DA^2-Net: Integrating SAM2 With Domain Adaption and Difference Aggregation for Remote Sensing Change Detection

Hailong Ning, *Member, IEEE*, Qi He, Tao Lei, *Senior Member, IEEE*, Xiaopeng Cao, Wuxia Zhang, Yanping Chen, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*—Visual foundation models (VFMs) have been widely applied in the field of remote sensing (RS). However, they still face two main challenges when applied to precise RS change detection (RSCD) tasks in complex scenes. Firstly, the nonnegligible domain shift between natural scene and RS scene limits the direct application of VFMs to the RSCD task. Second, most of the existing RSCD methods may suffer from the boundary displacement problem due to the inadequate exploration of temporal differences for bi-temporal features. To address the above issues, this study proposes a Segment Anything Model 2 (SAM2)-based domain adaptive and spatial difference aggregation network (DA^2-Net) for RSCD. The proposed DA^2-Net has two main advantages. First, a hierarchical low-rank adaptation (LoRA) strategy is presented by introducing low-rank matrices at key positions of SAM2, which can inject inductive biases from the RS domain into the network and alleviate the domain shift problem. Second, a difference adaptive enhancement module (DAEM) is designed to explore temporal differences for hierarchical bi-temporal features. The DAEM provides respective attention weights for different information through a dual branch of global difference awareness and local detail optimization. Experimental results on SYSU-CD, WHU-CD, and LEVIR-CD datasets demonstrate the superiority of DA^2-Net. Code is available at https://github.com/xuptheqi-hash/DA2Net.**

***Index Terms*—Domain adaptation, feature aggregation, remote sensing change detection (RSCD), vision foundation models (VFMs).**

Received 13 July 2025; revised 29 August 2025; accepted 29 September 2025. Date of publication 6 October 2025; date of current version 3 November 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62201452, Grant 62271296, and Grant 62471389; in part by the Technology Innovation Guidance Special Fund of Shaanxi Province under Grant 2024QY-SZX-17; in part by the Innovation Capability Support Plan Project in Shaanxi Province under Grant 2025RS-CXTD-012; and in part by Shaanxi Provincial Key Research and Develop Program General Project under Grant 2024SF-YBXM-572. *(Corresponding author: Tao Lei.)*

Hailong Ning, Xiaopeng Cao, Wuxia Zhang, and Yanping Chen are with the School of Computer Science and Technology, Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, and Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: ninghailong93@gmail.com; caoxp@xupt.edu.cn; wuxiazhang100@126.com; chenyp@xupt.edu.cn).

Qi He is with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: xuptheqi@163.com).

Tao Lei is with Shaanxi Joint Laboratory of Artificial Intelligence and the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: leitao@sust.edu.cn).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University of London, UB8 3PH Uxbridge, U.K., and also with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: asoke.nandi@brunel.ac.uk).

Digital Object Identifier 10.1109/TGRS.2025.3617980

I. Introduction

REMOTE sensing change detection (RSCD) is a significant research topic in the field of Earth observation [1], with the goal of detecting relevant semantic changes in bi-temporal remote sensing (RS) images of the same area. It has been widely applied in various scenarios, such as land planning [2], urban expansion [3], military reconnaissance [4], disaster monitoring [5], and environmental protection [6].

RSCD is a challenging task due to the interference of various noises, such as illumination variations, seasonal and environmental variations, which may cause pseudo changes. In the past decade, deep learning (DL) techniques have almost become the dominant architecture for RSCD due to their powerful ability to automatically extract high-level semantic features. Typical representatives of these methods are convolutional neural networks (CNNs) based and transformer-based RSCD methods. Generally, CNN-based RSCD methods exploit various mechanisms (e.g., dilated convolutions [7], multiscale features [8], [9], and various attention mechanisms [10], [11]) for learning discriminative semantic features. For example, Song et al. [12] built a spatial integration module using dilated convolutions. Although CNN-based methods have achieved practical success, they struggle to capture global information, deteriorating the RSCD performance. In recent years, due to the excellent global modeling capability, vision transformers (ViTs) [13] have demonstrated outstanding performance in various RS tasks [14], [15], particularly in RSCD. For example, Zhang et al. [16] proposed a pure transformer network with shifted windowing. Although achieving impressive performance, transformer-based RSCD methods exhibit limited ability in learning local information, leading to suboptimal detection result for changed object details. To this end, some studies have combined CNNs with transformers, aiming to explore more powerful RSCD variants. For example, Feng et al. [17] used CNN and transformer to extract features from each image simultaneously.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

In summary, these attempts have achieved a certain degree of success. However, with the expansion of the model scale, the limited data hinders further improvement of model performance. In fact, the data acquisition and labeling for RSCD is very time-consuming and labor-intensive. Despite much research efforts to alleviate the data dependence, such as few-shot learning [18], generating simulation data [19], and data augmentation [20], the network's overfitting for specific data sets continues to hinder its generalization performance.

A. Motivation

Recently, visual foundation models (VFMs) have gained considerable attention due to their task-agnostic characteristics across various downstream tasks. After training on large-scale natural datasets, VFMs acquire strong generalization and few-shot learning capabilities, which help reduce the reliance of downstream tasks on specific data. Although previous studies [21], [22], [23] have applied VFMs (e.g., CLIP [24], SAM [25]) to RSCD tasks, these VFMs still have some limitations. First, CLIP optimizes the similarity between images and text through global contrastive learning rather than focusing on pixel-level visual understanding. As a result, the visual features extracted by CLIP tend to be overly global and often introduce substantial background noise. When applied to RSCD, this leads to insensitivity to subtle changes and an increased likelihood of false change detections. Second, although SAM mitigates some of the aforementioned issues, it generates features with a single scale and relatively low resolution. Its computational and memory costs also become significant when processing low-resolution images. These limitations hinder its practical applicability, especially in scene-level RS scenarios where object scales vary significantly.

Segment Anything Model 2 (SAM2) [26] is a typical VFM designed to predict the related segmentation mask based on user-provided prompts. In contrast, SAM2 is specifically designed for dense prediction tasks and demonstrates a stronger capacity to preserve spatial structures and fine-grained semantic information. As an improved version of SAM, SAM2 adopts a hierarchical downsampling architecture to produce multiscale features, while achieving over six times faster inference speed. These advantages, which distinguish SAM2 from other VFMs, motivate this study to explore its adaptation to RSCD tasks. However, when applied to the RSCD task, SAM2 faces two main challenges.

1) *Domain Shift:* At the data level, there are significant differences between RS images and natural images in spatial resolution, imaging conditions, and object scale. RS images usually have low resolution, and the spectral overlap between ground objects is large, which makes it difficult to distinguish the foreground from the background. Directly employing SAM2 for the RSCD task forces the network to rely on its inherent general knowledge in natural scene, limiting its sensitivity to subtle changes in and RS scene. As shown in Fig. 1(a), directly employing SAM2 for the RSCD task leads to distinct false detections and missed detections due to the domain shift problem. Although VFMs adaptation strategies have made progress in mitigating domain shift, they still face certain limitations. Full tuning strategies are prone to overfitting [27], while adapter-based strategies increase model complexity [21]. In contrast, hierarchical low-rank adaptation (LoRA) [28] is a parameter-efficient fine-tuning (PEFT) method, has been scarcely explored in a systematic manner, limiting its full potential.
2) *Boundary Displacement:* At the feature level, simultaneously capturing category-discriminative information (to locate coarse change regions) and spatial-detail information (to preserve fine-grained boundaries) is crucial for exploring temporal differences between bi-temporal images. Simple feature fusion methods (e.g., concatenation) fail to adequately explore temporal differences for bi-temporal features. As shown in Fig. 1(b), inadequate exploration of temporal differences results in an obvious boundary displacement problem in detection results.

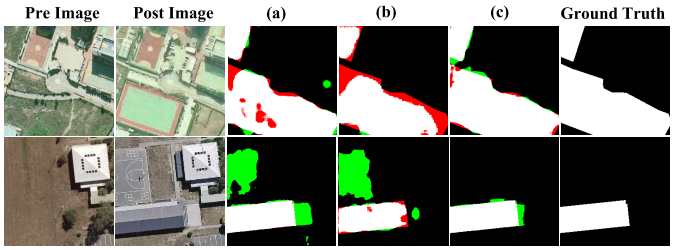


Fig. 1. Visual examples. (a) Results of directly adopting SAM2 as the feature extractor for RSCD. (b) Results of feature fusion with the concatenate operation. (c) Results of the proposed DA^2-Net. The rendered colors represent true positives (white), false positives (green), true negatives (black), and false negatives (red).

B. Overview

To alleviate the above challenges, this study proposes a SAM2-based domain adaptive and difference aggregation network (DA^2-Net) by exploring both domain adaptation and multiinformation cooperation. On the one hand, a hierarchical LoRA with PEFT strategy is designed to alleviate the domain shift problem. Specifically, the trainable low-rank decomposition matrices are introduced at key positions in the SAM2 image encoder, enabling the network to efficiently learn knowledge from the RS domain. On the other hand, a difference adaptive enhancement module (DAEM) is designed to alleviate the boundary displacement problem. Through a global difference awareness and local detail optimization dual-branch process, DAEM can adaptively aggregate bi-temporal features and enhance change perception. As shown in Fig. 1(c), the detection results by the proposed DA^2-Net have smoother edges and more uniform interiors than those in Fig. 1(a) and (b).

C. Contributions

In summary, the contributions of this study are threefold.

1) To bridge the knowledge gap between VFMs and the RSCD task, this study proposes DA^2-Net with a hierarchical LoRA fine-tuning strategy. Unlike other VFMs adaptation strategies, this study introduces low-rank matrices into multiple key layers of SAM2. It guides

the model to learn knowledge specific to the RS domain, thereby achieving domain adaptation of VFMs to the RS domain.

2) In order to coordinate category-discriminative and spatial-detail information, this study proposes DAEM. Unlike existing dual-branch fusion methods, DAEM generates complementary attention weights through a global difference awareness and local detail optimization dual-branch. It can adaptively enhance difference information and suppress the boundary displacement problem.
3) Extensive experiments demonstrate that DA^2-Net exhibits competitive cross-scene generalization and few-shot learning capabilities. Moreover, DA^2-Net can be easily extended to other multimodal vision tasks, such as RGB-SAR land use classification and RGB-thermal semantic segmentation.

D. Organization

The remainder of this study is organized as follows. Section II reviews mainstream DL-based RSCD methods, recent VFMs and adaptation strategies. In Section III, the framework of DA^2-Net is described in detail. The evaluation methods and experimental results are thoroughly explained and analyzed in Section IV. Section V discusses the interpretability, hyperparameters, and model efficiency. Finally, the work is summarized in Section VI.

II. Related Works

A. DL-Based CD Methods

The widespread application of DL techniques has driven the numerous DL-based RSCD networks. These networks can be roughly divided into CNN-based methods, transformer-based methods, and hybrid methods.

CNN-based RSCD methods [2], [29] leverage convolution operations to extract local spatial features and progressively build multiscale feature representations. Typically, the Siamese network is a mainstream architecture in RSCD tasks due to its ability to independently extract features from bi-temporal images. Specifically, Daudt et al. [30] first proposed two fully convolutional Siamese networks, FC-Conc and FC-Diff, which reuse shallow features to enrich deep feature representations, demonstrating their potential in RSCD. However, the two fully convolutional Siamese networks adopt limited feature fusion strategies and simple network structures, hindering the full modeling of differential information. To address this issue, Fang et al. [31] proposed SNUNet, which alleviates the loss of localization information in deep layers by employing dense skip connections between components. To capture richer contextual information, Lei et al. [2] subsequently designed a differential module with an attention mechanism. It can learn the differences between foreground and background, enhancing the internal consistency of the changed objects. Zhang et al. [32] combined a Siamese RSCD network with global attention and foreground-aware strategies, enhancing contextual learning and feature extraction capabilities. Considering the impact of environmental noise, Shi et al. [33] introduced deep supervision into the Siamese RSCD network. They utilized multiscale attention to provide more discriminative information to the network, making it more sensitive to fine-grained changes. Generally, CNN-based RSCD methods can efficiently extract local spatial features and adapt to various RSCD scenes through different attention mechanisms and multiscale representations. However, the convolution operation often neglects long-range dependencies between pixels [34], hindering accurate capture of large-scale semantic changes.

In recent years, the extensive application of transformers in the field of RS has brought new solutions to RSCD. The ViT can accurately capture global context information in images and establish associations between distant pixels through self-attention mechanisms. For example, Chen et al. [35] proposed BIT to efficiently model context information in the spatio-temporal domain by representing bi-temporal deep features as semantic tokens. Zhang et al. [16] proposed SwinSUNet, which processes bi-temporal images in parallel and extracts their multiscale features, achieving promising results. Bandara and Patel [36] integrated a hierarchical transformer encoder with a multilayer perceptron (MLP) decoder into the Siamese network architecture, effectively capturing multiscale remote details. While transformers effectively capture long-range dependencies, they may struggle to represent the fine-grained spatial detail structure of RS images, hindering the detection of small change objects.

In fact, the independent structures of CNNs and transformers limit the network's global-local modeling capabilities. Therefore, many RSCD networks are dedicated to leveraging the complementary advantages of CNNs and transformers. For example, Feng et al. [17] proposed ICIF-Net for capturing bi-temporal features with a dual-branch parallel structure of CNNs and transformers. Li et al. [37] presented ConvTransNet by integrating CNNs and transformers in parallel connection as a feature extraction module. Liu et al. [38] proposed AMTNet, which leverages a cross-fusion transformer to process the multiscale features output by ResNet [39].

Although the above DL-based RSCD methods have made significant progress, their data demands for training also increase with the expansion of the model scale. In RS scenarios with complex spatio-temporal relationships, their detection capabilities remain more limited.

B. Visual Foundation Models

VFMs refer to models that are pre-trained on large-scale datasets and with general knowledge. They exhibit strong generalization capabilities and can reduce data dependence for downstream tasks. For example, OpenAI proposed a vision-language model called CLIP [24], which has demonstrated comparable performance to fully supervised CNNs in zero-shot classification tasks. To enhance the cross-modal learning capabilities of CLIP, Dong et al. [40] proposed MaskCLIP, which incorporates masked self-distillation during training. However, CLIP and MaskCLIP struggle with the image segmentation task. To address this, Lan et al. [41] further introduced ClearCLIP for generating segmentation maps by removing noise-inducing components from CLIP.

With the advancement of computer vision, integrating multiple segmentation tasks into a single model has become increasingly attractive. For example, the SAM [25] is a segmentation model released by Meta that is capable of performing different types of segmentation tasks. Trained on ultra-large scale natural image data, SAM exhibits strong zero-shot generalization capabilities across various visual tasks. To optimize the computational efficiency and resource consumption of SAM, Meta introduced EfficientSAM [42]. By incorporating masked image pretraining, the model is enabled to operate efficiently even in resource-constrained environments. However, practical applications require handling time series data. Therefore, Meta released SAM2 [26], which extends SAM's capabilities from image segmentation to video segmentation.

In the RS domain, an increasing number of task-specific VFMs have been proposed. SAMRS [43] provides a large-scale RS image segmentation dataset, while Ma et al. [44] utilize SAM for fast RS image segmentation with minimal interaction. Additionally, several adaptation methods for VFMs have been proposed. For example, Zheng et al. [45] introduced a multicognitive visual adapter to facilitate the transfer learning of the SAM in RS semantic segmentation. Yuan et al. [46] designed a multimodal adapter for the RS image-text retrieval task, which enables cross-domain transfer of multimodal information through a shared-weight mechanism.

It is worth noting that current research on SAM2 remains limited; however, it still demonstrates strong application potential. In [47], the prompt-based segmentation capability of SAM2 was leveraged to convert an RS object tracking dataset into a video segmentation dataset. In [48], the authors leveraged the positive and negative click prompts of SAM2 to achieve high-resolution water body extraction. In [49], SAM2 was employed to generate pseudo-labels for object boundary regions, thereby improving the accuracy of multiview RS multiview segmentation. However, due to the significant domain shift between natural images and RS images, directly transferring SAM2 to RSCD fails to yield satisfactory performance. Moreover, temporal differences are particularly critical for RSCD, while SAM2 excels at segmentation but lacks the ability to compare temporal changes.

C. VFMs Adaptation Strategies

A reasonable integration of VFMs with downstream tasks can save substantial resources while significantly improving the performance ceiling and generalization ability of new models. With the continuous advancement of VFMs, three adaptation strategies have emerged, namely full tuning, adapters, and PEFT.

The essence of full tuning lies in leveraging existing knowledge to guide the learning process of downstream tasks, allowing all pre-trained weights to be updated during training. This approach enables direct sharing of low-level features across different tasks and can fit the feature distributions of out-of-distribution data. Owing to these advantages, full tuning has been widely applied in various RS scenarios, including segmentation [14], retrieval [50], and detection [51], achieving encouraging success. However, the widespread practice of full tuning was primarily on earlier pre-trained model (e.g., VGG and ResNet), which had relatively simple architectures and limited scale, and thus can no longer satisfy the performance demands of downstream applications. Although recent VFMs with hundreds of millions of parameters (e.g., SAM and CLIP) have emerged, the computational cost and risk of overfitting have constrained the development of full tuning.

Adapters are an effective alternative to full tuning. Its core idea is to introduce learnable modules into frozen VFMs to facilitate the optimization of downstream tasks under supervision signals. The standard adapter [52] typically consists of dimension reduction, activation, and dimension expansion, and is inserted between adjacent transformer layers. For RSCD tasks, BAN [21] treats the existing RSCD model as an external adapter, where multiscale features extracted by CLIP are fused with features extracted by the external adapter. SAM-CD [53] is another adapter-based RSCD method, which introduces convolutional modules into each layer of FastSAM and provides semantic information to the network through segmentation masks. The above methods typically treat VFMs as frozen feature extractors, lacking dynamic optimization of internal features, which limits their cross-domain understanding capabilities.

To overcome the limitations of adapters, it is necessary to perform PEFT within VFMs to mitigate the domain shift between natural images and RS imagery. In recent years, LoRA has attracted widespread attention as a PEFT method. LoRA decomposes a high-dimensional weight matrix into two low-rank matrices, updating only these low-rank matrices during training. This approach enables the network to learn inductive biases specific to the target data while significantly reducing the number of trainable parameters. In previous studies [54], [55], LoRA has been applied to RSCD tasks. However, these works merely inserted LoRA modules simply into every layer of the VFMs, lacking a systematic exploration of the optimal embedding locations and frequencies. Moreover, they did not thoroughly investigate the effects of LoRA's hyperparameters (e.g., rank, scaling factor, and dropout rate) on detection performance. Therefore, the full potential of LoRA in RSCD tasks remains to be further explored.

III. Method

A. Overall Framework

As shown in Fig. 2, the main components of DA^2-Net include: a hierarchical low-rank domain adaptive image encoder, a DAEM, and a residual convolutional decoder. First, bi-temporal images are input into the hierarchical low-rank domain adaptive image encoder for extracting multiscale contextual features. Each encoder block contains several transformer blocks. To reduce the domain knowledge gap, a trainable low-rank factorization matrix is introduced in the self-attention and MLP layers of the transformer block. Subsequently, DAEM processes detailed and discriminative information to output multilevel enhanced difference features. The difference features are then fed into the residual convolutional decoder to generate a feature map with

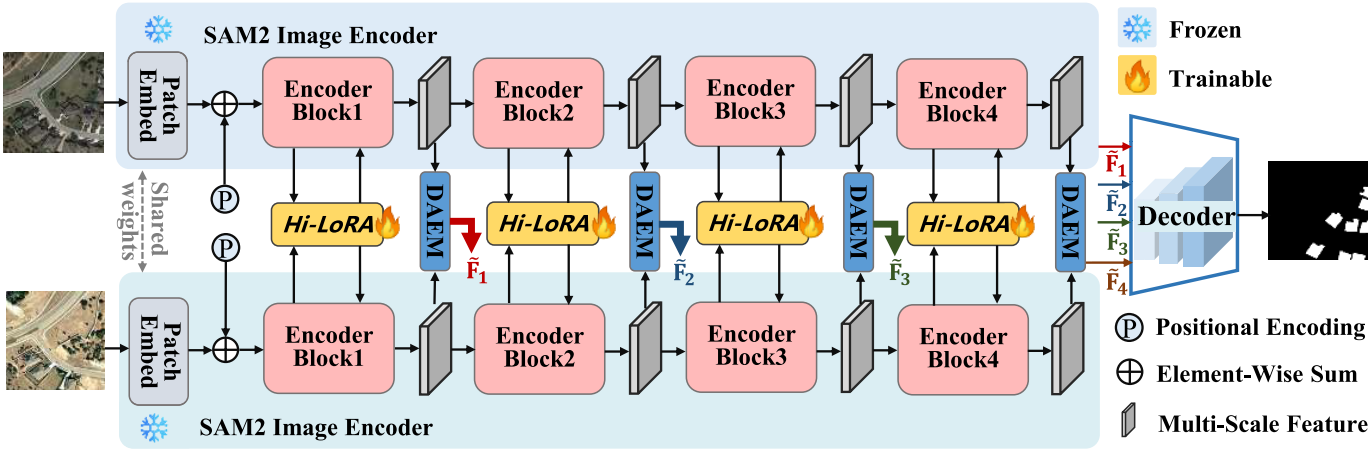


Fig. 2. Architecture of the proposed DA^2-Net. It mainly consists of three parts: the hierarchical low-rank domain adaptive image encoder, the DAEM, and the residual convolutional decoder.

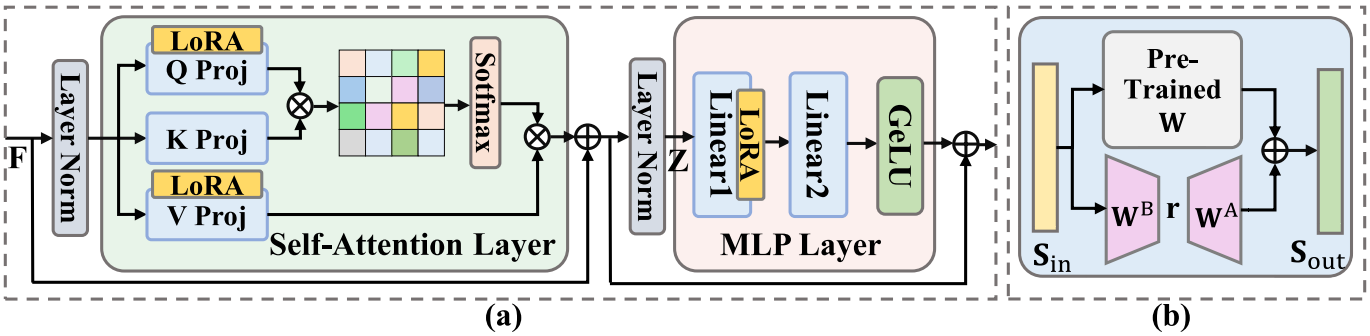


Fig. 3. Diagram of transformer block in encoder block. (a) Transformer block architecture after introducing hierarchical LoRA. (b) Schematic of the LoRA principle.

change-related semantic information. Finally, a convolutional layer is employed to obtain the RDCD results.

B. Hierarchical Low-Rank Domain Adaptive Image Encoder

SAM2 has exhibited outstanding performance in various downstream tasks after high-quality training on large-scale natural datasets. However, the domain gap between natural images and RS images restricts its performance in RSCD tasks. To extract the inductive bias of RS images and further enhance the generalization ability of SAM2, this study proposes a hierarchical LoRA strategy with PEFT. As shown in Fig. 3(a), driven by the functional heterogeneity of transformer components, LoRA is integrated into both the self-attention and MLP layers of the transformer module in the SAM2 image encoder. Specifically, in the self-attention layer, low-rank adjustments are applied to the weight matrices to optimize the adaptive weight distribution of the attention mechanism. This allows the model to capture the global relationships among input features more accurately. In the MLP layer, LoRA enables more effective adjustments to feature representations through nonlinear transformations, which facilitates improved adaptability of the model to the specific characteristics of RS data.

The principle of LoRA is illustrated in Fig. 3(b). Given the weight matrix $\mathbf{W} \in \mathbb{R}^{C_1 \times C_2}$ of a certain layer in the transformer block, an additional branch is introduced on one side of $\mathbf{W}$. The branch is responsible for decomposing $\mathbf{W}$ into two lower rank matrices $\mathbf{W}^{\mathrm{A}} \in \mathbb{R}^{r \times C_2}$ and $\mathbf{W}^{\mathrm{B}} \in \mathbb{R}^{C_1 \times r}$, where $r \ll \min\{C_1, C_2\}$. During training, the weights of $\mathbf{W}$ are frozen, and $\mathbf{W}^{\mathrm{A}}$ and $\mathbf{W}^{\mathrm{B}}$ are trained to approximate the updates to $\mathbf{W}$. Let the input feature be $\mathbf{S}_{\mathrm{in}}$, and the output feature be $\mathbf{S}_{\mathrm{out}}$. The process can be expressed as

$$\mathbf{S}_{\mathrm{out}} = \hat{\mathbf{W}}\mathbf{S}_{\mathrm{in}} \tag{1}$$

$$\hat{\mathbf{W}} = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{W}^{\mathrm{B}}\mathbf{W}^{\mathrm{A}} \tag{2}$$

where $\hat{\mathbf{W}}$ denotes the weight matrix after introducing LoRA, and $\Delta\mathbf{W}$ represents the weight matrix that is updated. In the multihead self-attention of the transformer module, the cosine similarity between different positions is computed to determine which features should be weighted. In this study, LoRA is applied to query and value projection layers for modulating the attention scores, thus adjusting the model to focus on different regions. The self-attention computation process after introducing LoRA is as follows:

$$\mathrm{Attention}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{C_1}}\right)\mathbf{V} \tag{3}$$

$$\mathbf{Q} = \hat{\mathbf{W}}_q\mathbf{F} = \mathbf{W}_q\mathbf{F} + \mathbf{W}_q^{\mathrm{B}}\mathbf{W}_q^{\mathrm{A}}\mathbf{F} \tag{4}$$

$$\mathbf{K} = \mathbf{W}_k\mathbf{F} \tag{5}$$

$$\mathbf{V} = \hat{\mathbf{W}}_v\mathbf{F} = \mathbf{W}_v\mathbf{F} + \mathbf{W}_v^{\mathrm{B}}\mathbf{W}_v^{\mathrm{A}}\mathbf{F} \tag{6}$$

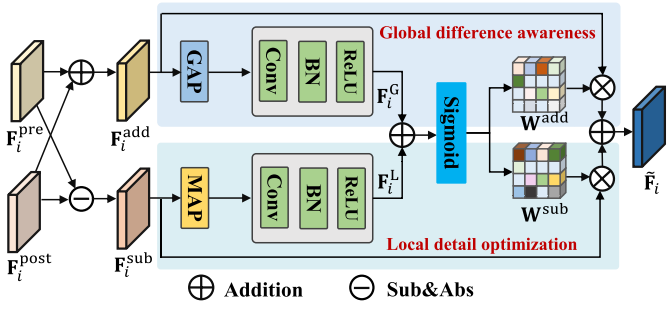


Fig. 4. Illustration of DAEM.

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are the weight matrices of the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ projection layers in SAM2. C_1 denotes the dimension of single-head attention, while $\mathbf{W}_q^{\mathrm{A}}$, $\mathbf{W}_q^{\mathrm{B}}$, $\mathbf{W}_v^{\mathrm{A}}$, and $\mathbf{W}_v^{\mathrm{B}}$ are the trainable low-rank matrices. $\mathbf{F}$ represents the multiscale contextual features extracted by the SAM2 image encoder.

In the MLP layer of the transformer module, nonlinear transformations enable the model to learn more complex feature representations. To capture the complex relationships between input features for the RS data, LoRA is introduced in the first linear layer of the MLP. The computation process of the linear layer after introducing LoRA is as follows:

$$\mathrm{MLP}(\mathbf{Z}) = \mathbf{W}_2\left(\hat{\mathbf{W}}_1\mathbf{Z} + \mathbf{b}_1\right) + \mathbf{b}_2 \tag{7}$$

$$\hat{\mathbf{W}}_1 = \mathbf{W}_1 + \mathbf{W}_1^{\mathrm{B}}\mathbf{W}_1^{\mathrm{A}} \tag{8}$$

where $\mathbf{Z}$ is the input sequence to the MLP layer. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices of the first and second linear layers in the MLP layer, respectively. $\mathbf{W}_1^{\mathrm{A}}$ and $\mathbf{W}_1^{\mathrm{B}}$ are the low-rank matrices from the decomposition of $\mathbf{W}_1$. $\mathbf{b}_1$ and $\mathbf{b}_2$ represent the bias terms.

C. Difference Adaptive Enhancement Module

Traditional feature fusion methods (e.g., differential or concatenate approaches) often overlook the synergistic relationship between change awareness and detail capture. This makes it challenging for RSCD networks to simultaneously preserve edge details of changed objects and internal tightness. To address this issue, this study proposes the DAEM. Unlike single-branch feature fusion strategies [35], [37], [36], [56], DAEM can accurately capture bi-temporal difference information through the collaboration of two branches: global difference awareness and local detail optimization.

The DAEM employs two distinct branches to capture category-discriminative information via global attention and spatial-detail information via local attention, respectively. Each branch generates corresponding attention weights for subsequently adaptively integrating bi-temporal features and obtaining difference features. As shown in Fig. 4, the bi-temporal features are denoted as $\mathbf{F}_i^{\mathrm{pre}}$ and $\mathbf{F}_i^{\mathrm{post}}$ ($i = 1, 2, 3, 4$ indexes the ith encoder block). The difference features enhanced by DAEM are denoted as $\tilde{\mathbf{F}}_i$. The operation steps of DAEM are as follows: First, $\mathbf{F}_i^{\mathrm{pre}}$ and $\mathbf{F}_i^{\mathrm{post}}$ are element-wise added to obtain $\mathbf{F}_i^{\mathrm{add}}$, and then $\mathbf{F}_i^{\mathrm{pre}}$ and $\mathbf{F}_i^{\mathrm{post}}$ are element-wise subtracted to obtain $\mathbf{F}_i^{\mathrm{sub}}$. Subsequently, $\mathbf{F}_i^{\mathrm{add}}$ and $\mathbf{F}_i^{\mathrm{sub}}$ are fed into the global difference awareness and optimization branches, respectively, for learning category-discriminative and spatial-detail information.

In the global difference awareness branch, the global average pooling operation is first applied to each feature channel in $\mathbf{F}_i^{\mathrm{add}}$ for learning the global attention feature $\mathbf{F}_i^{\mathrm{G}}$. The operation reduces the spatial dimensions of $\mathbf{F}_i^{\mathrm{add}}$ from $[C, H, W]$ to $[C, 1, 1]$, where C, H, and W represent the channel number, height, and width of the feature map, respectively. The global attention feature $\mathbf{F}_i^{\mathrm{G}}$ can be computed as follows:

$$\mathbf{F}_i^{\mathrm{add}} = \begin{cases} \mathbf{F}_i^{\mathrm{pre}} \oplus \mathbf{F}_i^{\mathrm{post}}, & i = 1 \\ \mathrm{UP}\left(\mathbf{F}_i^{\mathrm{pre}}\right) \oplus \mathrm{UP}\left(\mathbf{F}_i^{\mathrm{post}}\right), & i > 1 \end{cases} \tag{9}$$

$$\mathbf{F}_i^{\mathrm{G}} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv1}\left(\mathrm{GAP}\left(\mathbf{F}_i^{\mathrm{add}}\right)\right)\right)\right) \tag{10}$$

where $\oplus$ denotes element-wise addition, UP stands for a $2\times$ upsampling operation, GAP represents the global average pooling operation, Conv1 refers to a 1×1 convolution operation, BN stands for batch normalization, and ReLU(X) = max{0, X}.

In the local detail optimization branch, a max pooling operation is employed to extract the maximum value within each local receptive field to emphasize salient local details. The local attention feature $\mathbf{F}_i^{\mathrm{L}}$ is calculated as follows:

$$\mathbf{F}_i^{\mathrm{sub}} = \begin{cases} |\mathbf{F}_i^{\mathrm{pre}} \ominus \mathbf{F}_i^{\mathrm{post}}|, & i = 1 \\ |\mathrm{UP}\left(\mathbf{F}_i^{\mathrm{pre}}\right) \ominus \mathrm{UP}\left(\mathbf{F}_i^{\mathrm{post}}\right)|, & i > 1 \end{cases} \tag{11}$$

$$\mathbf{F}_i^{\mathrm{L}} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv1}\left(\mathrm{MAP}\left(\mathbf{F}_i^{\mathrm{sub}}\right)\right)\right)\right) \tag{12}$$

where $\ominus$ denotes element-wise subtraction, $|\cdot|$ represents the absolute value operation, and MAP stands for max pooling operation.

Then, $\mathbf{F}_i^{\mathrm{G}}$ and $\mathbf{F}_i^{\mathrm{L}}$ are added together and passed through a sigmoid function to obtain the attention weight matrices $\mathbf{W}_i^{\mathrm{add}}$ and $\mathbf{W}_i^{\mathrm{sub}}$. The details are as follows:

$$\mathbf{W}_i^{\mathrm{add}} = \mathrm{Sigmoid}\left(\mathrm{Conv1}\left(\mathbf{F}_i^{\mathrm{G}} \oplus \mathbf{F}_i^{\mathrm{L}}\right)\right) \tag{13}$$

$$\mathbf{W}_i^{\mathrm{sub}} = \mathbf{1} - \mathbf{W}_i^{\mathrm{add}}. \tag{14}$$

Finally, $\mathbf{W}_i^{\mathrm{add}}$ and $\mathbf{W}_i^{\mathrm{sub}}$ are utilized as the global and local attention weights, respectively, for obtaining the final enhanced features. Specifically, the extracted global attention feature and local attention feature are multiplied by their corresponding pixel-level attention weights and then integrated by element-wise addition to obtain the enhanced feature $\tilde{\mathbf{F}}_i$. It can be calculated as follows:

$$\tilde{\mathbf{F}}_i = \left(\mathbf{W}_i^{\mathrm{add}} \odot \mathbf{F}_i^{\mathrm{G}}\right) \oplus \left(\mathbf{W}_i^{\mathrm{sub}} \odot \mathbf{F}_i^{\mathrm{L}}\right) \tag{15}$$

where $\odot$ denotes element-wise multiplication.

D. Residual Convolutional Decoder and Optimization Strategy

To obtain the final RSCD results, this study introduces a simple yet efficient residual convolutional decoder. The decoder details are shown in Fig. 5. Initially, the $\tilde{\mathbf{F}}_i$ ($i = 1, 2, 3, 4$) are concatenated along the channel dimension and conducted a 1×1 convolution operation. Subsequently, to

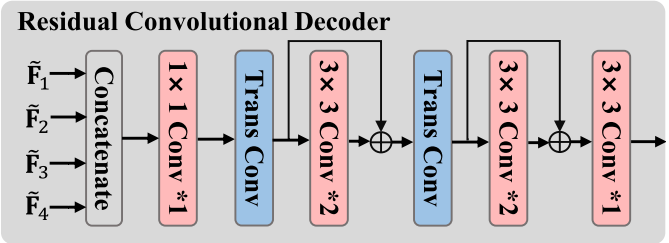


Fig. 5. Illustration of residual convolutional decoder.

avoid the checkerboard effect [57] during upsampling, a cross-structured combination of "transposed convolution + 3 × 3 residual convolution" is adopted to achieve feature upsampling and refinement. Finally, a 3 × 3 convolution operation is applied to obtain the final change prediction probability map, where the first channel corresponds to no change and the second channel corresponds to the change class probability. To obtain the binary change map, an argmax operation is performed along the channel dimension.

The RSCD is essentially a dense prediction task. During training, this study optimizes the network by minimizing the cross-entropy loss. The cross-entropy loss is as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right) \tag{16}$$

where y_i is the true class label of the sample, $\hat{y}_i$ is the predicted class label, and N is the total number of pixels in each sample.

IV. Experiments

A. Dataset and Evaluation Metrics

1) Dataset Description: The proposed DA^2-Net is validated on three public CD datasets: SYSU-CD [33], WHU-CD [58], and LEVIR-CD [59].

SYSU-CD dataset contains 20 000 pairs of high-resolution (0.5 m/pixel) aerial images with spatial dimensions of 256 × 256, captured between 2007 and 2014. The types of changes in the dataset include newly constructed urban buildings, suburban expansion, pre-construction groundwork, vegetation changes, road expansion, and offshore construction. In our experiments, the default data split is adopted, with the training, validation, and test sets consisting of 12 000, 4000, and 4000 image pairs, respectively.

WHU-CD dataset contains a pair of high-resolution (0.2 m/pixel) aerial images with spatial dimensions of 32 507 × 15 354. The WHU-CD dataset is specifically designed for building CD. Since no default partitioning strategy is provided, this study follows the settings used in [21], where the original images are cropped with non-overlapping 256 × 256 patches. The training, validation, and test sets consist of 5947, 743, and 744 image pairs, respectively.

LEVIR-CD contains 637 pairs of high-resolution (0.5 m/pixel) images with spatial dimensions of 1024 × 1024. The period span for each pair of images ranges from 5 to 14 years. This dataset primarily focuses on changes related to building construction and demolition. In our experiments, the default data split is adopted, and the original images are cropped into non-overlapping 256 × 256 patches. The training, validation, and test sets consist of 7120, 1024, and 2048 image pairs, respectively.

2) Evaluation Metrics: To quantitatively evaluate the effectiveness of DA^2-Net, this study adopted a comprehensive set of performance metrics, including the $F1$-score ($F1$), precision (Pre.), recall (Rec.), overall accuracy (OA), and intersection over union (IoU)

$$\text{Pre.} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{17}$$

$$\text{Rec.} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{18}$$

$$F1 = 2\cdot\frac{\text{Pre}\cdot\text{Rec}}{\text{Pre}+\text{Rec}} \tag{19}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{20}$$

$$\text{OA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{21}$$

where rmTP, rmTN, rmFP, and rmFN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

B. Implementation Details

DA^2-Net leverages the SAM2-large image encoder as the feature extractor and outputs multiscale features at the 2nd, 8th, 44th, and 48th transformer module layers. During the training phase, various data augmentation strategies, such as random flipping and random rotation are applied to increase the diversity of the training dataset. The pre-trained weights of the network's backbone are kept frozen to preserve the inherent general knowledge. The network is trained on an NVIDIA 4090 GPU. The low-rank matrices $\mathbf{W}^{\text{A}}$ and $\mathbf{W}^{\text{B}}$ are initialized to 0 and 1, and the remaining trainable parameters are randomly initialized. Based on empirical settings, the learning rate is set to 2.1e^{-4}, and the model is trained for 150 epochs. The learning rate is linearly decayed until the final epoch, using the AdamW optimizer with weight decay of 0.01 and beta values of (0.9, 0.999). To achieve optimal model performance, the hierarchical LoRA introduction frequency $N = 2$, and the rank value $R = 16$ (for all datasets). On the SYSU-CD and LEVIR-CD datasets, the scale factor $\alpha = 16$ and the dropout rate $\beta = 0$, while on the WHU dataset, the scale factor $\alpha = 32$ and the dropout rate $\beta = 0.1$. The above hyperparameter discussion and analysis are presented in Sections IV-D3 and V-B.

C. Comparison With State-of-the-Art Methods

In this study, the proposed DA^2-Net is compared with 11 representative CD methods, including CNN-based methods (FC-EF [30], FC-Diff [30], USSFC [60]), transformer-based methods (BIT [35], ChangeFormer [36], LSAT [61], ELGC-Net [56]), and VFMs-based methods (ChangeViT [27], BAN [21], SAM-CD [53], Meta-CD [62]).

FC-EF is a fully convolutional RSCD network that fuses bi-temporal images at an early stage and then adopts a U-structure to perform RSCD. FC-Diff is a variant of FC-EF that fuses the bi-temporal features output by the Siamese network during

TABLE I

QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON SYSU-CD, WHU-CD, AND LEVIR-CD DATASETS. ALL THE VALUES ARE IN %

Type	Method	SYSU-CD					WHU-CD					LEVIR-CD				
		Pre.	Rec.	IoU	F1	OA	Pre.	Rec.	IoU	F1	OA	Pre.	Rec.	IoU	F1	OA
CNN	FC-EF	74.32	75.84	60.09	75.07	86.02	71.63	67.25	53.11	69.37	97.61	86.91	80.17	71.53	83.40	98.39
	FC-Diff	79.13	61.21	56.96	72.57	82.11	47.33	77.66	41.66	58.81	95.63	89.53	83.31	75.92	86.31	98.67
	USSFC	82.18	74.49	64.13	78.15	90.18	93.34	89.39	84.03	91.32	99.32	90.46	89.75	82.00	90.11	98.99
Trans.	CFormer	81.15	71.65	61.43	76.11	89.39	93.95	88.54	83.76	91.16	99.54	92.05	88.81	82.48	99.40	99.04
	BIT	81.54	79.11	65.41	78.33	90.42	93.34	89.39	84.03	91.32	99.31	89.24	89.37	80.68	89.31	98.92
	ELGC-Net	81.42	77.69	65.99	79.51	90.55	92.72	92.57	86.30	92.64	99.06	92.25	89.67	83.40	90.94	99.06
	LSAT	81.92	79.69	67.77	80.79	91.06	89.02	93.41	86.16	92.56	**99.56**	91.54	88.11	81.48	89.79	99.09
VFMs	ChangeViT	83.29	**80.25**	**69.12**	**81.74**	91.55	94.61	92.17	**87.58**	**93.37**	99.48	91.95	91.08	84.36	91.52	**99.14**
	SAM-CD	85.56	73.22	65.17	78.91	90.77	91.27	91.34	84.01	91.31	99.31	**93.71**	**91.92**	85.01	**92.09**	98.92
	BAN	**87.95**	74.31	67.44	80.55	**91.91**	**97.05**	89.76	87.37	93.26	99.47	**93.52**	91.16	**85.49**	91.48	99.10
	Meta-CD	**87.97**	69.52	63.48	77.66	90.57	94.58	**93.54**	86.78	92.80	99.29	92.00	89.53	83.07	90.75	99.07
	DA^2-Net	86.81	**82.04**	**72.94**	**84.35**	**92.82**	**95.99**	**94.96**	**91.34**	**95.47**	**99.64**	93.28	**91.51**	**85.87**	**92.16**	**99.21**

TABLE II

IMPACT OF THE COMBINATION OF DIFFERENT COMPONENTS ON THE PERFORMANCE OF DA^2-NET [PARAM (M), PRE. (%), IOU (%), AND $F1$ (%)]. ✓ ADDS THE COMPONENT, AND × REMOVES THE COMPONENT. WITHOUT DAEM, THE DIFFERENCE FEATURES ARE OBTAINED BY ELEMENT-WISE SUBTRACTION. MARK * DENOTES TRAINABLE PARAMETERS IN NETWORK

Method	Hierarchical LoRA		DAEM		Param*	SUSU-CD			WHU-CD		
	sLoRA	mLoRA	global	local		Pre.	IoU	F1	Pre.	IoU	F1
Strategy1	×	×	×	×	5.27	81.57	64.17	78.17	91.24	78.68	88.80
Strategy2	✓	×	×	×	6.09	81.96	66.75	81.41	92.98	86.91	94.49
Strategy3	✓	✓	×	×	6.52	81.78	69.93	82.31	94.08	88.33	93.85
Strategy4	✓	✓	✓	×	6.79	**83.54**	**70.47**	**82.90**	**94.89**	**90.76**	**94.55**
DA^2-Net	✓	✓	✓	✓	6.85	**86.81**	**72.94**	**84.35**	**95.99**	**91.34**	**95.47**

the decoding phase to perform RSCD. USSFC is an efficient ultra-lightweight spatial-spectral feature collaborative network, which introduces a 3-D attention mechanism for flexible capture of transformed features. BIT converts bi-temporal image features into semantic tokens and leverages a transformer encoder to model contextual tokens. ChangeFormer (denoted by CFormer) effectively models multiscale long-range information by using a hierarchical transformer encoder and a lightweight MLP decoder. LSAT is a lightweight structure-aware transformer network, which utilizes cross-dimensional interactive self-attention to replace the ordinary self-attention module in the visual transformer to focus on key regions. ELGC-Net introduces an efficient local-global context aggregator module in the transformer encoder. It captures enhanced global context and local spatial information through novel pooling-transpose attention and deep convolution mechanisms. ChangeViT integrates DINOv2 [63] with ResNet, it unleashes the potential of VFMs in RSCD through full-parameter fine-tuning. BAN is a model-agnostic concept that extracts general features from a frozen VFMs, which are then aligned and injected into existing RSCD models. SAM-CD utilizing the visual encoder of FastSAM to extract visual representations from RS scenes, further improving RSCD performance by incorporating segmentation semantic information. Meta-CD introduces an additional convolutional encoder to FastSAM, significantly improving the model's cross-scene generalization capability.

1) Quantitative Comparison: The performance of DA^2-Net and the other methods on the three datasets is summarized in Table I. The higher IoU and $F1$ indicate better detection performance, with the best result in each column highlighted in red and bold, and the second best in blue. Distinctly, DA^2-Net performs the best on the SYSU-CD dataset, with Recall, IoU, and $F1$ improved by 1.79%, 3.82%, and 2.61%, respectively, compared to ChangeViT. DA^2-Net also shows excellent performance on the WHU-CD dataset, with Recall and IoU improved by 5.20% and 3.97%, respectively, compared to BAN. The results on the LEVIR-CD dataset, reveal that DA^2-Net excels in both IoU and OA, with IoU and OA improved by 2.80% and 0.14%, respectively, compared to Meta-CD. Overall, VFMs-based methods outperform traditional CNN or transformer-based approaches. This is because VFMs leverage additional prior knowledge, which enables better performance for RSCD. The experimental results above demonstrate that DA^2-Net outperforms other VFMs-based RSCD methods, validating the superiority of the proposed method.

2) Visual Analysis: To further demonstrate the superiority of the proposed DA^2-Net, a visual comparison is shown in Fig. 6. It can be observed that most methods exhibit a boundary displacement problem in the detection of large-scale objects [Fig. 6(a)–(c) and (g)], while DA^2-Net shows complete interiors and smooth boundaries. Specifically, in cases with dense change objects [Fig. 6(f) and (i)], DA^2-Net detects more complete results compared to other methods, with virtually no

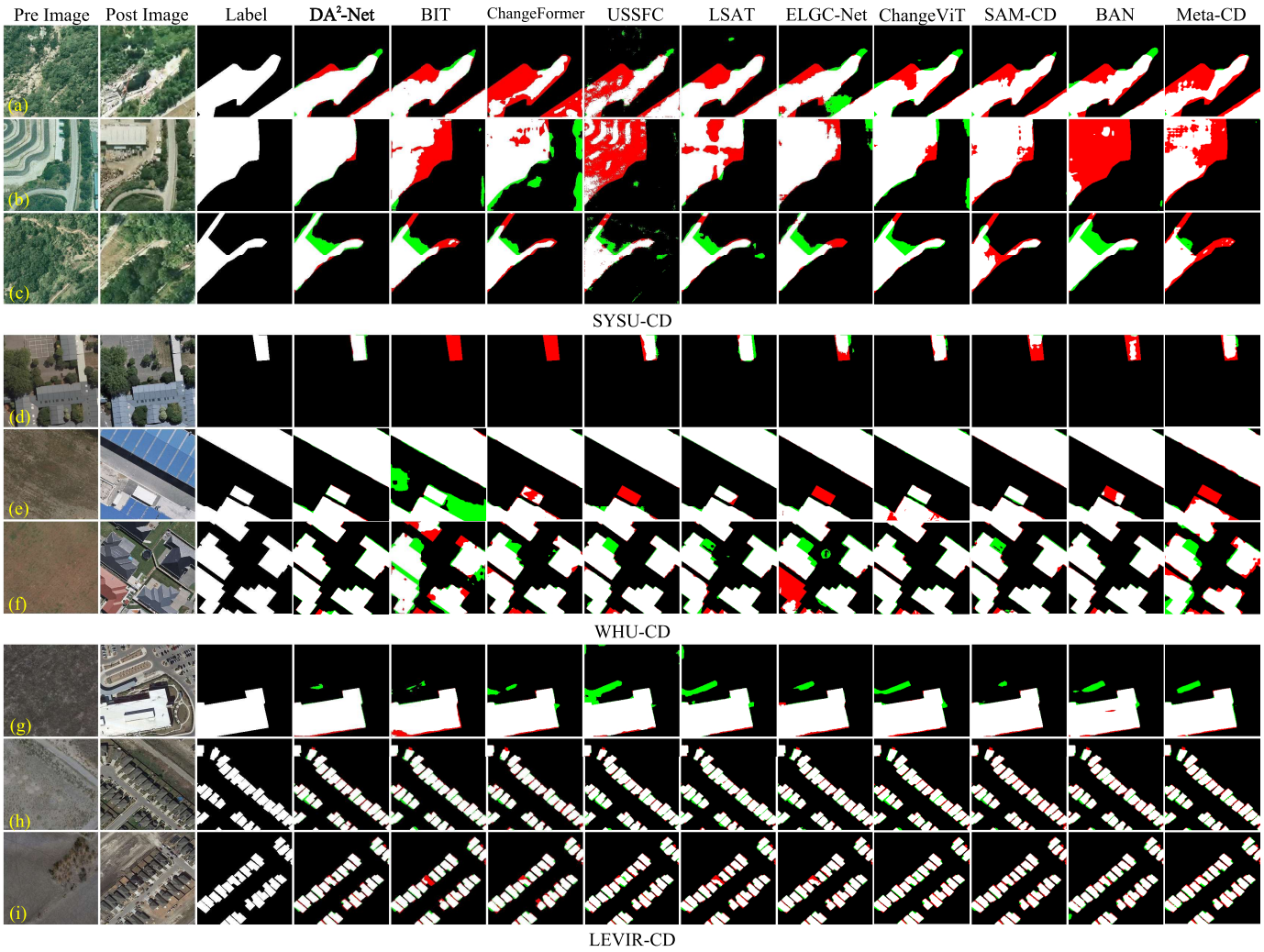


Fig. 6. Visual comparison of different methods on SYSU-CD, WHU-CD, and LEVIR-CD datasets. To present the results more clearly, white, green, black, and red represent TP, FP, TN, and FN, respectively.

missed or false detections. In cases with sparse change objects [Fig. 6(d) and (h)], DA^2-Net accurately detects all change objects and provides clear details, whereas other methods have false detections. Notably, although other methods all detect the change of the blue building in Fig. 6(e), they fail to accurately identify the white building that is similar to the surrounding environment. This suggests that DA^2-Net has greater semantic perception capabilities for changing objects in complex environments. Overall, the visualization results underscore the superiority of DA^2-Net in detecting structurally complex and densely populated change scenes, demonstrating clear advantages over existing methods.

D. Ablation Experiments

1) Different Components: To validate the effectiveness of the proposed hierarchical LoRA strategy and DAEM, ablation experiments were conducted for each component. Specifically, sLoRA and mLoRA represent the application of LoRA in the self-attention layer and the MLP layer of the transformer module, respectively. Considering both the scale of the datasets and the number of change objects, experiments were carried out on the WHU-CD and SYSU-CD datasets, and the results are shown in Table II. The pre-trained SAM2 image encoder with weight sharing was used as the baseline. From the experimental results, it can be observed that the model's performance improves progressively with the inclusion of different components.

Comparing the first three strategies in Table II, the baseline model cannot effectively identify the change region due to the relative lack of RS domain knowledge contained in SAM2. With the progressive introduction of sLoRA and mLoRA, these trainable parameters significantly enhance the model's ability to learn RS domain knowledge. On the WHU-CD dataset, compared to the Strategy1, precision, IoU, and $F1$ of Strategy3 are increased by 2.84%, 9.65%, and 5.05%, respectively. This demonstrates the effectiveness of the hierarchical LoRA strategy. Observing Strategy3 and Strategy2, the naive difference fusion strategy exhibits limited representational power for features, resulting in lower model accuracy. Strategy4 integrates a global difference awareness branch, thereby strengthening the network's ability to capture and discriminate

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED HIERARCHICAL LORA AGAINST OTHER VFMS ADAPTATION METHODS. TESTED ON SYSU-CD

VFMs	Fine-tune	Param*	Pre.	IoU	F1
DINOv2	Full	27.13	55.51	51.09	67.63
	Freeze	5.45	57.68	52.40	68.77
	Adapter	6.05	59.26	54.88	70.48
	LoRA	5.74	61.52	**56.75**	72.68
	DoRA	5.75	61.82	56.40	71.97
	Mona	6.91	**63.19**	56.32	**72.91**
	Ous	5.67	58.19	53.29	69.53
ESAM	Full	29.25	79.90	64.78	78.63
	Freeze	5.45	83.46	66.18	79.65
	Adapter	6.05	84.21	68.66	79.68
	LoRA	5.74	85.02	68.47	79.83
	DoRA	5.75	84.36	68.90	82.05
	Mona	6.91	85.07	68.30	81.17
	Ous	5.67	**84.30**	**70.78**	**82.89**
SAM2	Full	217.76	79.42	65.17	78.91
	Freeze	5.61	82.31	69.51	82.01
	Adapter	7.32	85.80	72.83	84.28
	LoRA	7.30	86.37	72.13	83.47
	DoRA	7.34	86.00	72.42	84.01
	Mona	13.82	83.02	50.84	67.41
	Ous	6.85	**86.81**	**72.94**	**84.35**

TABLE IV
IMPACT OF HIERARCHICAL LORA INTRODUCTION FREQUENCY AND RANK VALUE ON THE PERFORMANCE OF DA^2-NET [PARAM (M), PRE. (%), IOU (%), AND $F1$ (%)]. TESTED ON SYSU-CD AND LEVIR-CD

Method	Param*	SYSU-CD			LEVIR-CD		
		Pre.	IoU	F1	Pre.	IoU	F1
N=1 R=8	6.89	**88.79**	72.58	84.11	91.58	84.36	92.11
N=1 R=16	8.17	85.69	72.25	83.10	92.91	85.06	92.51
N=1 R=32	10.73	88.25	66.79	80.09	93.01	85.47	92.39
N=2 R=8	6.23	85.35	71.84	83.61	92.89	85.11	**92.65**
N=2 R=16	6.85	86.81	**72.94**	**84.35**	**93.28**	**85.87**	92.16
N=2 R=32	8.10	87.75	71.27	83.22	93.10	85.45	92.03
N=3 R=8	5.91	86.02	72.10	83.78	93.15	84.96	91.87
N=3 R=16	6.20	87.72	71.99	83.71	92.69	85.16	92.36
N=3 R=32	6.79	86.82	71.75	83.55	93.24	85.04	91.16

genuine changes. With the introduction of the local detail optimization branch, DAEM is capable of adaptively weighting spatial detail features and class discriminative features, thereby enhancing the model's spatial localization capabilities. Compared to Strategy3, the IoU of DA^2-Net on the SYSU-CD dataset increased by 3.01%, which sufficiently validates the effectiveness of DAEM. Additionally, the learnable parameters of DA^2-Net constitute only 1.5% of the encoder network. The introduction of a small number of parameters has enabled the domain adaptation of SAM2's general knowledge to the RS domain, significantly enhancing the performance of RSCD models.

2) Different VFMs and Fine-tuning: To evaluate the generalization capability of the proposed hierarchical LoRA and DAEM across different VFMs, replaced the backbone of DA^2-Net with two representative VFMs. Specifically, the EfficientSAM-s (denoted by ESAM in Table III) and DINOv2-s variants were introduced, retaining only their image encoders. As shown in Table III, the proposed method achieved a notable performance gain when integrated with EfficientSAM-s, reaching an IoU of 70.78% on the SYSU-CD dataset. In contrast, the performance improvement was less pronounced when using DINOv2-s as the backbone, which may be attributed to the greater heterogeneity between its training data and RS images. These results demonstrate that the proposed method exhibits robust adaptability, it can deliver effective performance gains without reliance on a specific backbone architecture.

Furthermore, we compared the hierarchical LoRA (Ours) with several popular VFMs fine-tuning strategies, including full-parameter fine-tuning (Full), freezing, Adapter [52], LoRA [28], DoRA [64], and Mona [65]. As shown in Table III, full-parameter fine-tuning despite involving the largest number of trainable parameters, yielded the lowest performance. This may result from overfitting to specific samples during full-parameter fine-tuning, which compromises the general representation ability of the VFMs. In contrast, freezing all backbone parameters (Freeze) preserved the semantic parsing capabilities of the original VFMs, leading to improved performance. Introducing lightweight fine-tuning approaches enabled the injection of RS-specific inductive biases with a limited number of trainable parameters. Among them, Mona introduced the highest parameter count and helped DINOv2 achieve the highest precision (63.19%) and $F1$ (72.91%). In the cases where EfficientSAM and SAM2 were used as backbones, hierarchical LoRA required the fewest trainable parameters while delivering the best performance. Notably, it outperformed DoRA on EfficientSAM, achieving an $F1$ of 82.89%. When applied to SAM2, our method improved precision by 0.81%.

3) LoRA Introduction Frequency and Rank Value: Inspired by [66], it is not necessary to introduce trainable parameters into each layer of VFMs. Additionally, the rank value has a direct impact on both model performance and resource consumption in the context of LoRA fine-tuning. To validate the combined effect of these two factors on model performance, extensive experiments were conducted, and the results are shown in Table IV. N indicates the frequency of introducing hierarchical LoRA once every N transformer blocks in the encoder, and R represents the rank value. It can be observed that the model performance reaches its peak when $N = 2$ and $R = 16$. Notably, comparing the three strategies listed in rows 1, 5, and 9, it can be seen that the model performs best when $R = 16$, despite having nearly identical parameters. The analysis is as follows: the first strategy, despite having broader parameter coverage, suffers from the limited optimization capability of hierarchical LoRA under lower rank conditions. This makes it difficult to efficiently model contextual features. The third

TABLE V

IoU (%) AND OA (%) OF DA^2-NET AND SEMI-SUPERVISED CD METHODS UNDER DIFFERENT PROPORTIONS OF TRAINING DATA. TESTED ON LEVIR-CD AND WHU-CD

Method	LEVIR-CD				WHU-CD			
	5%	10%	20%	40%	5%	10%	20%	40%
S4GAN	64.0/97.89	67.0/98.11	73.4/98.51	75.4/98.62	18.3/96.69	62.6/98.15	70.8/98.60	76.4/98.96
SemiCDNet	67.6/98.17	71.5/98.42	74.3/98.58	75.5/98.63	51.7/97.71	62.0/98.16	66.7/98.28	75.9/98.93
SemiCD	72.5/98.47	75.5/98.63	76.2/98.68	77.2/98.72	65.8/98.37	68.1/98.47	74.8/98.84	77.2/98.96
UniMatch	**80.7/98.95**	**82.0/99.02**	**81.7/99.02**	**82.1/99.03**	**80.2/99.15**	**81.7/99.22**	**81.7/99.18**	**85.1/99.35**
DA^2-Net	**80.0/98.92**	**81.1/98.97**	**82.5/99.06**	**84.1/99.12**	**80.1/99.10**	**81.4/99.20**	**85.0/99.42**	**88.1/99.46**

TABLE VI

CROSS-SCENE GENERALIZATION PERFORMANCE OF DIFFERENT METHODS. LEVIR-CD TO WHU-CD. ALL THE VALUES ARE IN %

Type	Method	Pre.	IoU	F1	OA
CNN	FC-EF	21.97	21.36	35.20	89.07
	FC-Diff	32.16	24.18	33.96	88.32
	USSFC	35.70	25.46	40.59	94.53
Trans.	CFormer	55.26	52.13	68.54	97.92
	BIT	48.55	41.26	58.42	95.85
	ELGC-Net	74.68	56.59	72.28	97.86
	LSAT	53.96	33.42	50.10	96.31
VFMs	ChangeViT	72.85	58.04	73.45	97.87
	SAM-CD	77.67	60.11	75.08	98.08
	BAN	78.25	60.31	75.24	98.10
	Meta-CD	78.18	62.92	77.24	98.21
	DA^2-Net	**87.69**	**65.96**	**79.49**	**98.51**

TABLE VII

COMPARISON OF DA^2-NET WITH OTHER METHODS IN TERMS OF LEARNABLE PARAMETERS [PARAMETERS (M)], FLOPS (G), AND INFERENCE TIME (S). THE $F1$ (%) SCORE IS BASED ON THE RESULTS FROM THE WHU-CD DATASET

Type	Method	Param*	FLOPs	Infer	F1
CNN	FC-EF	1.54	3.57	0.046	69.37
	FC-Diff	1.66	5.09	0.099	58.81
	USSFC	1.52	9.72	0.309	83.98
Trans.	CFormer	11.94	17.50	0.234	91.16
	BIT	3.50	21.26	0.208	91.32
	ELGC-Net	10.56	186.96	0.293	92.64
	LSAT	18.93	7.64	0.247	92.56
VFMs	ChangeViT	41.14	77.61	2.424	93.37
	SAM-CD	2.59	17.19	2.538	91.31
	BAN	4.23	353.72	0.261	93.26
	Meta-CD	13.67	47.73	3.50	92.80
	DA^2-Net	6.85	286.25	0.731	**95.47**
	DA^2-Net-t	0.82	27.85	0.100	**94.33**

strategy exhibits strong expressive capability in a single transformer block. However, the sparse distribution of parameters leads to insufficient optimization of some key feature layers, weakening the ability to express both global and local features. The proposed DA^2-Net is based on the second strategy, where LoRA introduces a balanced setting of frequency and rank values to achieve optimal model performance.

E. Performance Evaluation With Limited Labeled Data

The annotation of RSCD datasets is typically time-consuming and labor-intensive. Semi-supervised methods mitigate this issue by leveraging a small set of labeled samples and a large number of unlabeled samples for training [67], achieving good detection performance. To verify the learning ability of DA^2-Net in scenes with limited labeled data, this section compares it with four other advanced semi-supervised RSCD methods: S4GAN [68], SemiCDNet [10], SemiCD [69], and UniMatch [70]. In this experiment, DA^2-Net and the methods above were all trained using 5%, 10%, 20%, and 40% of the labeled data from the training set. This comparison setting is more challenging for DA^2-Net, as semi-supervised methods typically rely on joint learning from both labeled and unlabeled data, along with complicated training strategies. In contrast, DA^2-Net is trained solely with a small fraction of labeled data using a standard supervised learning paradigm.

The experimental results are shown in Table V. UniMatch is an advanced semi-supervised semantic segmentation network that also demonstrates excellent performance in the RSCD task. It achieves the best performance when using 5% and 10% of the training data. Under conditions where training data is scarce, due to the lack of RS-related knowledge in low-rank matrices, DAEM, and the decoder, they still require sufficient data for learning. It leads to DA^2-Net having lower accuracy than UniMatch when using 5% and 10% of the training data. When training data is abundant, the potential of each component is fully exploited, and the highest accuracy is achieved when using 20% and 40% of the training data. In summary, under the same data conditions, the performance of DA^2-Net has already surpassed most semi-supervised CD networks, demonstrating the potential of the proposed DA^2-Net under conditions with a limited number of labeled samples.

F. Performance Evaluation in Cross-Scene

To further evaluate the generalization capability of DA^2-Net, this study conducts a zero-shot performance test. Without any fine-tuning or domain adaptation, the model is directly applied to a completely unseen dataset. Specifically, the model

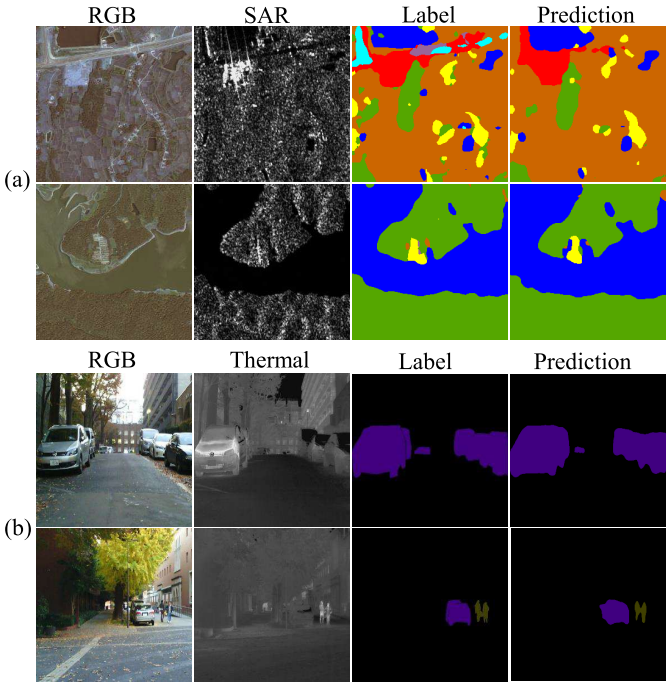


Fig. 7. Visualization results of DA^2-Net on different downstream tasks. (a) RGB-SAR land use classification. (b) RGB-thermal semantic segmentation.

is trained on the LEVIR-CD dataset and tested directly on the WHU-CD dataset. The LEVIR-CD dataset primarily focuses on urban building changes, whereas the WHU-CD dataset covers both urban and rural environments. This setup simulates a practical RSCD scenario in which no labeled data is available in the target domain.

The experimental results are summarized in Table VI. Overall, early convolutional networks exhibited limited generalization performance, likely due to their relatively small number of parameters. With continuous architectural improvements, the three transformer-based methods demonstrated significantly enhanced generalization capabilities. Driven by powerful data engines, the remaining methods based on VFMs also achieved strong performance, with all $F1$ exceeding 70%. The proposed DA^2-Net outperforms all other methods across multiple evaluation metrics, achieving improvements of 9.44% in precision and 5.65% in IoU compared to BAN. These results demonstrate the superior zero-shot generalization capability of DA^2-Net and underscore its strong potential for practical applications.

G. Performance Evaluation in Cross-Task

This section evaluates the generalization performance of DA^2-Net on various multimodal visual tasks.

1) RGB-SAR Land Use Classification: Integrating RGB and SAR data for fine-grained land category extraction holds practical significance. To demonstrate the generalizability of the proposed method in the RGB-SAR land use classification task, the two inputs of DA^2-Net were replaced, respectively, with RGB and SAR images. The DA^2-Net was then retrained on the WHU-OPT-SAR dataset [71], following the same data partitioning protocol as in [71]. As shown in Fig. 7(a), DA^2-Net produces accurate predictions, effectively distinguishing various classes and generating complete boundaries.

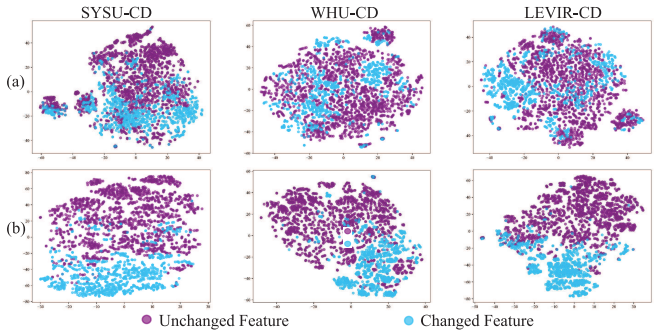


Fig. 8. Internal features of SAM2 are visualized using t-SNE, where blue and purple points represent changed and unchanged features, respectively. (a) SAM2 image encoder. (b) SAM2 image encoder with hierarchical LoRA.

2) RGB-Thermal Semantic Segmentation: Urban road segmentation is fundamental to autonomous driving technologies, and the fusion of RGB and thermal infrared data facilitates robust semantic understanding under challenging conditions such as low illumination and occlusion. In this study, further experiments were conducted on the MFNet dataset [72], with the training, validation, and test sets split at a ratio of 2:1:1. As shown in the visualization results in Fig. 7(b), DA^2-Net demonstrates a certain degree of cross-task generalization capability.

V. Discussion

A. Effectiveness of the Hierarchical LoRA and the DAEM

The hierarchical LoRA strategy introduces LoRA into the self-attention and MLP layers of the transformer blocks to achieve domain adaptation between VFMs and RS images. To further demonstrate the effectiveness of the hierarchical LoRA strategy, this study compares the distribution of deep features in SAM2 before and after its introduction. Specifically, post-temporal images rich in land-cover categories are fed into the SAM2 image encoder to extract high-level semantic features. These features are then spatially divided into changed and unchanged regions, followed by dimensionality reduction and visualization. As shown in Fig. 8(a), when SAM2 is used directly as a feature extractor, the changed and unchanged features exhibit substantial overlap in the feature space. The inter-class boundaries are indistinct, indicating insufficient discriminative information, which limits the network's ability to capture changes. In contrast, as illustrated in Fig. 8(b), the introduction of hierarchical LoRA enhances inter-class separability, leading to more distinct intra-class feature distributions. The above results indicate that the hierarchical LoRA effectively mitigates domain shift, thereby enhancing change features and suppressing unchanged features.

The proposed DAEM is designed to coordinate category-discriminative and spatial detail information to mitigate the boundary shift problem in RSCD. To further demonstrate the effectiveness of DAEM, several feature visualizations

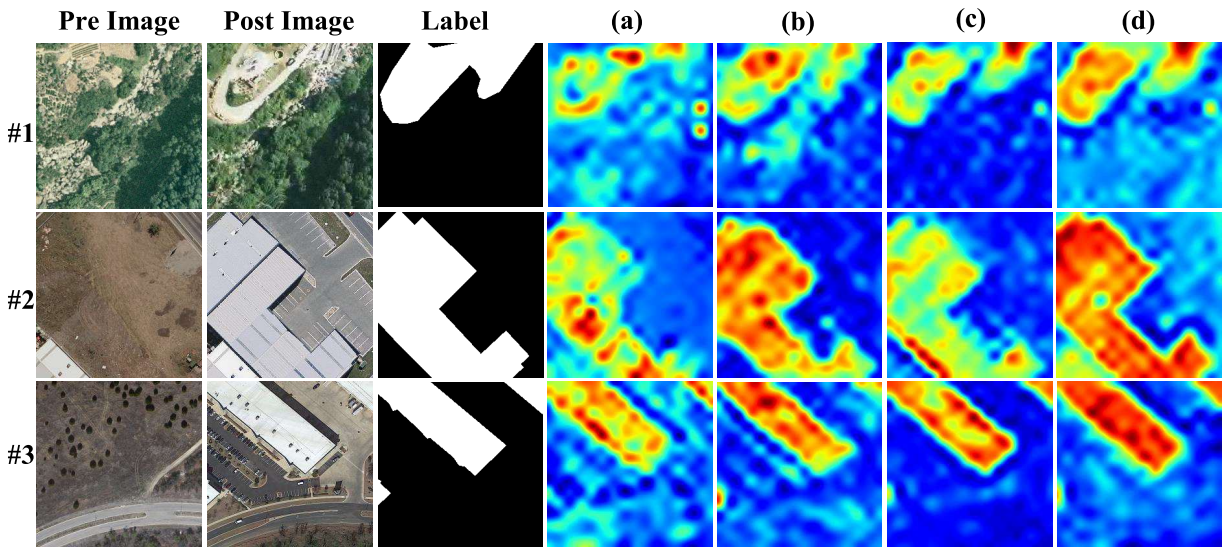


Fig. 9. Visualization examples of features in DAEM. (a) Feature map obtained by element-wise subtraction of bi-temporal features. (b) Feature map from the global difference awareness branch in DAEM. (c) Feature map from the local detail optimization branch in DAEM. (d) Feature maps obtained from the final output of DAEM. The samples are from SYSU-CD, WHU-CD, and LEVIR-CD, respectively.

are presented in Fig. 9. Specifically, Fig. 9(a) shows the difference features obtained through element-wise subtraction. This simple fusion strategy fails to provide meaningful change information. Fig. 9(b) illustrates the output from the global difference awareness branch of DAEM, where the network captures the approximate change regions but lacks accurate localization of object boundaries. In Fig. 9(c), although the interior activation of the targets is relatively weak, the edges appear smoother and more structurally enriched. As shown in Fig. 9(d), DAEM not only preserves the internal consistency of the targets but also enhances the delineation of their boundaries. These visualizations collectively demonstrate that DAEM effectively leverages the complementary strengths of different branches, thereby strengthening the network's ability to detect substantial changes.

B. Sensitivity of Hyperparameters

The hierarchical LoRA involves three key hyperparameters: the rank value R, the scaling factor α, and the dropout rate β. The R controls the number of trainable parameters, and experimental results presented in Section IV-D3 indicate that the model achieves optimal performance when $R = 16$. The parameter α adjusts the magnitude of the output from the low-rank adapters (the output scaled to $(\alpha/R)\mathbf{W}^{\mathrm{B}}\mathbf{W}^{\mathrm{A}}$), while the β serves to mitigate overfitting. To investigate the sensitivity of DA^2-Net to variations in α and β, this study evaluated its performance under multiple configurations.

1) Scaling Factor α: The impact of the α on network performance is shown in Fig. 10(a). Excessively large or small values of α lead to a notable decline in performance. The network achieves the best results on the SYSU-CD and LEVIR-CD datasets when $\alpha = 16$, while optimal performance on the WHU-CD dataset is obtained when $\alpha = 32$. This discrepancy may stem from variations in scene complexity and change patterns among different datasets.

2) Dropout Rate β: The impact of the β on network performance is shown in Fig. 10(b). It can be observed that DA^2-Net exhibits no significant performance fluctuations across different values of β. When $\beta = 0.2$, the network achieves the best overall performance on the SYSU-CD dataset. For the WHU-CD dataset, the best performance is obtained at $\beta = 0.1$, while for the LEVIR-CD dataset, the optimal results are achieved when $\beta = 0$. These results indicate that the sensitivity to β varies across datasets, suggesting that β should be adaptively tuned to achieve optimal RSCD performance.

C. Model Efficiency

Table VII presents the computational cost of DA^2-Net and other methods. It can be observed that FC-EF has the fewest parameters, lowest FLOPs, and shortest inference time, but its $F1$ is only 58.81%. USSFC achieves a good balance across these three metrics. LSAT has the most learnable parameters, but its accuracy is not the highest. Although SAM-CD has the longest inference time, its learnable parameters and the number of FLOPs operations are considerable. BAN exhibits the highest model complexity, with FLOPs reaching 353.72. Overall, VFMs-based methods generally achieve higher $F1$, which can be attributed to the strong generalization ability of VFMs. The intricate nature of VFMs leads to higher FLOPs and inference time for their adapter networks compared to traditional CNN or transformer networks. Among the comparison methods, the proposed DA^2-Net achieves an optimal balance between learnable parameters and performance. With only 6.85 million trainable parameters, DA^2-Net achieves an impressive $F1$ of 94.48%. It indicates that DA^2-Net offers a reasonable computational cost and excellent detection performance.

To further improve the practical deployment potential of the model, this study proposes DA^2-Net-tiny (denoted by DA^2-Net-t in Table VII), a lightweight version of DA^2-Net. DA^2-Net-t adopts SAM2-tiny as its backbone, and the embedding dimension of the decoder is reduced to 1/4 that of DA^2-Net. As a result, it has the fewest trainable parameters among all methods (0.82 M) and a computational cost of

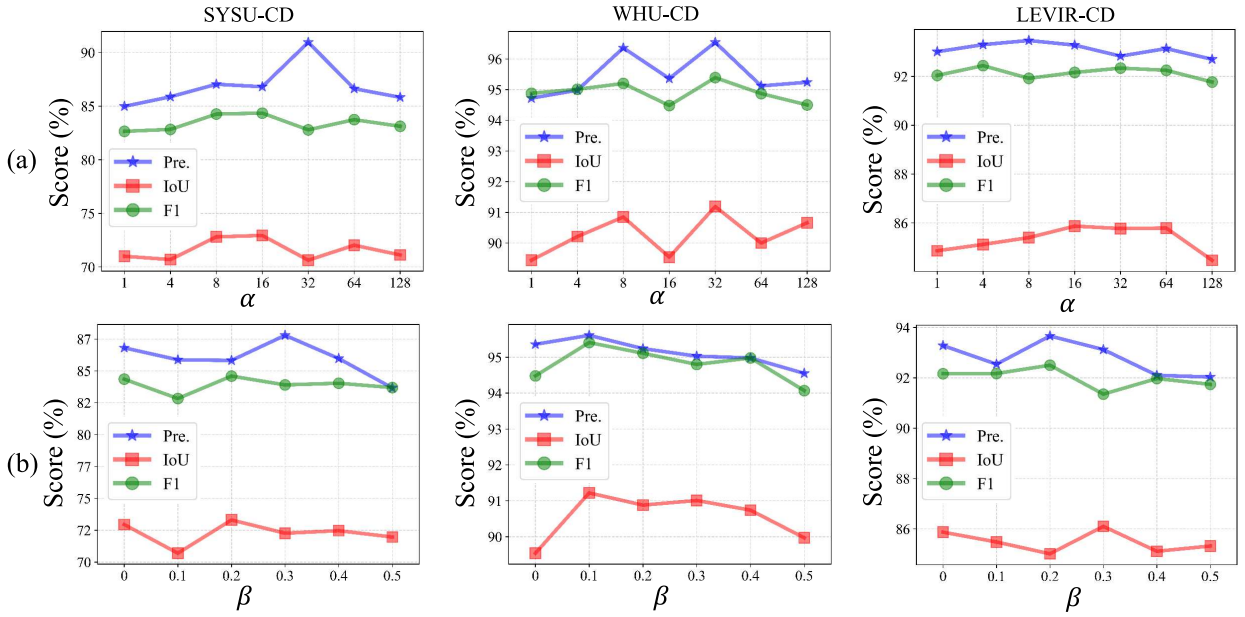


Fig. 10. Sensitivity analysis of hyperparameters. (a) Scale factor α. (b) Dropout rate β.

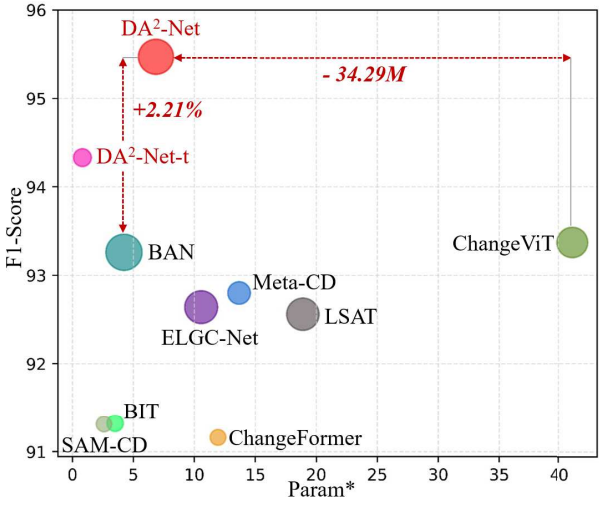


Fig. 11. Comparison of performance and trainable parameter efficiency across different methods.

only 27.85 G FLOPs. DA^2-Net-t not only significantly reduces computational cost but also delivers satisfactory performance, achieving an $F1$ on the WHU-CD dataset that is second only to DA^2-Net. Fig. 11 illustrates the superiority of the proposed method. Compared with ChangeViT, the DA^2-Net reduces the number of parameters by 34.29 M while achieving a 2.1% performance gain. Moreover, the model efficiency of DA^2-Net-t also surpasses that of prevailing RSCD methods.

VI. Conclusion

This study presents DA^2-Net for RSCD, which integrates SAM2 domain adaptation with the difference aggregation. By introducing hierarchical LoRA into SAM2, the proposed DA^2-Net can bridge the knowledge gap between natural and RS images, thereby achieving effective domain adaptation. In addition, the DAEM can adaptively aggregate category-discriminative and spatial-detail information through generating attention weights, thereby alleviating the boundary displacement problem and further enhancing the precision of RSCD. Through extensive comparative experiments and ablation studies, it is demonstrated that DA^2-Net significantly enhances the applicability of SAM2 in complex RS scenes. However, the proposed method still exhibits relatively high FLOPs and overall parameters. In future work, techniques such as knowledge distillation will be explored to develop more lightweight models for deployment on edge devices.

References

[1] H. Sun, Y. Yao, L. Zhang, and D. Ren, "Spatial focused bitemporal interactive network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5639115.

[2] T. Lei et al., "Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4507013.

[3] X. Pan, J. Lai, Y. Jin, X. Zhou, and J. Zheng, "STENet: A spatial selection and temporal evolution network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4410915.

[4] X. Zheng, H. Cui, and X. Lu, "Multiple source domain adaptation for multiple object tracking in satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626911.

[5] T. Yan, Z. Wan, P. Zhang, G. Cheng, and H. Lu, "TransY-Net: Learning fully transformer networks for change detection of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4410012.

[6] S. Yuan, R. Zhong, C. Yang, Q. Li, and Y. Dong, "Dynamically updated semi-supervised change detection network combining cross-supervision and screening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[7] B. Yang, Y. Mao, L. Liu, X. Liu, Y. Ma, and J. Li, "From trained to untrained: A novel change detection framework using randomly initialized models with spatial–channel augmentation for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402214.

[8] 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.

[9] Y. Yuan, H. Guo, and J. Gao, "Distance-aware network for physical-world object distribution estimation and counting," *Pattern Recognit.*, vol. 157, Jan. 2025, Art. no. 110896.

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

[11] Y. Chen, Z. Ye, H. Sun, T. Gong, S. Xiong, and X. Lu, "Global–local fusion with semantic information-guidance for accurate small object detection in UAV aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4701115.

[12] X. Song, Z. Hua, and J. Li, "Context spatial awareness remote sensing image change detection network based on graph and convolution interaction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 3000316.

[13] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[14] Y. Chen, Y. Wang, S. Xiong, X. Lu, X. X. Zhu, and L. Mou, "Integrating detailed features and global contexts for semantic segmentation in ultrahigh-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4703914.

[15] J. Gao, L. Zhao, and X. Li, "NWPU-MOC: A benchmark for fine-grained multicategory object counting in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5606614.

[16] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[17] 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.

[18] X. Wang, S. Li, X. Zhao, and K. Zhao, "BiG-FSLF: A cross heterogeneous domain few-shot learning framework based on bidirectional generation for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516213.

[19] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[20] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, Sep. 2021.

[21] K. Li, X. Cao, and D. Meng, "A new learning paradigm for foundation model-based remote-sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610112.

[22] D. Zhang, F. Wang, L. Ning, Z. Zhao, J. Gao, and X. Li, "Integrating SAM with feature interaction for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4513011.

[23] S. Dong, L. Wang, B. Du, and X. Meng, "ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 208, pp. 53–69, Feb. 2024.

[24] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[25] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[26] N. Ravi et al., "SAM 2: Segment anything in images and videos," 2024, *arXiv:2408.00714*.

[27] D. Zhu, X. Huang, H. Huang, Z. Shao, and Q. Cheng, "ChangeViT: Unleashing plain vision transformers for change detection," 2024, *arXiv:2406.12847*.

[28] E. J. Hu et al., "LoRa: Low-rank adaptation of large language models," in *Proc. ICLR*, vol. 1, no. 2, 2022, p. 3.

[29] T. Liu et al., "AEKAN: Exploring superpixel-based autoencoder Kolmogorov-Arnold network for unsupervised multimodal change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5601114.

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

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

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

[33] 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, 2022, Art. no. 5604816.

[34] S. Zuo, Y. Xiao, X. Chang, and X. Wang, "Vision transformers for dense prediction: A survey," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109552.

[35] 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.

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

[37] W. Li, L. Xue, X. Wang, and G. Li, "ConvTransNet: A CNN–transformer network for change detection with multiscale global–local representations," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610315.

[38] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 599–609, Aug. 2023.

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

[40] X. Dong et al., "MaskCLIP: Masked self-distillation advances contrastive language-image pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10995–11005.

[41] M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, and W. Zhang, "ClearCLIP: Decomposing clip representations for dense vision-language inference," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 15105. Cham, Switzerland: Springer, 2025, pp. 143–160.

[42] Y. Xiong et al., "EfficientSAM: Leveraged masked image pretraining for efficient segment anything," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16111–16121.

[43] D. Wang et al., "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 8815–8827.

[44] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5636916.

[45] L. Zheng, X. Pu, and F. Xu, "Tuning a SAM-based model with multi-cognitive visual adapter to remote sensing instance segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 2737–2748, 2024.

[46] Y. Yuan, Y. Zhan, and Z. Xiong, "Parameter-efficient transfer learning for remote sensing image–text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5619014.

[47] Z. Shan, Y. Liu, L. Zhou, C. Yan, H. Wang, and X. Xie, "ROS-SAM: High-quality interactive segmentation for remote sensing moving object," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 3625–3635.

[48] T. Zhang, Y. Ren, W. Li, C. Qin, L. Jiao, and H. Su, "CSW-SAM: A cross-scale algorithm for very-high-resolution water body segmentation based on segment anything model 2," *ISPRS J. Photogramm. Remote Sens.*, vol. 228, pp. 208–227, Oct. 2025.

[49] Z. Qi, H. Chen, H. Zhang, Z. Zou, and Z. Shi, "Efficient semantic splatting for remote sensing multiview segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5621415.

[50] H. Ning, B. Zhao, and Y. Yuan, "Semantics-consistent representation learning for remote sensing image–voice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700614.

[51] J. Gao, H. Yang, D. Zhang, Y. Yuan, and X. Li, "Imbalanced aircraft data anomaly detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 61, no. 2, pp. 1422–1432, Apr. 2025.

[52] X. Xiong et al., "SAM2-UNet: Segment anything 2 makes strong encoder for natural and medical image segmentation," 2024, *arXiv:2408.08870*.

[53] 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.

[54] K. Chen et al., "Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2024, pp. 8581–8584.

[55] M. Wang, L. Zhou, K. Zhang, X. Li, M. Hao, and Y. Ye, "ESAM-CD: Fine-tuned efficientsam network with LoRa for weakly supervised remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4708616.

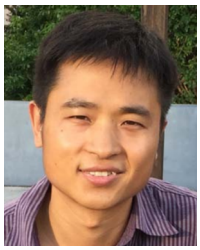
[56] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "ELGC-Net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701611.
[57] H. Ning, B. Zhao, Z. Hu, L. He, and E. Pei, "Audio–visual collaborative representation learning for dynamic saliency prediction," *Knowl.-Based Syst.*, vol. 256, Sep. 2022, Art. no. 109675.
[58] 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, Jan. 2019.
[59] 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, p. 1662, May 2020.
[60] 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.
[61] T. Lei et al., "Lightweight structure-aware transformer network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[62] J. Gao, D. Zhang, F. Wang, L. Ning, Z. Zhao, and X. Li, "Combining SAM with limited data for change detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5614311.
[63] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.
[64] S.-Y. Liu et al., "DoRA: Weight-decomposed low-rank adaptation," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 1–22.
[65] D. Yin, L. Hu, B. Li, Y. Zhang, and X. Yang, "5%>100%: Breaking performance shackles of full fine-tuning on visual recognition tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2025, pp. 20071–20081.
[66] B. Yu et al., "Visual tuning," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–38, 2024.
[67] X. Zheng, H. Cui, C. Xu, and X. Lu, "Dual teacher: A semisupervised cotraining framework for cross-domain ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613312.
[68] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2021.
[69] W. Gedara Chaminda Bandara and V. M. Patel, "Revisiting consistency regularization for semi-supervised change detection in remote sensing images," 2022, *arXiv:2204.08454*.
[70] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7236–7246.
[71] X. Li et al., "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102638.
[72] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115.

Hailong Ning (Member, IEEE) received the bachelor's degree in biomedical engineering from Xi'an Jiaotong University (XJTU), Xi'an, China, in 2016, and the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2021.

He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an. His research interests include pattern recognition, machine learning, computer vision, and multimodal learning.

Qi He received the bachelor's degree from Xi'an University of Posts and Telecommunications, Xi'an, China, in 2023, where he is currently pursuing the M.S. degree with the School of Computer Science and Technology.

His research interests include image processing and pattern recognition.



Tao Lei (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2011.

From 2012 to 2014, he was a Post-Doctoral Research Fellow with the School of Electronics and Information, Northwestern Polytechnical University. From 2015 to 2016, he was a Visiting Scholar with the Quantum Computation and Intelligent Systems Group, University of Technology Sydney, Ultimo, NSW, Australia. He is currently a Professor with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an. He has authored or co-authored more than 80 research articles, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the International Conference on Acoustics, Speech, and Signal Processing, the IEEE International Conference on Image Processing, and the IEEE International Conference on Automatic Face and Gesture Recognition. His research interests include image processing, pattern recognition, and machine learning.

Xiaopeng Cao received the Ph.D. degree in computer science and technology from Xidian University, Xi'an, China, in 2016.

He is currently a Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an. His research interests include computer graphics and machine learning.

Wuxia Zhang received the bachelor's degree in information display and opto-electronic technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the master's and Ph.D. degrees in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2012 and 2019, respectively.

From 2012 to 2016, she was a Software Engineer with Xi'an Huawei Technologies Company Ltd., Xi'an, China. She is currently an Associate Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an. Her research interests include remote sensing and machine learning, especially remote sensing detection and deep networks, with their applications in remote sensing.

Yanping Chen received the Ph.D. degree in computer architecture from Xi'an Jiaotong University, Xi'an, China, in 2007.

She is currently a Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an. Her research interests include service mining, industrial intelligence, and network management.

Asoke K. Nandi (Life Fellow, IEEE) received the Ph.D. degree in physics from the University of Cambridge (Trinity College), Cambridge, U.K., in 1979.

He held academic positions in several universities, including the University of Oxford, Oxford, U.K.; Imperial College London, London, U.K.; the University of Strathclyde, Glasgow, U.K.; the University of Liverpool, Liverpool, U.K.; and a Finland Distinguished Professorship. In 2013, he moved to Brunel University of London, Uxbridge, U.K. In 1983, he co-discovered the three fundamental particles known as W^+, W^-, and Z^0, providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 awarded the prize to two of his team leaders for their decisive contributions. He made fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big Data." He has authored over 650 technical publications, including 310 journal articles and six books, entitled *Image Segmentation: Principles, Techniques, and Applications* (Wiley, 2022), *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (Wiley, 2020), *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Blind Estimation Using Higher-Order Statistics* (Springer, 1999), and *Automatic Modulation Recognition of Communication Signals* (Springer, 1996). The H-index of his publications is 91 (Google Scholar), and ERDOS number is 2. His research interests lie in signal processing and machine learning, with applications to machine health monitoring, functional magnetic resonance data, gene expression data, communications, and biomedical data.

Prof. Nandi is a fellow of the Royal Academy of Engineering and six other institutions, including the IEEE. In 2023, he was honored by the Academia Europaea and the Academia Scientiarum et Artium Europaea. He has received many awards, including the IEEE Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers in 1999, and the Mountbatten Premium of the Institution of Electrical Engineers in 1998. He was an IEEE Distinguished Lecturer (EMBS, 2018 and 2019).