-F_{e2}^i is used as K and V. That aims to match the key information of prominent and differential features to enhance the noteworthy differential value, namely the semantic CA.

The entire self-attention can be represented as follows:

$$\begin{cases} Q_1^i = f_{d=1}^{1*1}\left(F_{e1}^i + F_{e2}^i\right) W_{q1}^i \\ K_1^i = f_{d=1}^{1*1}\left(F_{e1}^i - F_{e2}^i\right) W_{k1}^i \\ V_1^i = f_{d=1}^{1*1}\left(F_{e1}^i - F_{e2}^i\right) W_{v1}^i. \end{cases} \quad (2)$$

The final output $F_{\mathrm{SL}}^i = Att(Q_1^i, K_1^i, V_1^i)$, whose details are shown as follows:

$$F_{\mathrm{SL}}^i = \sigma\left(\frac{Q_1^i K_1^i}{\sqrt{d}}\right) V_1^i. \quad (3)$$

C. Illustration of the Module CSALFFM

With the net-depth increase, the extracted features gradually transitions from low-level (such as edges, textures, etc.) to HLSF. Considering the RS complexity, enhancing the feature correlation is obviously beneficial for the model to understand the semantic information of CAs. Except for the same-level, the correlation between adjacent layer is the strongest throughout the feature extraction process. Inspired by this, we propose the CSALFFM to achieve potential valid information from neighbouring maps.

As shown in Fig. 5, the CSALFFM input consists of two sets of adjacent feature pairs (i.e.,$\{F_1^i, F_1^{i+1}; F_2^i, F_2^{i+1}\}$). First,

1×1 convolution and transposed convolution operations are performed on the deep features $\{F_1^{i+1}, F_2^{i+1}\}$. On one hand, the channel number is adjusted to be the same for the two layers. On the other hand, the size of adjacent feature maps is unified, while spatial attention mechanism is used to enhance their prominent features.

The entire process is shown as follows:

$$\overline{F}_{SA1}^{i} = SA\left\{\overline{f}\left(f_{d=1}^{1*1}\left(F_1^{i+1}\right)\right) + F_1^{i}\right\} \tag{4}$$

where $\overline{f}(\cdot)$ represents the transposed convolution operator.

Simultaneously, since deep features have the more wider receptive field, transforming them into channel weights can highlight the effective information of shallow features. Therefore, 1×1 convolution is applied to F_1^{i+1} to compress it to the same channel number as F_1^i, and global average pooling is used to fuse and compress the spatial dimension information. Then, linear operations and activation functions are applied to transform F_1^{i+1} into channel weights.

The entire process is shown as follows:

$$\overline{F}_{CA1}^{i} = gap\left(f_{d=1}^{1*1}(F_1^{i+1})\right) * F_1^{i}. \tag{5}$$

Finally, following the residual-connection paradigm, F_1^i is added to the spatial-related features and channel-related features described above, resulting in the fused adjacent layer features $\overline{F}_1^i$, which is shown as follows:

$$\overline{F}_1^{i} = F_1^{i} + \overline{F}_{SA1}^{i} + \overline{F}_{CA1}^{i}. \tag{6}$$

Similarly, $\overline{F}_{SA2}^{i}$, $\overline{F}_{CA2}^{i}$, $\overline{F}_2^i$ can be obtained.

After fully fusing neighboring features, CSALFFM further extracts their global information (cascading), significant information (adding), and differential information (subtracting). Then, self-attention are performed to fully explore the potential HLSF of the current adjacent layer.

The entire process is represented as follows:

$$\begin{cases} \overline{Q}^{i} = f_{d=1}^{1*1}\left(\overline{F}_1^{i} \copyright \overline{F}_2^{i} \copyright (\overline{F}_1^{i} + \overline{F}_2^{i})\right)\overline{W}_q^{i} \\ \overline{K}^{i} = f_{d=1}^{1*1}(\overline{F}_1^{i} - \overline{F}_2^{i})\overline{W}_k^{i} \\ \overline{V}^{i} = f_{d=1}^{1*1}\left(\overline{F}_1^{i} - \overline{F}_2^{i}\right)\overline{W}_v^{i}. \end{cases} \tag{7}$$

Final output $\bar{F}_{\mathrm{CS}}^{i} = Att(\overline{Q}^{i}, \overline{K}^{i}, \overline{V}^{i})$, namely

$$\overline{F}_{\mathrm{CS}}^{i} = \sigma\left(\frac{\overline{Q}^{i}\overline{K}^{i^T}}{\sqrt{d}}\right)\overline{V}^{i}. \tag{8}$$

D. Illustration of the Module MSGFFM

Local relevance and contextual information have been enhanced in SLDFEM and CSALFFM, respectively. To establish the global long-range dependency relationship of DTIs, we propose the MSGFFM based on transformer decoder.

For MSFs refined by the SLDFEM and CSALFFM, as shown in Fig. 6, we intend to use the residual-paradigm to fuse and heighten them.

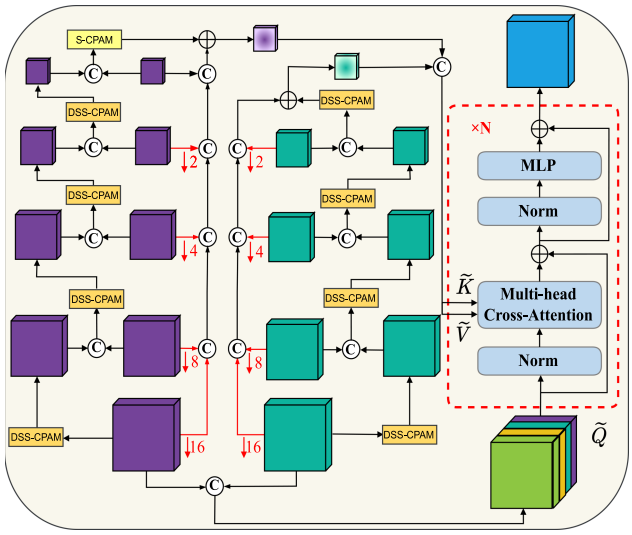


Fig. 6. Overall structure diagram of the module MSGFFM.

First, the large-scale features are gradually strengthened to achieve the MSF fusion are as follows:

$$\widetilde{F}_{\mathrm{SL}}^{i} = \begin{cases} g_2(F_{\mathrm{SL}}^{i}), & i = 1 \\ g_2(F_{\mathrm{SL}}^{i} \copyright \widetilde{F}_{\mathrm{SL}}^{i-1}), & i \in [2, M-1] \\ g_1(F_{\mathrm{SL}}^{i} \copyright \widetilde{F}_{\mathrm{SL}}^{i-1}), & i = M \end{cases} \tag{9}$$

where $g_2(\cdot)$ is the MSF enhancement function, which adopts the specially designed different-scale spatial channel parallel attention module (DSS-CPAM) as shown in Fig. 4(b). The structure of DSS-CPAM is similar to S-CPAM, but with the addition of a max pooling layer before the 3×3 convolution to downsample and unify their feature scale.

Then, the different-scale features $(F_{\mathrm{SL}}^{i}, \overline{F}_{\mathrm{CS}}^{i})$ are adjusted to the same size and directly sent to the fusion end, acting as skip-connection (i.e., original features)

$$\widetilde{F}_{\mathrm{SL}} = \widetilde{F}_{\mathrm{SL}}^{M} + C\left(d_{2^{M-i}}(F_{\mathrm{SL}}^{i})\right) \tag{10}$$

where $C(\cdot)$ refers to the cascading function, and $d_n(\mathcal{X})$ represents performing n-fold downsampling on the tensor $\mathcal{X}$.

In this way, the multiscale fusion results $\widetilde{F}_{\mathrm{SL}}$ and $\widetilde{F}_{\mathrm{AL}}$ extracted by SLDFEM and CSALFFM are obtained. Then, we apply the multihead cross-attention mechanism to model the global contextual relationships. Specifically, $\widetilde{F}_{\mathrm{SL}}$ and $\widetilde{F}_{\mathrm{AL}}$, which contain more deep semantic features, are cascaded and fed into the transformer decoder as K and V. The decoder requirement Q is crucial for transformer. It needs to focus on both the feature scale (related to the transformer output size) and the multilevel physical and semantic meaning, which could help to achieve a balance between deep and shallow information. Therefore, the large-sized maps that appear in the entire feature extraction and enhancement process are treated as Q

$$\begin{aligned} F_Q = C\big(&F_1^1 - F_2^1, F_{\mathrm{SL}}^1, \bar{F}_{\mathrm{CS}}^1, \\ &u_2(F_1^2 - F_2^2), u_2(F_{\mathrm{SL}}^2), u_2(\bar{F}_{\mathrm{CS}}^2)\big) \end{aligned} \tag{11}$$

where $u_n(\mathcal{X})$ represents performing n-fold upsampling on the tensor $\mathcal{X}$.

The entire attention can be represented as follows:

$$\begin{cases} \widetilde{Q} = (F_Q)\widetilde{W}_q \\ \widetilde{K} = (\widetilde{F}_{\text{SL}} \copyright \widetilde{F}_{\text{AL}})\widetilde{W}_k \\ \tilde{V} = (\tilde{F}_{\text{SL}} \copyright \tilde{F}_{\text{AL}})\widetilde{W}_v. \end{cases} \tag{12}$$

The output of transformer decoder $F_o = Att(\widetilde{Q}, \widetilde{K}, \widetilde{V})$

$$F_o = \sigma\left(\frac{\widetilde{Q}\widetilde{K}^T}{\sqrt{d}}\right)\widetilde{V} \tag{13}$$

$$\begin{aligned} &\text{MSA}\left(F_Q, \widetilde{F}_{\text{SL}} \copyright \widetilde{F}_{\text{AL}}\right) \\ &= Concat(head_1, \ldots, head_h)\widetilde{W}_o \end{aligned} \tag{14}$$

$$\begin{aligned} &head_j = Att\{F_Q\widetilde{W}_q^j, \\ &(\widetilde{F}_{\text{SL}} \copyright \widetilde{F}_{\text{AL}})\widetilde{W}_k^j, (\widetilde{F}_{\text{SL}} \copyright \widetilde{F}_{\text{AL}})\widetilde{W}_v^j\} \end{aligned} \tag{15}$$

where $\widetilde{W}_q, \widetilde{W}_k, \widetilde{W}_v$, and $\widetilde{W}_o$ represent linear projection, and h represents the number of attention heads. In addition, the MLP consists of two linear layers with a GELU activation function in between.

To get the final CD result PV, the detection head performs 3×3 convolution on the output of transformer decoder F_o, i.e., $PV = \text{softmax}(f_{d=1}^{3*3}(F_o))$.

E. Loss Function

1) Main Loss Function: An inherent problem in CD is the class imbalance between positive and negative samples. Regardless of the dataset, the number of negative samples (i.e., UA pixels) almost always exceeds that of positive samples (i.e., CA pixels). This often leads to neural networks ignoring the learning for positive samples and instead focusing on the less-important information from negative samples.

To address this issue, the main loss function combines binary cross-entropy (BCE) and dice coefficient to guide the training process in this article.

We mainly measure the difference between the probability distributions of two given random variables through the BCE, while the dice coefficient is used to test the similarity between different sets of variables (e.g., the DTIs pixels).

The main loss function is as follows:

$$L_{\text{main}} = \lambda_1 * L_{\text{bce}} + \lambda_2 * L_{\text{dice}} \tag{16}$$

$$L_{\text{bce}} = -y \cdot \log y'_n - (1-y) \cdot \log(1-y'_n) \tag{17}$$

$$L_{\text{dice}} = 1 - (2 \cdot y \cdot \text{softmax}(y'))/(y + \text{softmax}(y')) \tag{18}$$

where y represents the GT, and y' stands for the model predicted value.

2) Auxiliary Loss Function: To supervise the feature learning in the intermediate layers, a cross-entropy loss function is set for each fusion module. Considering the size nonuniformity, the CD labels are downsampled to match the size of the intermediate layers.

The auxiliary loss function is as follows:

$$L_{\text{AUX}} = \sum_{i=1}^{M} L_i(Z_i, f\left(F_{\text{SL}}^i\right)) + \sum_{i=1}^{M-1} \overline{L}_i\left(Z_i, f\left(\overline{F}_{\text{CS}}^i\right)\right) \tag{19}$$

where L_i and $\overline{L}_i$ both refer to cross-entropy loss function; Z_i represents a multiscale label, which is achieved by downsampling GT according to the intermediate PV size.

3) The Final Loss is Attained:

$$L_{\text{total}} = \lambda_1 * L_{\text{bce}} + \lambda_2 * L_{\text{dice}} + \lambda_3 * L_{\text{AUX}} \tag{20}$$

where $\lambda_i, i \in \{1, 2, 3\}$ are the regularization coefficients.

IV. Experimental Setup

A. Dataset

To validate the effectiveness of WSMsFNet, we conducted experiments on three representative HRRSCD datasets (i.e., LEVIR-CD, CDD, and SYSUCD). Each dataset consists of a change map and two HRRS images captured at different times in the same area. The details of the three datasets are as follows:

LEVIR-CD dataset: The LEVIR-CD dataset consists of 637 pairs of HRRS with the size of 1024×1024. These images are from 20 different areas in several cities in Texas, with the spatial resolution of 0.5 m. The main change type in this dataset is about building. For the experiment, we cropped each image into nonoverlapping blocks of size 256×256. The image number of the training, validation, and test set is 7120, 1024, and 2048 pairs, respectively.

CDD dataset: The CDD dataset consists of 16 000 pairs images with the size of 256×256, including training, verification, and test sets of 10 000, 3000, and 3000 pairs, respectively. The spatial resolution of CDD ranges from 0.3 to 1.0 m. It contains changes in different objects such as buildings, roads, and vehicles, while ignoring changes caused by factors such as seasonality and brightness.

SYSUCD dataset: This dataset consists of 20 000 pairs of aerial images captured in Hong Kong between 2007 and 2014, with the size of 256×256 and the spatial resolution of 0.5 m. The main change types in this dataset are as follows:

1) newly built urban buildings;
2) suburban expansion;
3) preconstruction groundwork;
4) vegetation change;
5) road expansion;
6) coastal construction.

In addition, the 20 000 image pairs are divided into a training set (10 000 pairs), a validation set (4000 pairs), and a testing set (4000 pairs).

B. Experimental Parameters

The WSMsFNet is trained and tested using a single NVIDIA RTX3090 GPU. We adapt the Adam optimizer to reform the model, with an initial learning rate of 0.001 linearly decaying to 0. In addition, validation is conducted after each training cycle, and the best model on the validation set is used to evaluate the test

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS ON VARIOUS DATASETS

	LEVIR-CD P. / Re. / F1. / OA. / mIoU	CDD P. / Re. / F1. / OA. / mIoU	SYSUCD P. / Re. / F1. / OA. / mIoU
STANet	83.42/90.44/86.79/98.59/87.60	88.37/90.07/89.21/97.43/88.82	79.73/**75.00**/77.29/89.61/75.18
DTCDSCN	**90.73**/88.24/**89.47**/**98.94**/**89.92**	93.52/93.27/93.40/98.44/92.93	79.16/73.50/76.22/89.18/74.25
SRCDNet	90.84/81.14/85.72/98.45/86.69	93.25/91.81/92.52/98.22/92.04	82.89/72.31/77.24/88.48/74.30
MSPSNet	90.80/**88.70**/89.74/98.96/90.15	85.53/74.23/79.48/95.47/80.49	77.61/74.45/76.00/88.91/73.92
BIT	89.35/86.95/88.14/98.80/88.77	96.72/91.51/**94.05**/98.63/**93.61**	80.41/72.97/76.51/89.43/74.60
Change Former	86.26/84.91/85.58/98.54/86.63	92.13/78.98/85.05/96.72/85.19	81.30/70.40/75.46/89.20/73.82
ICIFNet	76.65/81.25/78.88/98.30/81.69	94.71/88.56/91.53/98.07/91.11	84.60/70.71/77.03/90.06/75.36
DMINet	79.23/77.61/78.41/98.33/81.39	95.22/90.41/92.75/98.33/92.31	86.07/71.13/**77.89**/**90.48**/**76.17**
USSFC-Net	90.13/84.19/87.06/98.73/87.86	**95.49**/91.29/93.34/98.40/92.87	81.66/68.24/74.35/88.90/72.97
WSMsFNet(1:1:1)	90.19/90.67/90.43/99.02/90.75	95.18/**92.97**/94.06/**98.61**/93.62	**84.41**/78.68/80.50/91.01/78.16
WSMsFNet(**optimum**)	— / — / — / — / —	96.26/93.84/95.03/98.84/94.62	82.16/79.74/80.93/91.14/78.53

*All values are reported in percentage(%). Color convention:Best, 2-nd best, **3-ed best**.
*The indicator values shown in the last line are derived from our WSMsFNet of optimal regularization term configuration.

set. The backbone layer M is uniformly set to 5, and $C_b^i, C_s^i \in \{64, 64, 128, 256, 512\}$.

C. Evaluation Metrics

To comprehensively reflect model performance, we evaluate the experimental results using five evaluation metrics: precision (P), recall (Re), F1 score, overall accuracy (OA), and mean intersection over union (mIoU). The definitions of these metrics are as follows:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{21}$$

$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{22}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{P} \times \mathrm{Re}}{\mathrm{P} + \mathrm{Re}} \tag{23}$$

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{24}$$

$$\mathrm{mIoU} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN} + \mathrm{FP}}\right). \tag{25}$$

V. EXPERIMENTAL RESULTS

A. Comparative Methods

To validate the WSMsFNet effectiveness, we compare it with several advanced CD methods, including STANet [63], DTCDSCN [64], SRCDNet [65], MSPSNet [66], ChangeFormer [56], BiT [58], ICIF [32], DMINet [67], and USSFCNet [68] etc. For fairness, we train and test the aforementioned CD networks using their publicly available codes and default hyperparameters on our unified platform. Eventually, the quantitative comparison on the datasets is shown in Table I, and the model visual results are presented in Figs. 7, 8, and 9. To facilitate distinction, we use different colors to represent TP (white), TN (black), FP (red), and FN (green). In the case study, we choose various sizes and quantities of change targets for comparison, which is convenient for comprehensive evaluation of model performance.

B. Quantitative Comparison

Table I shows the overall performance comparison of different algorithms on three datasets. The quantitative metrics indicate that our WSMsFNet, after tuning with regularization terms, consistently outperforms other algorithms. For example, F1 and mIoU surpass BIT 2.29/0.98/4.42 points, 1.98/1.01/3.93 points in three datasets, respectively.

Furthermore, our CNN backbone only uses ResNet34 and does not involves more complex structures such as ResNet50, FPN, and Unet. This may be attributed to our model ability to fuse multiscale spatio-temporal information and enhance feature representation by modeling global contextual relations.

C. Results and Discussion on the LEVIR-CD Dataset

Fig. 7 showcases the visual comparison of different method on the LEVIR-CD. Many algorithms tend to struggle with false negatives for small targets, such as SRCD, MSPS, ChangeFormer, ICIFNet, DMINet, and USSFCNet. When multiple CD targets are present, algorithms may experience varying degrees of false positives near the edges, especially STANet, SRCD, and USSFCNet. For complex and densely arranged structures, issues of false positives (MSPS, BIT, ChangeFormer) and false negatives (STANet, SRCD, ICIFNet, and DMINet) often arise near the boundaries. Therefore, the completeness and smoothness of the building boundaries could intuitively reflect the model performance. Complex scenes can also interfere with the CD performance for certain algorithms, such as MSPS, BIT, ChangeFormer, and DMINet. Regardless of the target size and amount, the WSMsFNet demonstrates superior performance in terms of target completeness and boundary precision. This might be attributed to the model effective fusion of MSF.

D. Results and Discussion on the CDD Dataset

Their detection results on the CDD dataset are shown in Fig. 8. Compared to the other two datasets, the CDD expresses more complex scenes and greater disturbance in the CAs. In addition, the distribution of CAs is uneven, and the type and size of change targets vary critically. Visually, the BIT, Changeformer, and

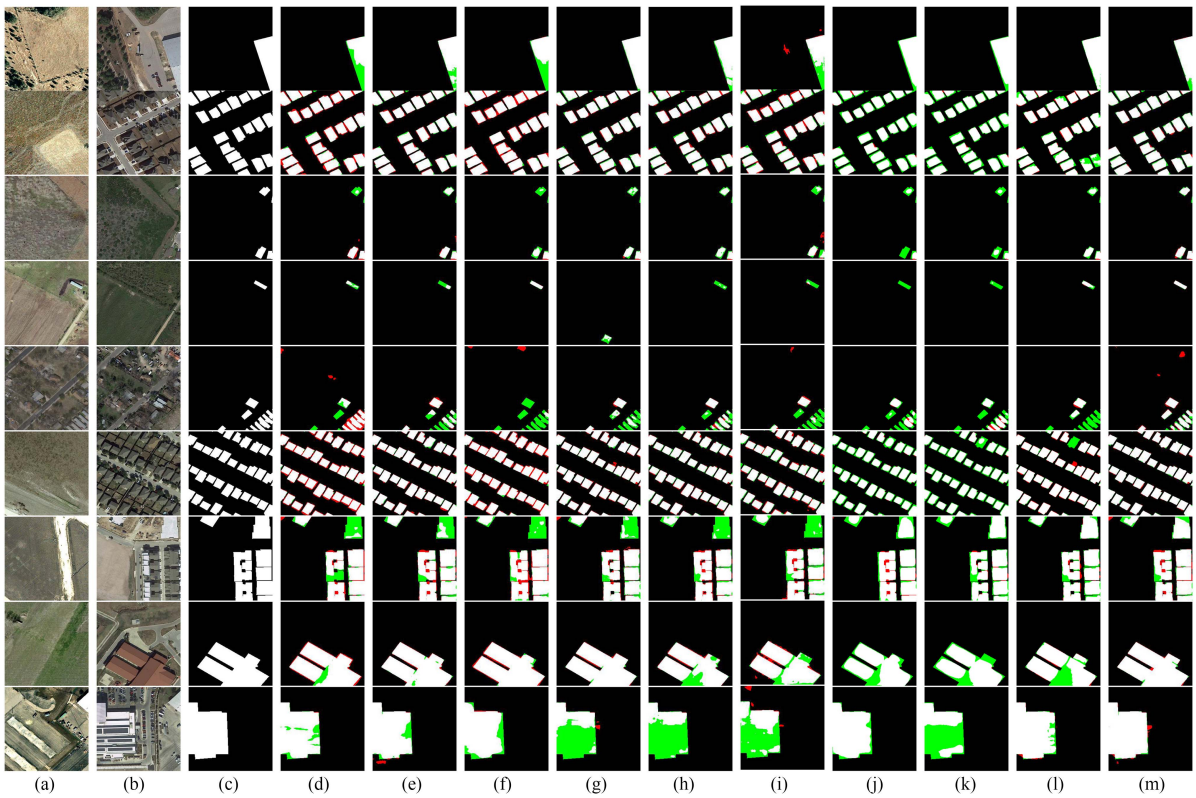


Fig. 7. Detection results on the LEVIR-CD using different methods. (a) A. (b) B. (c) Lable. (d) STANet. (e) DTCDSCN. (f) SRCDNet. (g) MSPSNet. (h) BIT. (i) Change Former. (j) ICIFNet. (k) DMINet. (l) USSFCNet. (m) WSMsFNet.

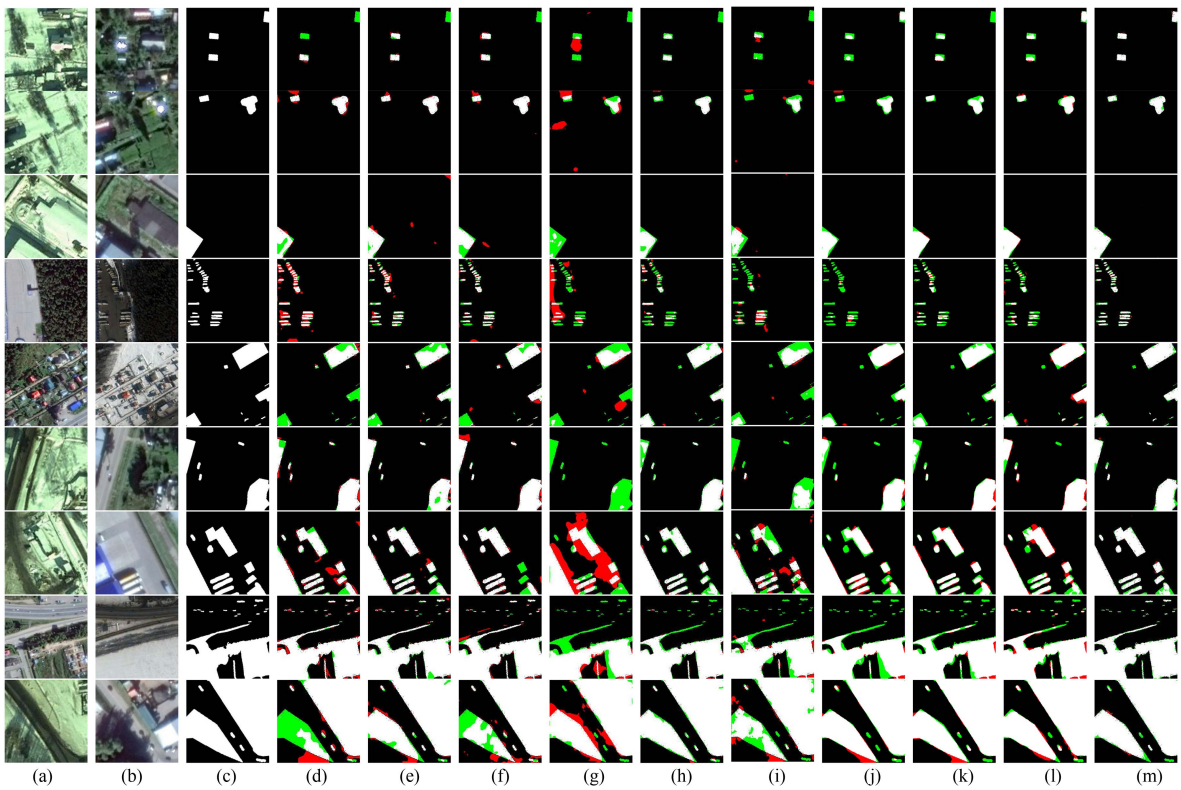


Fig. 8. Detection results on the CDD using different methods. (a) A. (b) B. (c) Lable. (d) STANet. (e) DTCDSCN. (f) SRCDNet. (g) MSPSNet. (h) BIT. (i) Change Former. (j) ICIFNet. (k) DMINet. (l) USSFCNet. (m) WSMsFNet.

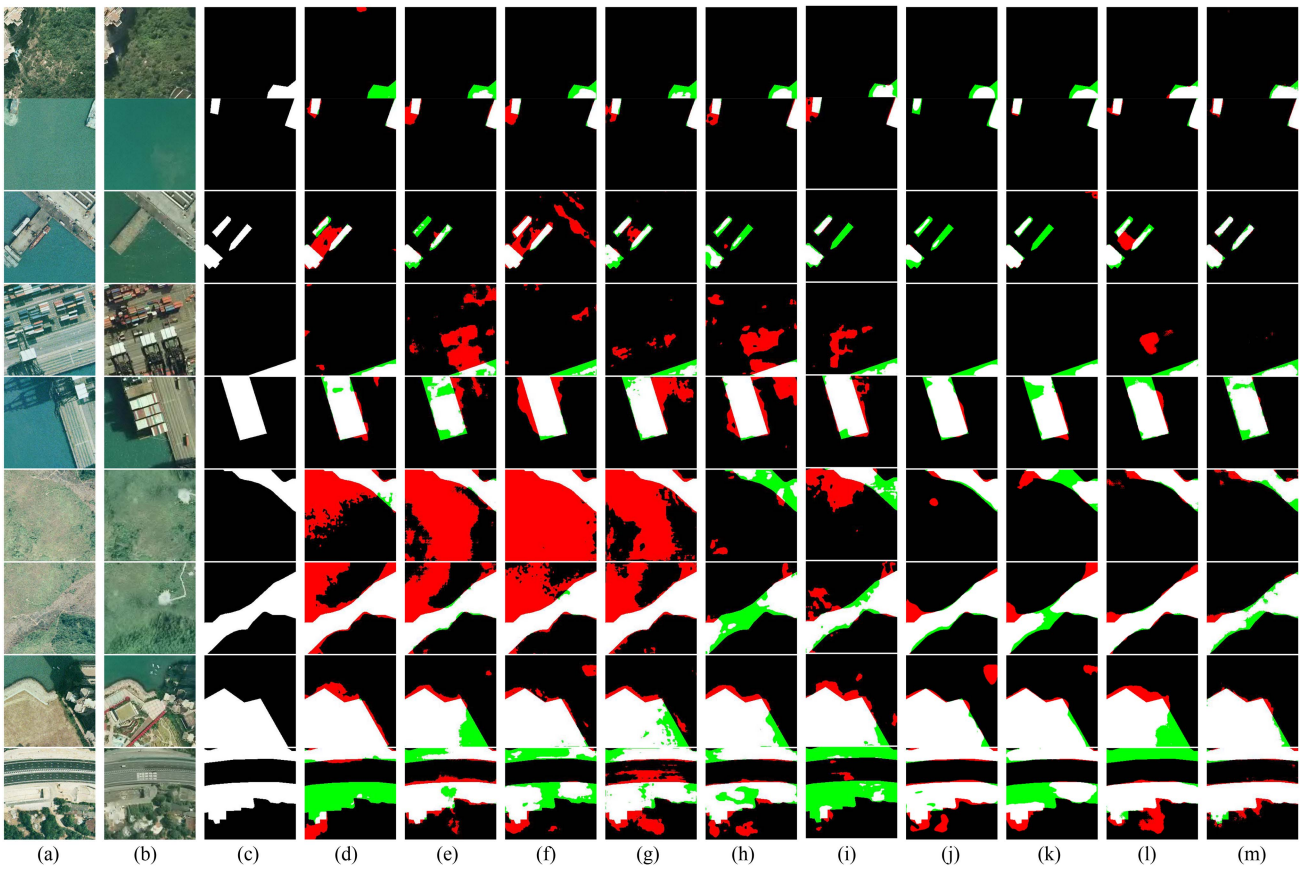


Fig. 9. Detection results on the SYSUCD using different methods. (a) A. (b) B. (c) Lable. (d) STANet. (e) DTCDSCN. (f) SRCDNet. (g) MSPSNet. (h) BIT. (i) Change former. (j) ICIF. (k) DMINet. (l) USSFCNet. (m) WSMsFNet.

USSFCNet exhibit significant false negatives overall. STANet and MSPSNet show certain false positives at different scales. Models such as MSPS, BIT, ChangeFormer, ICIFNet, DMINet, and USSFCNet all are prone to missing smaller targets scattered in the scene. For larger CAs, many methods exhibit false negatives at the target edges (e.g., MSPS and ICIFNet, etc.) or detection errors (e.g., MSPS and STANet, etc.). In comparison, WSMsFNet shows better adaptability to detecting those wider ranges of change targets.

E. Results and Discussion on the SYSUCD Dataset

Fig. 9 showcases the visual CD results on the SYSUCD. Due to the greater number and complexity of change types in the SYSUCD dataset, some models exhibit relatively poorer CD performance compared to LEVIR-CD. Under similar lighting conditions, there are noticeable false negatives in the small target areas for DTCDSCN, BIT, Changeformer, ICIFNet, and DMINet. For big target, the scene complexity causes the detected target boundaries to be affected in different degrees. When disturbed by luminosity, STANet, DTCDSCN, SRCDNet, MSPS, and BIT all show significant false positives, while STANet, ChangeFormer, and DMINet exhibit more prominent false negatives. Overall, the WSMsFNet finds CA with clear boundaries, fewer false positives, and false negatives, showing better adaptability when facing some wider ranges of change types. This may be attributed that WSMsFNet could extract HLSF through MSF fusion and result in better robustness.

F. Regularization Parameter Setting

To evaluate the contribution of each loss term, we conducted ablation experiment about the hyperparameters $\lambda 1$, $\lambda 2$, and $\lambda 3$, whose results are shown in Table II. For the LEVIR-CD, there is a slight difference in overall performance under various tradeoff parameters. It can be seen that the cross-entropy, dice loss, and deep supervision auxiliary loss have relatively balanced effects on the model and dataset. We chose the weight-ratio with the highest Re (1:1:1) as the optimal parameter, which is 2.2% higher in recall than the lowest value (90.6 versus 88.4). For the CDD, the optimal parameter configuration (2:2:1) is tremendously significant. The proportions of cross-entropy and dice loss are the same and both higher than the deep supervision loss. Similarly, the optimal regularization coefficient for the SYSUCD is (2:2:3). They clearly outperform other configurations in terms of comprehensive metrics (i.e., F1 and mIoU). This indicates that the level of requirement for deep supervision varies for different datasets.

G. Ablation Experiments

As shown in Table III, we conducted ablation experiments to verify the effectiveness of the module. The experiment shows

TABLE II
EXPERIMENTAL RESULTS OF REGULARIZATION COEFFICIENTS

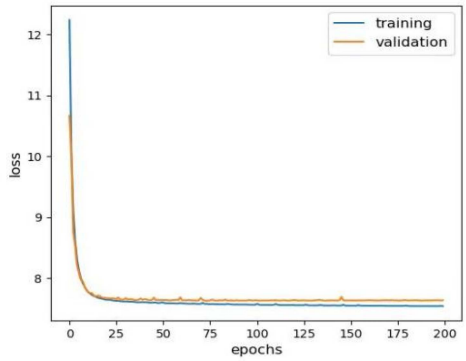
λ_1	λ_2	λ_3	LEVIR-CD P. / Re. / F1. / OA. / mIoU	CDD P. / Re. / F1. / OA. / mIoU	SYSUCD P. / Re. / F1. / OA. / mIoU
1	1	1	90.19/**90.67**/90.43/99.02/90.75	95.18/92.97/94.06/98.61/93.62	**84.41**/78.68/80.50/91.01/78.16
2	1	1	91.12/89.85/90.48/99.17/**90.90**	95.54/92.15/93.81/98.56/93.37	84.31/75.76/79.80/90.95/77.69
2	3	3	91.01/89.97/90.48/**99.20**/**90.90**	95.16/93.00/94.07/98.61/93.63	81.40/78.41/79.87/90.68/77.53
2	1	3	91.60/89.18/90.38/**99.20**/90.81	95.89/93.38/94.62/98.74/94.19	77.33/80.24/78.76/89.79/76.18
2	3	1	91.05/89.92/90.48/**99.20**/**90.90**	96.10/93.66/94.86/98.80/94.44	79.45/**81.52**/80.47/90.67/77.88
2	2	1	**91.63**/88.40/89.99/99.17/90.47	96.26/**93.84**/**95.03**/**98.84**/**94.62**	82.42/78.13/80.22/90.91/77.91
2	2	3	91.46/88.78/90.08/99.18/90.55	96.78/93.03/94.87/98.81/94.45	82.16/79.74/**80.93**/**91.14**/**78.53**
2	1	2	90.94/89.19/90.06/99.17/90.53	**96.90**/93.02/94.92/98.82/94.51	83.40/76.17/79.62/90.80/77.47
2	3	2	91.49/89.59/**90.53**/99.04/90.85	95.17/93.36/94.26/98.65/93.82	79.15/79.52/79.34/90.23/76.86

*All values are reported in percentage(%). Color convention:**Best.**

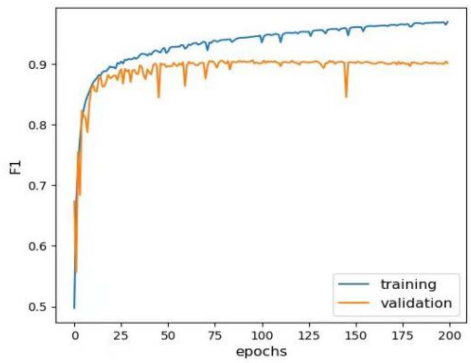
TABLE III
MODULE ABLATION EXPERIMENT

Model SLDFFM/CSALFFM/MSGFFM	LEVIR F1 / mIoU	CDD F1 / mIoU	SYSUCD F1 / mIoU
✓ / ✗ / ✗	88.90/89.40	91.10/90.69	78.74/76.24
✗ / ✓ / ✗	87.81/88.51	89.10/88.79	78.20/76.05
✗ / ✗ / ✓	75.88/79.32	81.91/82.43	70.32/69.46
✓ / ✓ / ✗	89.03/89.55	91.02/90.62	79.39/78.06
✓ / ✗ / ✓	88.95/89.47	90.17/89.79	80.20/77.76
✗ / ✓ / ✓	88.62/89.19	90.06/89.68	79.19/77.09
✓ / ✓ / ✓	**90.43/90.75**	**94.06/93.62**	**80.50/78.16**

*All values are reported in percentage(%). Color convention:**Best.**



(a)



(b)

Fig. 10. Convergence and accuracy of WSMsFNet on the LEVIR-CD during training/validation sets. (a) The overall trend of loss value regarding training/validation sets. (b) The overall trend of F1 score regarding training/validation sets.

that regardless of which dataset, each module has made its own contribution to improving model performance.

H. Convergence Analysis

To visualize the training process, we test the convergence and accuracy of WSMsFNet on the LEVIR-CD. Fig. 10(a) shows that the model loss rapidly decreases within the first 25 epochs, both for the testing and validation sets. The validation set loss has stabilized after 100 epochs, indicating good convergence of the WSMsFNet. Similarly, the F1 score rapidly increases within the first 25 epochs, and that remains stable after 100 epochs on the validation set. It indicates that the WSMsFNet is convergent, stable, and effective. This may be attributed that the WSMsFNet could learn effective MSF and global contextual information, which accurately represents the interest areas for the detected CAs.

I. Network Visualization

To better illustrate the learning effectiveness, a sample from the test set is used to visualize the heatmaps of each stage in WSMsFNet. The heatmaps provide the intuitive explanation for the network learning about the changing targets. The visualization results are shown in Fig. 11. Given DTIs, the hierarchical features are first extracted from shallow to deep levels using ResNet. Then, the designed intralevel and adjacent-level fusion modules concentrate the attention mechanism on the interest regions. It is apparent that the intralevel fused features are more refined, while the adjacent-level fused features have the richer scale, demonstrating the effectiveness of these modules. For the decoding stage, the output represented by "Q" showcases the contour and position information of the CAs, highlighting the requirements of CD; while the "K and V" mainly contains spatial or HLSF that could exist in the CAs. Meanwhile, the attention towards unchanged targets is noticeably reduced in the decoding stage.

Overall, WSMsFNet effectively learns semantic features in a hierarchical manner to highlight the changed targets.

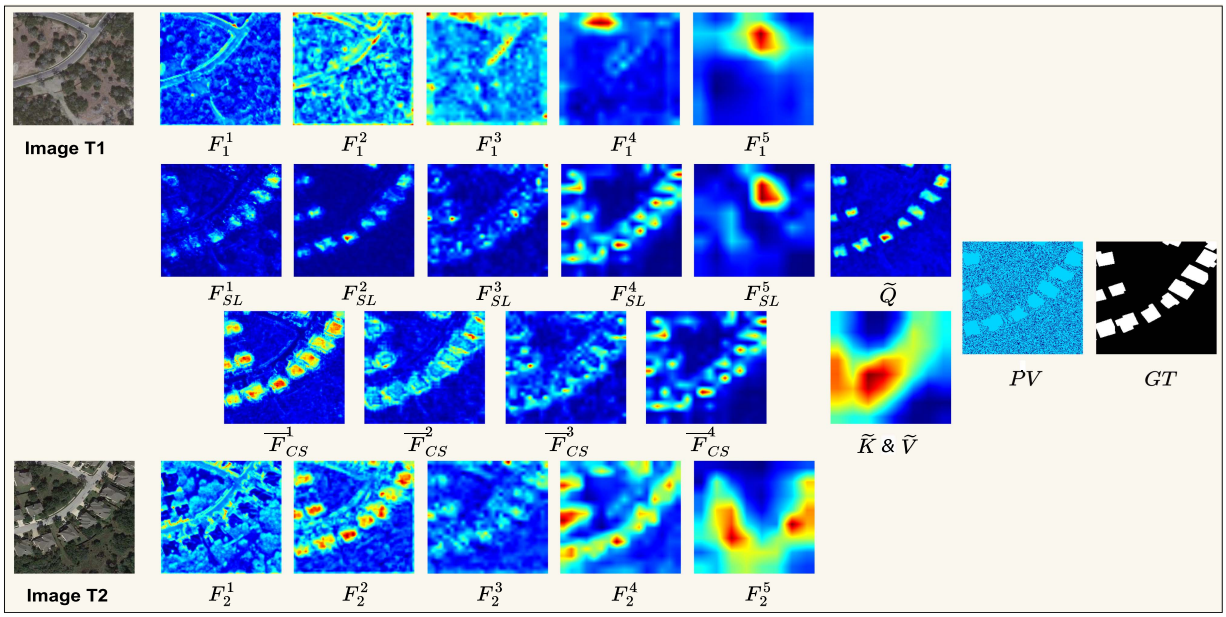


Fig. 11. Visualization of key modules in WSMsFNet. Each thermal map is generated by adding and normalizing channel-level elements to the corresponding feature map tensor. For the convenience of observation, each visualization image is upsampled to 256×256 through bilinear interpolation.

VI. Discussion

To suit the CA nonuniformity, the WSMsFNet aims to refine the multiscale features of the dual-temporal image to realize RSCD in the region of interest. In the feature fusion stage for each scale, this article fully regards the global-local correlation relationship modeling. In addition, the full-process supervision also makes each feature extraction and fusion module more efficient. Experiments show that our model performs better than others in several metrics (F1 and OA, etc.).

However, the WSMsFNet still exists some limitations: One is the problem of model light-weighting. This model parameter is up to 33.5 MB, and streamlining parameters under the premise of guaranteeing performance is the focus of our future research. The others is the model robustness problem. When the model suffers from serious noise, light interference, or even specialized malicious attacks, specific measures of improving robustness should be invoked to cope with such situations.

VII. Conclusion

Aiming at the variability and potential contextual relationships of CAs in dual-phase RS images, this article proposes the CD depth network WSMsFNet based on MSF fusion and Transformer variants. First, ResNet34 has been adopted as backbone to extract multiscale local features from DTRSI. Then, we introduce modules such as SLDFEM, CSALFFM, and MSGFFM to achieve MSF fusion and enhance the representation of global contextual information. Finally, an auxiliary loss function is designed to supervise the learning of intermediate layer features. Moreover, the experiments on the LEVIR-CD, CDD, and SYSUCD datasets demonstrate that WSMsFNet achieves favorable results in terms of comprehensive metrics (F1, mIoU) and qualitative comparisons. This verifies the strong adaptability of WSMsFNet in detecting different types of change targets.

References

[1] H. Li et al., "Selective transfer based evolutionary multitasking optimization for change detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 3, pp. 2197–2212, Jun. 2024.

[2] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.

[3] Y. Sun, L. Lei, D. Guan, J. Wu, G. Kuang, and L. Liu, "Image regression with structure cycle consistency for heterogeneous change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1613–1627, Feb. 2024.

[4] 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.

[5] C. Mulverhill, N. C. Coops, and A. Achim, "Continuous monitoring and sub-annual change detection in high-latitude forests using harmonized landsat sentinel-2 data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 309–319, 2023.

[6] R. Zhou et al., "A unified deep learning network for remote sensing image registration and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5101216.

[7] C. Wang, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Introspective deep metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 1964–1980, Apr. 2024.

[8] M. Fu and F.-X. Wu, "Qlabgrad: A hyperparameter-free and convergence-guaranteed scheme for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 12072–12081.

[9] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[10] Y. Chen, M. Mancini, X. Zhu, and Z. Akata, "Semi-supervised and unsupervised deep visual learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1327–1347, Mar. 2024.

[11] Y. Fan, Y. Yu, W. Lu, and Y. Han, "Weakly-supervised video anomaly detection with snippet anomalous attention," in *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5480–5492, Jul. 2024.

[12] S. Chen, K. Yang, and R. Stiefelhagen, "DR-TANet: Dynamic receptive temporal attention network for street scene change detection," in *2021 IEEE Intell. Veh. Symp.*, 2021, pp. 502–509.

[13] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," *IEEE Trans. Image Process.*, vol. 30, pp. 55–67, 2021.

[14] S. Pang, J. Lan, Z. Zuo, and J. Chen, "SFGT-CD: Semantic feature-guided building change detection from bitemporal remote-sensing images with transformers," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2500405.

[15] J. Long, M. Li, X. Wang, and A. Stein, "Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 211, pp. 318–335, 2024.

[16] Y. Zuo et al., "Robust instance-based semi-supervised learning change detection for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4404815.

[17] M. Zhao, X. Hu, L. Zhang, Q. Meng, Y. Chen, and L. Bruzzone, "Beyond pixel-level annotation: Exploring self-supervised learning for change detection with image-level supervision," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5614916.

[18] J. Qu, W. Dong, Y. Yang, T. Zhang, Y. Li, and Q. Du, "Cycle-refined multidecision joint alignment network for unsupervised domain adaptive hyperspectral change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024, to be published, doi: 10.1109/TNNLS.2023.3347301.

[19] B. Cui, C. Liu, and J. Yu, "BGSINet-CD: Bitemporal graph semantic interaction network for remote-sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5002205.

[20] W. Peng, W. Shi, M. Zhang, and L. Wang, "FDA-FFNet: A feature-distance attention-based change detection network for remote sensing image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2224–2233, 2024.

[21] Y. Shangguan, J. Li, Z. Chen, L. Ren, and Z. Hua, "MultiScale attention fusion graph network for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4402618.

[22] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and X. Gao, "DCENet: Diff-feature contrast enhancement network for semi-supervised hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5511514.

[23] Z. Mao, Z. Luo, and Y. Tang, "Remote sensing building change detection with global high-frequency cues guidance and result-aware alignment," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6005105.

[24] Y. Huang, L. Zhang, W. Qi, R. Song, C. Huang, and Y. Cen, "Semi-supervised hypermatch-driven cross temporal and spatial interaction transformer for hyperspectral change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6426–6443, 2024.

[25] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting segment anything model for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611711.

[26] L. Miao et al., "SNUNet3+: A full-scale connected siamese network and a dataset for cultivated land change detection in high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4400818.

[27] Y. Tang et al., "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600213.

[28] F. Gu, P. Xiao, X. Zhang, Z. Li, and D. Muhtar, "FDFF-Net: A full-scale difference feature fusion network for change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2161–2172, 2024.

[29] D. Zheng, Z. Wu, J. Liu, C.-C. Hung, and Z. Wei, "Detail enhanced change detection in VHR images using a self-supervised multiscale hybrid network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3181–3196, 2024.

[30] J. Wang, Y. Zhong, and L. Zhang, "Contrastive scene change representation learning for high-resolution remote sensing scene change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5618118.

[31] J. Ma, J. Duan, X. Tang, X. Zhang, and L. Jiao, "Eatder: Edge-assisted adaptive transformer detector for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5602015.

[32] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.

[33] J. Li, S. Li, and F. Wang, "LCDNet: Lightweight change detection network with dual-attention guidance and multiscale feature fusion for remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6000705.

[34] X. Zhao, K. Zhao, S. Li, C. Song, and X. Wang, "GAMPF: A full-scale gated message passing framework based on collaborative estimation for VHR remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5603413.

[35] H. Lin, R. Hang, S. Wang, and Q. Liu, "DiFormer: A difference transformer network for remote sensing change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6003905.

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

[37] A. Raza, H. Huo, and T. Fang, "EUNet-CD: Efficient UNet++ for change detection of very high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3510805.

[38] 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, 2022, Art. no. 8000605.

[39] J. Lei, Y. Gu, W. Xie, Y. Li, and Q. Du, "Boundary extraction constrained siamese network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621613.

[40] S. Zhu, Y. Song, Y. Zhang, and Y. Zhang, "ECFNet: A Siamese network with fewer FPS and fewer FNS for change detection of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6001005.

[41] Y. Deng et al., "Feature-guided multitask change detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9667–9679, 2022.

[42] 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, 2022, Art. no. 5610613.

[43] J. Qu, S. Hou, W. Dong, Y. Li, and W. Xie, "A multilevel encoder–decoder attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518113.

[44] G. Pei and L. Zhang, "Feature hierarchical differentiation for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6514105.

[45] C. Zhao, L. Ma, L. Wang, T. Ohtsuki, P. T. Mathiopoulos, and Y. Wang, "SAR image change detection in spatial-frequency domain based on attention mechanism and gated linear unit," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 4002205.

[46] X. Zhao et al., "Fractional fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2314–2326, Feb. 2024.

[47] Y. Xu, B. Du, and L. Zhang, "Robust self-ensembling network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3780–3793, Mar. 2024.

[48] F. Wang, J. Ji, and Y. Wang, "Remote sensing image semantic segmentation based on cascaded transformer," *IEEE Trans. Artif. Intell.*, 2024, to be published, doi: 10.1109/TAI.2024.3363685.

[49] H. Chen, F. Shen, D. Ding, Y. Deng, and C. Li, "Disentangled cross-modal transformer for RGB-D salient object detection and beyond," *IEEE Trans. Image Process.*, vol. 33, pp. 1699–1709, 2024.

[50] J. Liu, X. Wang, M. Guo, R. Feng, and Y. Wang, "Shadow detection in remote sensing images based on spectral radiance separability enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3438–3449, May 2024.

[51] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "TTST: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 738–752, 2024.

[52] Y. Alkendi, R. Azzam, A. Ayyad, S. Javed, L. Seneviratne, and Y. Zweiri, "Neuromorphic camera denoising using graph neural network-driven transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 4110–4124, Mar. 2024.

[53] Y. Su, J. Deng, R. Sun, G. Lin, H. Su, and Q. Wu, "A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 313–325, 2024.

[54] T. Liu et al., "Tracking with saliency region transformer," *IEEE Trans. Image Process.*, vol. 33, pp. 285–296, 2024.

[55] Z. Luo et al., "Exploring point-BEV fusion for 3D point cloud object tracking with transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 5921–5935, Sep. 2024.

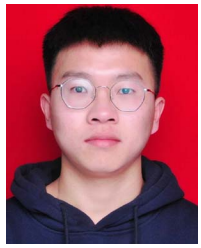
[56] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IGARSS 2022-2022 IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[57] F. Song, S. Zhang, T. Lei, Y. Song, and Z. Peng, "MSTDSNet-CD: Multiscale swin transformer and deeply supervised network for change detection of the fast-growing urban regions," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6508505.

[58] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[59] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 18537–18546.
[60] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-CNN architectures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 14388–14397.
[61] H. Fang, S. Guo, X. Wang, S. Liu, C. Lin, and P. Du, "Automatic urban scene-level binary change detection based on a novel sample selection approach and advanced triplet neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601518.
[62] W. G. C. Bandara and V. M. Patel, "Deep metric learning for unsupervised remote sensing change detection," 2023, *arXiv:2303.09536*.
[63] 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, Art. no. 1662.
[64] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[65] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403718.
[66] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multiscale siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406512.
[67] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.
[68] T. Lei et al., "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402114.

Bin Wang received the bachelor's degree in automation from the Anhui University of Science and Technology (AUST), Huainan, China, in 2010, the M.S. degree in control theory and control engineering from the Hefei University of Technology (HFUT), Hefei, China, in 2013, and the Ph.D. degree in complex system modeling and simulation from the North University of China (NUC), Taiyuan, China, in 2020.

He is currently an Assistant Professor with the Department of Computer Science and Technology, North University of China. His research interests include deep learning and remote sensing image change detection.



Xiaohu Jiang received the bachelor's degree in computer science and technology from Shanxi Datong University (AUST), Datong, China, in 2022. He is currently working toward the M.S. degree in artificial intelligence with the Department of Computer Science and Technology, North University of China (NUC), Taiyuan, China.

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

Pinle Qin received the M.S. and Ph.D. degrees in computer application technology from the Dalian University of Technology, Dalian, China, in 2003 and 2008, respectively.

He is currently a Professor with the Department of Computer Science and Technology, North University of China, Taiyuan, China. His research interests include deep learning, image process, video splicing, and fusion.

Jianchao Zeng received the B.S. degree in industrial automation from Taiyuan Heavy Machinery Institute, Taiyuan, China, in 1982, and the M.S. and Ph.D. degrees in engineering from Xi'an Jiaotong University, Xi'an, China, in 1985 and 1990, respectively.

He is currently a Professor with the Department of Computer Science and Technology, North University of China, Taiyuan, China. He is also the Vice President of the North University of China. He has authored or coauthored more than 200 international journal and conference papers. His current research interests include modeling and control of complex systems, intelligent computation, swarm intelligence, and swarm robotics.

Dr. Zeng is the Director of the China Simulation Federation. He is also the Vice President of the Technical Committee on System Simulation of the Chinese Association of Automation.