Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DINO-CD:Marrying DINOv3 with Dual-Branch Complementary Reconstruction for Remote Sensing Change Detection

Xiaopeng Cao

cxp2760@163.com

Xi'an University of Posts and Telecommunications

Bingwen Dou

Xi'an University of Posts and Telecommunications

Hailong Ning

Xi'an University of Posts and Telecommunications

Qi He

Xi'an University of Posts and Telecommunications

Yizhuo Dong

Xi'an University of Posts and Telecommunications

Sisi Wang

Xi'an University of Posts and Telecommunications

Anlei Wang

Xi'an University of Posts and Telecommunications

Research Article

Keywords: Remote Sensing Change Detection, Vision Foundation Models, Frequency Learning, Dual-Branch

Posted Date: April 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9195501/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DINO-CD:Marrying DINOv3 with Dual-Branch Complementary Reconstruction for Remote Sensing Change Detection

Xiaopeng Cao[1*], Bingwen Dou[1], Hailong Ning[1], Qi He[1], Yizhuo Dong[1], Sisi Wang[1], Anlei Wang[1]

[1*]School of Computer Science, Xi'an University of Posts and Telecommunications, Chang'an West Street, Xi'an, 710121, Shaanxi, China.

*Corresponding author(s). E-mail(s): cxp2760@163.com;
Contributing authors: a791515391@163.com;

Abstract

The goal of remote sensing change detection (RSCD) is to accurately distinguish land-surface changes. Recently, Vision foundation models (VFMs), such as the self-supervised DINOv3, have demonstrated strong semantic parsing capabilities across various visual tasks. However, directly applying DINOv3 to RSCD remains challenging. First, DINOv3 produces low-resolution, single-scale feature representations, which are unfavorable for pixel-wise change classification. Second, DINOv3 is not tailored for RSCD and thus struggles to model geospatial differences along the temporal dimension. To address these issues, Marrying DINOv3 with Dual-Branch Complementary Reconstruction for Remote Sensing Change Detection (DINO-CD) is proposed. Specifically, a dual-branch complementary reconstruction module is introduced to enable interaction between DINO and CNN branches for complementary feature enhancement, producing multi-scale representations that preserve global semantic consistency while strengthening local detail discrimination. In addition, a hierarchical difference aggregation module is designed to comprehensively enhance temporal discrepancies via explicit frequency learning and multi-scale directional aggregation, thereby highlighting effective change regions and reducing false alarms and missed detections. Extensive experiments on multiple datasets demonstrate that DINO-CD outperforms state-of-the-art methods, achieving gains of 6% in F1-score and 8% in IoU on the CLCD dataset.

Keywords: Remote Sensing Change Detection,Vision Foundation Models,Frequency Learning,Dual-Branch.

1 Introduction

With the continuous evolution of Earth observation technologies, satellite platforms, and sensor engineering, high-resolution optical remote sensing imagery has achieved significant improvements in both spatial and temporal resolutions, providing a solid data foundation for fine-grained characterization of land-cover evolution and human activity processes. Remote sensing change detection (RSCD) [1] aims to automatically identify change regions and change types in land objects (i.e., changes in category or state) based on remote sensing images of the same area acquired at different times, and it constitutes a key technical component in a wide range of applications, including land resource management [2], urban expansion monitoring [3], land-cover change assessment [4], ecological and environmental regulation [5], and disaster loss assessment [6]. Affected by multiple sources of uncertainty, such as illumination variations, seasonal transitions, vegetation phenology differences, and imaging environment fluctuations, RSCD tasks are often accompanied by pronounced radiometric and textural disturbances. These non-semantic interferences are easily misjudged as real land-object changes, thereby triggering large-scale pseudo-changes and posing severe challenges to accurate change identification. Traditional hand-crafted methods struggle to cope with such complex variations, whereas deep learning (DL) methods, benefiting from end-to-end learning and their ability to extract more robust and generalizable high-level semantic features, have become the mainstream approach in contemporary RSCD.

CNN-based methods mainly improve network architectures through receptive-field expansion, multi-scale representation learning, and attention enhancement, thereby strengthening the semantic discrimination and spatial localization of changed regions in RSCD [7–10]. Owing to their stable performance in local texture representation and spatial detail characterization, CNN-based approaches have constituted the mainstream technical route in RSCD research. In parallel, Vision Transformers (ViTs) [11], benefiting from self-attention mechanisms, exhibit clear advantages in capturing long-range dependencies and global contextual information, and have gradually evolved into core backbone networks that run alongside traditional CNNs, and even dominate in certain tasks. Extensive studies indicate that the Transformer family has progressively demonstrated stronger capabilities in image classification [12], object detection [13], semantic segmentation [14], image matting [15], and human pose estimation [16], making it a highly competitive alternative to conventional CNN architectures. This technological trend has also been extending to remote sensing vision. Existing research has verified the effectiveness of Transformer architectures in modeling complex spatial relationships and long-range semantic dependencies across various remote sensing tasks [17], [18], and has further attempted to introduce them into RSCD to explicitly characterize global semantic discrepancies between bi-temporal remote sensing images. However, given the importance of local details and boundary information in remote sensing imagery, some studies have further explored collaborative CNN-Transformer

fusion strategies. By integrating local perception capability with global contextual modeling, these variants aim to build RSCD models with more robust performance and stronger generalization ability.

Although DL methods have achieved significant performance gains in recent years, they still face several key issues that remain insufficiently resolved in complex remote sensing scenarios. First, existing methods are commonly constrained by the trade-off between global semantic modeling and local detail characterization. CNN-based approaches perform reliably in capturing local texture patterns and spatial details; however, limited by their receptive fields, they remain insufficient in modeling long-range spatial dependencies and maintaining global semantic consistency. In contrast, Transformer-based methods can effectively model long-range dependencies and strengthen global contextual understanding via self-attention, but their computational complexity grows quadratically with spatial resolution, often resulting in substantial computation and GPU memory overhead for high-resolution remote sensing imagery. In this context, vision foundation models (VFMs) with large-scale pretraining capability learn robust global semantic representations from massive data, providing new possibilities for simultaneously modeling long-range dependencies and scene-level semantic consistency. Such a paradigm is expected to, to some extent, alleviate the limitations of existing methods in the global-local modeling trade-off and robustness under complex remote sensing conditions.

With the rapid advancement of large-scale pretrained Vision Transformers [19–24], self-supervised DINOv3 in particular has demonstrated strong representational power and impressive cross-dataset transferability on generic vision tasks [21]. Compared with VFMs that have already been introduced into RSCD, such as SAM and CLIP, DINOv3 places greater emphasis, in both model design and pretraining objectives, on feature quality for dense prediction. It produces patch-level dense representations with good spatial alignment, rather than being primarily tailored to specific downstream goals such as prompt-driven segmentation or cross-modal semantic matching. From the perspective of feature form and modeling objective, the robust global semantic representations and highly consistent dense embeddings provided by DINOv3 better match the fundamental requirements of RSCD, namely spatial continuity and cross-temporal feature comparability. This makes DINOv3 a more promising candidate as a feature encoder for RSCD.

However, directly adopting DINOv3 as a plug-and-play Transformer backbone for RSCD still faces the following challenges: (1).Limitations of the original DINOv3 features. DINOv3 is biased toward global semantic representations, and its features are typically output as single-scale patch embeddings, which is unfavorable for pixel-wise change classification. Naively obtaining multi-scale features by interpolation tends to introduce over-smoothing, resulting in blurred boundaries and missed detections of fine structures, such as narrow roads, elongated edges, and scattered small patches. (2).Lack of explicit spatiotemporal difference modeling. Although DINOv3 provides strong global semantic representations, its outputs are typically biased toward features of a single time point and do not explicitly model spatiotemporal discrepancies between bi-temporal images. While one can merge bi-temporal information using simple feature fusion, such a strategy cannot fully exploit cross-temporal differences, especially when

the time interval is large. Consequently, the model may struggle to accurately capture fine-grained changed regions, particularly subtle changes under complex backgrounds, which often leads to missed detections and false predictions for small targets.

To address the above challenges, This paper propose DINO-CD: Marrying DINOv3 with Dual-Branch Complementary Reconstruction for Remote Sensing Change Detection. On the one hand, a dual-branch encoding architecture is constructed with a pretrained DINOv3 branch and a CNN branch, and a Dual-Branch Complementary Reconstruction (DBCR) module is introduced on intermediate features to fully exploit the structural complementarity between the two representations, thereby mitigating the limitations of DINOv3 in RSCD. Specifically, the bi-temporal remote sensing images are fed into the DINOv3 and CNN branches, respectively, to extract four hierarchical feature levels. DBCR then performs spatial complementary reconstruction and channel complementary reconstruction at each corresponding level, deeply integrating the global semantic priors from DINOv3 with the high-resolution boundaries and texture details from the CNN, and producing complementary features that are both globally consistent and locally fine-grained. On the other hand, a Hierarchical Difference Aggregation (HDA) module is developed to compensate for DINOv3's deficiency in temporal discrepancy modeling. Given bi-temporal multi-scale features, HDA first introduces frequency-domain pre-enhancement before differencing to construct difference features. It then adopts hierarchical enhancement and grouped temporal interactions, combining local context modeling with directional receptive fields to aggregate long-range change cues and fine-structured information, and finally uses the difference features to modulate the temporally interacted features in a pixel-wise manner.

The main contributions of this work are summarized as follows:

1. To alleviate the limitations of the original DINOv3 feature representations, this paper proposes DINO-CD, a unified network that integrates DINOv3 and CNNs. Unlike existing VFM adaptation strategies, DINO-CD leverages DBCR to establish a hierarchical, bidirectional complementary relationship between the global semantic representations of DINOv3 and the local structural details captured by the CNN branch.
2. To strengthen spatiotemporal discrepancy modeling, a HDA module is introduced. Unlike existing difference generation modules, HDA constructs difference features via frequency-domain enhancement and then uniformly aggregates directional discrepancy cues within a hierarchical architecture, enabling progressive enhancement across stages.
3. Extensive experiments demonstrate that DINO-CD achieves competitive performance on multiple public RSCD benchmarks and exhibits stronger robustness, especially in challenging scenarios with complex backgrounds and low-contrast changes.

2 Related Work

In recent years, RSCD has achieved remarkable progress driven by deep learning, and representative methods can be broadly categorized into two lines, namely CNN-based approaches and Transformer-based approaches.

2.1 Deep Learning Based Change Detection Methods

CNN-based methods:Such methods typically adopt FCN, U-Net, and their variants as backbones, and build shared-weight or non-shared Siamese encoders to extract features from bi-temporal images separately. Then, change cues are derived via feature concatenation, differencing, or similarity measurement, and a decoder outputs the final change mask. FC-Conc and FC-Diff proposed by Daudt et al. [25] were among the first to validate, within a unified fully convolutional Siamese framework, the effectiveness of the paradigm of Siamese encoding coupled with difference representation, laying a foundational CNN-based formulation for change detection. Peng et al. [26] independently encoded bi-temporal features and fused them in the feature space to mine change cues, further promoting the adoption and extension of Siamese structures in RSCD. Building upon this line, subsequent studies have continuously evolved around three aspects, namely the expressiveness of difference cues, the preservation of fine structures, and robust modeling of contextual relationships. On the one hand, to mitigate the loss of localization information caused by downsampling and to improve boundary delineation in high-resolution scenes, researchers have widely introduced stronger cross-level interactions and multi-scale fusion mechanisms. For example, Fang et al. employed a nested U-Net with dense skip connections in SNUNet-CD [27], enabling shallow details and deep semantics to be sufficiently propagated and fused, thereby enhancing the representation of small changes and boundary regions. On the other hand, regarding difference representation as the core component, an increasing number of methods attempt to explicitly inject temporal interactions and discrepancy cues in a more direct or lightweight manner. Fang et al. [28] further proposed exchanging bi-temporal features inside the network to achieve an almost parameter-free difference representation, providing a simple yet effective strategy for lightweight discrepancy modeling. DAFNet proposed by Ma et al. [29] re-weights changed regions via difference feature extraction and attention modulation, while suppressing pseudo-changes caused by illumination, shadows, and seasonal or phenological variations.In addition, to alleviate the issue that relying solely on local differencing can be easily distracted by complex background noise, some works further introduce global context representations and scale-consistency constraints to improve scene-level discrimination. GAS-Net proposed by Zhang et al. [30] enhances long-range dependency modeling via global-aware feature learning, reducing the tendency of false alarms under large-scale and cluttered backgrounds. MFDS-Net by Huang et al. [31] adopts multi-scale deep supervision together with feature enhancement mechanisms, imposing consistent constraints on changed regions across scales, which improves the response consistency to fine-grained changes and boundaries.

Overall, CNN-based methods are relatively mature in local structure modeling and multi-scale feature fusion, and can effectively capture change cues in many

scenarios. However, their receptive-field expansion and global semantic modeling remain inherently limited, and they may still suffer from false-change diffusion, unstable boundaries, and missed detections of remotely correlated regions when facing large-area imagery, cross-region correlated changes, or complex background disturbances.

Transformer-based methods:With the widespread success of self-attention in computer vision, Vision Transformers have gradually been introduced into RSCD to compensate for the limitations of CNNs in modeling global relationships and long-range dependencies. BiT proposed by Chen et al. [32] follows a hybrid paradigm of CNN feature extraction, token compression, and Transformer encoding. Specifically, it first employs ResNet [33] to encode bi-temporal features, compresses high-level change information into a small set of semantic tokens, and then uses a Transformer encoder to capture long-range contextual dependencies in the token space. Finally, these globally informed tokens are mapped back to pixel-level features and refined by convolutional representations, thereby enhancing global change modeling while preserving local details. Building upon this line, end-to-end Transformer frameworks have been further explored, including Siamese designs that explicitly model bi-temporal interactions and single-stream attempts that unify bi-temporal representations. ChangeFormer [18] introduces multi-scale Transformer representations and a dedicated decoding design within a Siamese framework to achieve end-to-end change prediction, demonstrating the effectiveness of Transformers for RSCD in terms of global context modeling and multi-scale semantic representation.To address the computational bottleneck induced by high-resolution, large-area remote sensing imagery, researchers have paid increasing attention to efficient attention designs, typically adopting locally constrained attention, such as window-based attention, or sparse attention to reduce computation and GPU memory cost. Siam-Swin-Unet [34] combines the hierarchical window attention of Swin Transformer with a U-shaped architecture to balance local details and global semantics and to improve boundary quality. STCD [35] introduces window-based hierarchical attention into a Siamese change detection framework to mitigate the heavy cost of global self-attention, achieving a better efficiency-accuracy trade-off. Noman et al. [36] further propose shuffled sparse attention, which reduces the cost of global modeling by focusing on sparse informative regions.Meanwhile, some works emphasize stronger temporal interactions and relational reasoning to improve the discriminability of difference representations. For example, RCDT [37] adopts cross-attention as the core mechanism to perform temporal interaction and change relation reasoning, and DEST [38] explicitly enhances discrepancy information within the Swin framework to improve change discrimination under complex backgrounds. To better characterize multi-scale changes, MCAT [39] proposes a single-stream multi-scale framework, introducing dynamic change-aware attention and cross-scale aggregation to enhance adaptability to changes of different sizes. Despite the clear advantages of Transformer-based methods in global context representation, they still face a common challenge in RSCD, namely that Transformer features are biased toward high-level semantics and tend to be over-smoothed. As a result, they may be less effective in fine-grained modeling of slender structures such as roads and rivers, building edges, and scattered small patches, which can lead to blurred boundaries and missed detections of small targets.

2.2 VFMs in RSCD

VFMs typically refer to visual models that are pretrained on large-scale datasets, endowed with strong general-purpose representations, and can be readily transferred to a wide range of downstream tasks. Benefiting from massive training data and high-capacity architectures, these models often exhibit excellent generalization, which can substantially reduce the reliance of downstream tasks on large amounts of manually annotated samples. For example, the Segment Anything Model (SAM) [22] and a series of self-supervised Vision Transformer models, namely the DINO family [19–21], are pretrained on an extremely large segmentation dataset and large-scale natural image corpora, respectively. They have shown remarkable cross-task transferability in object segmentation and general feature representation, and are thus increasingly regarded as important prior carriers for remote sensing image analysis.Motivated by these advantages, an increasing number of studies have attempted to introduce VFMs into RSCD, exploring various fusion strategies centered on how to efficiently leverage their prior knowledge while keeping most parameters frozen. Some works are built upon SAM and refine its high-level semantic features with lightweight structures. For instance, Zhang et al. [40] proposed SFCD-Net, which freezes the SAM encoder and introduces parameter-efficient fine-tuning strategies together with a bi-temporal feature interaction module, enabling enhanced representation of fine-grained changes and boundary details under limited computational resources. Building on this idea, Ning and He et al. [41] further proposed DA2-Net, which combines SAM2 with domain adaptation and difference aggregation mechanisms to mitigate the domain shift from natural scenes to remote sensing imagery and to strengthen temporal discrepancy representation, thereby improving change detection performance in complex scenarios.Meanwhile, beyond SAM-based approaches, self-supervised pretrained VFMs from the DINO family have also been introduced into RSCD to more efficiently exploit their general representations. Zhao et al. [42] proposed AdapterCD, which freezes a DINOv2-pretrained image Transformer encoder and inserts only a small number of adapter layers for parameter-efficient fine-tuning, achieving a favorable balance between accuracy and efficiency on multiple change detection benchmarks. Along a similar line of leveraging large-scale pretrained priors such as DINOv2, Zhu et al. [43] further proposed ChangeViT, which adopts DINOv2 as the backbone and introduces a detail-capturing module and a feature injector to inject multi-scale details extracted by a CNN into high-level semantic features, thereby balancing large-scale change perception and fine-grained boundary localization.

Overall, existing studies have validated the value of vision foundation models in change detection from different perspectives, indicating that their strong semantic priors can improve robustness under complex backgrounds and cross-scene settings. However, new issues have also been revealed. Self-supervised ViTs represented by the DINO family typically output single-scale, high-dimensional semantic embeddings, and thus provide relatively insufficient explicit multi-scale spatial details. In contrast, segmentation-oriented models represented by SAM still require further investigation on how to simultaneously model temporal discrepancies and preserve fine-grained spatial structures when transferred directly to bi-temporal change representation. These observations collectively suggest that it remains an important and ongoing research

direction to retain the semantic-prior advantages of VFMs while compensating for their limitations in multi-scale spatial structure modeling and temporal difference representation in VFM-based change detection.

3 Methodology

3.1 Overview

As shown in Figure 1, the overall framework consists of a DINOv3 image encoder, a convolutional spatial prior module, a DBCR module, a HDA module, and a decoder. First, the bi-temporal images are separately fed into DINOv3 and the convolutional spatial prior module to extract global semantic features and local detail features, respectively. Then, DBCR converts the single-scale semantic features produced by DINOv3 into multi-scale representations and integrates them with the spatial prior features in a complementary manner. Next, HDA performs multi-scale differencing and contextual enhancement on the complemented features, producing change features with more salient structures. Finally, the decoder fuses multi-scale contextual features to generate a pixel-wise change probability map with strong discriminability.

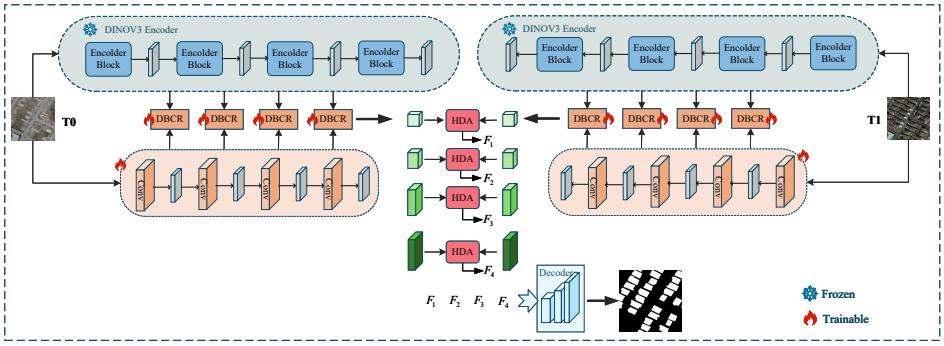


Fig. 1: The proposed DINO-CD network. The framework consists of a fully frozen DINOv3 image encoder, CSP, DBCR, HDA, and the decoder.

3.2 Dual-Branch Complementary Reconstruction Module

DINOv3 is employed to extract multi-level semantic features, while a convolutional spatial prior (CSP) layer is used to capture multi-scale local structural information; both are jointly fed into the DBCR module. The bi-temporal images $I_1 \in \mathbb{R}^{H\times W\times 3}$ and $I_2 \in \mathbb{R}^{H\times W\times 3}$ are first processed by the frozen DINOv3 encoder, from which four groups of intermediate-layer features $\{F^{(l)}\}$ are extracted.

$$F^l \in \mathbb{R}^{C\times H_i\times W_i}, \quad i = 1, \ldots, 4. \tag{1}$$

These layers correspond to different feature levels, ranging from shallow textures and edge structures to deep semantic abstractions, thereby covering the local-detail

sensitivity and global semantic stability required for change detection. In parallel, the convolutional spatial prior (CSP) module applies convolution and pooling operations to the original images, producing multi-scale features C_1, C_2, C_3, C_4 at spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input, respectively. To make the two-branch features compatible in both spatial size and channel dimension, the single-scale feature output of DINOv3 is first upsampled to the corresponding CSP scale and then projected to the same channel dimension as C_i using a 1×1 convolution followed by normalization. Accordingly, the projected semantic feature S_i is kept channel-consistent with the CSP output at each scale, yielding a matched feature pair (S_i, C_i) per scale, which represent the DINOv3 semantic branch and the convolutional prior branch, respectively. The formulation is given as follows:

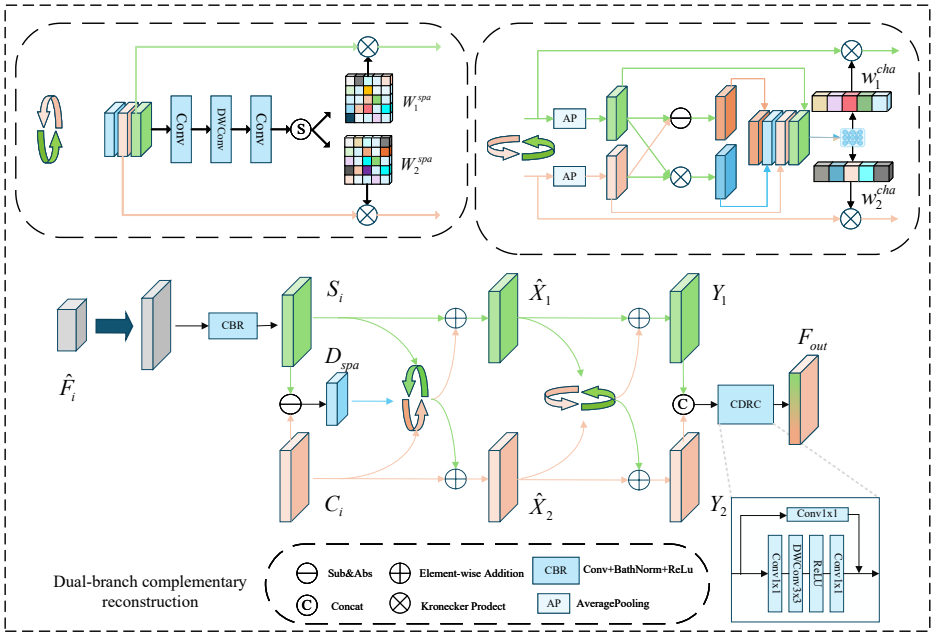


Fig. 2: The structure of the proposed DBCR model.

$$C_1, C_2, C_3, C_4 = CSP(I) \tag{2}$$

$$\hat{F}_i = \mathrm{Interp}\big(F^{(l)}, \mathrm{size}\big(H(C_i), W(C_i)\big)\big) \tag{3}$$

$$S_i = \mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left(\hat{F}_i\right)\right), \quad S_i \in \mathbb{R}^{D_i \times H_i \times W_i} \tag{4}$$

where CSP$(\cdot)$ denotes the convolution pooling operation; Interp$(\cdot)$ denotes the interpolation operation; size$(\cdot,\cdot)$ returns the height and width of the feature C_i. $\mathrm{Conv}_{1\times1}(\cdot)$ denotes a 1x1 convolution; and BN$(\cdot)$ denotes batch normalization.

Since S_i and C_i exhibit significant discrepancies in both spatial layout and channel characteristics, a DBCR module is designed to perform spatial complementary reconstruction and channel complementary reconstruction on each scale-wise pair (S_i, C_i), thereby establishing a consistent feature reference system between the two branches. For an arbitrary scale, the input features are denoted as $X_1 = S_i$ and $X_2 = C_i$. The spatial complementary reconstruction first explicitly constructs the absolute difference between the two features and concatenates it with them along the channel dimension. Then, a 1×1 convolution is applied for channel reduction, followed by a depthwise separable convolution to capture local spatial relationships. Finally, another 1×1 convolution and a Sigmoid function are used to generate spatial weight maps for both branches. The resulting spatial weight maps are split into W_1^{spa} and W_2^{spa}, corresponding to X_1 and X_2, respectively. Moreover, a set of learnable fusion coefficients $a_1, a_2 > 0$ is introduced to prevent excessive complementary interaction from corrupting the original semantic features; with a ReLU operation and a normalization constraint, their sum is enforced to be 1, so as to control the relative strength between spatial and channel complementarity. Guided by the explicit difference map, mutual spatial feature enhancement is performed for the two branches, which explicitly suppresses misaligned regions while preserving the respective structural advantages of each branch. The details are as follows:

$$D_{spa} = |X_1 - X_2| \tag{5}$$

$$U_{spa} = [X_1, X_2, D_{spa}] \tag{6}$$

$$W_{spa} = \sigma\big(Conv_{1\times1}\big(DWConv_{3\times3}\big(Conv_{1\times1}\big(U_{spa}\big)\big)\big)\big) \tag{7}$$

$$W_1^{spa}, W_2^{spa} = \mathrm{Split}(W_{spa}) \tag{8}$$

$$\hat{X}_1 = X_1 + a_1 W_2^{spa} \cdot X_2 \tag{9}$$

$$\hat{X}_2 = X_2 + a_1 W_1^{spa} \cdot X_1 \tag{10}$$

where $|\cdot|$ denotes the absolute value operator; $[\cdot]$ denotes concatenation; $\mathrm{DWConv}_{3\times3}$ denotes a depthwise separable convolution; σ denotes the sigmod operation; and $\hat{X}_1$, $\hat{X}_2$ are the spatially aligned features after spatial complementarity.

Building upon spatial complementary reconstruction, channel-wise semantic discrepancies between the two branches are further characterized. Specifically, global average pooling is applied to $\hat{X}_1$ and $\hat{X}_2$, and four types of statistical interactions are constructed to explicitly couple the channel relations between the two feature streams. The resulting concatenated descriptor is then fed into an MLP to produce channel complementary weights for both branches, which are squashed into $(0, 1)$ via a Sigmoid function. Here, a_2 is a learnable parameter that controls the strength of channel complementarity. The final channel complementary formulation is given as follows:

$$g_1 = avg(\tilde{X}_1),\ g_2 = avg(\tilde{X}_2) \tag{11}$$

$$z = (g_1, g_2, |g_1 - g_2|, g_1 \otimes g_2) \tag{12}$$

$$[w_1^{cha}, w_2^{cha}] = \mathrm{MLP}(z) \tag{13}$$

$$Y_1 = \tilde{X}_1 + a_2 w_2^{cha} \cdot \tilde{X}_2 \tag{14}$$

$$Y_2 = \tilde{X}_2 + a_2 w_1^{cha} \cdot \tilde{X}_1 \tag{15}$$

where avg$(\cdot)$ denotes average pooling, $\otimes$ denotes element-wise multiplication, MLP$(\cdot)$consists of two linear layers, Y_1 and Y_2 denote the features after channel complementary reconstruction.

Although Y_1 and Y_2 have been complemented in both spatial and channel dimensions, channel redundancy may still remain. To obtain a more compact representation suitable for subsequent modules, they are concatenated along the channel dimension and the fused feature is fed into a 1×1 convolution, a 3×3 depthwise separable convolution, a ReLU activation, and another 1×1 convolution for semantic restoration, followed by normalization to produce the final embedded feature. Meanwhile, to avoid semantic drift introduced by depthwise convolution, a residual branch is added. The formulation is given as follows:

$$F_{emb} = Norm\big(Conv_{1\times1}\big(ReLU\big(DWConv_{3\times3}\big(Conv_{1\times1}(X)\big)\big)\big)\big) \tag{16}$$

$$F_{out} = Norm\big(F_{emb} + Conv_{1\times1}(X)\big) \tag{17}$$

3.3 Hierarchical Difference Aggregation Module

In Figure 3, the bi-temporal features obtained from dual-branch complementary reconstruction are denoted as f_i^1 and f_i^2, where $i \in \{1, 2, 3, 4\}$ corresponds to the multi-scale features output by the feature matching module. To further enhance the representation of discrepancy information, pre-enhancement is performed on the bi-temporal features before constructing the initial differences. Specifically, a discrete wavelet transform (DWT) is applied to f_i^1 and f_i^2, decomposing each feature into a low-frequency component (LL) and three high-frequency components (LH/HL/HH). While preserving the semantic stability of the low-frequency component, the high-frequency components are enhanced with depthwise separable convolutions to strengthen edge information. Then, average pooling and max pooling are performed, and channel-attention weights are generated via fully connected layers followed by a Sigmoid function; these weights are used to recalibrate the subband features channel-wise and are added to the original subband features in a residual manner. Afterward, an inverse discrete wavelet transform (IDWT) is applied to reconstruct the frequency-enhanced features. Finally, pixel-wise subtraction, absolute-value computation, and a 3×3 convolution are conducted to obtain the frequency-enhanced difference feature D_i. The formulation is given as follows:

$$(f_i^{LL}, f_i^{LH}, f_i^{HL}, f_i^{HH}) = \mathrm{DWT}(f_i) \tag{18}$$

$$\widetilde{f}_L = \mathrm{ReLU}\big(BN\big(DWConv_{3\times3}(f_L)\big)\big), \quad L \in \{LH, HL, HH\} \tag{19}$$

$$\widetilde{f}_{LL} = f_{LL} \tag{20}$$

$$w_H = \sigma\big(MLP\big(\mathrm{Cat}\big(\mathrm{avg}(\widetilde{f}_H), \max(\widetilde{f}_H)\big)\big)\big), \quad H \in \{LH, HL, HH, LL\} \tag{21}$$

$$\bar{f}_H = \widetilde{f}_H \cdot w_H + \widetilde{f}_H \tag{22}$$

$$\hat{f}_i = \mathrm{IDWT}(\bar{f}_{LH}, \bar{f}_{HL}, \bar{f}_{HH}, \bar{f}_{LL}) \tag{23}$$

$$D_i = Conv_{3\times3}\big(\hat{f}_i^1 - \hat{f}_i^2\big) \tag{24}$$

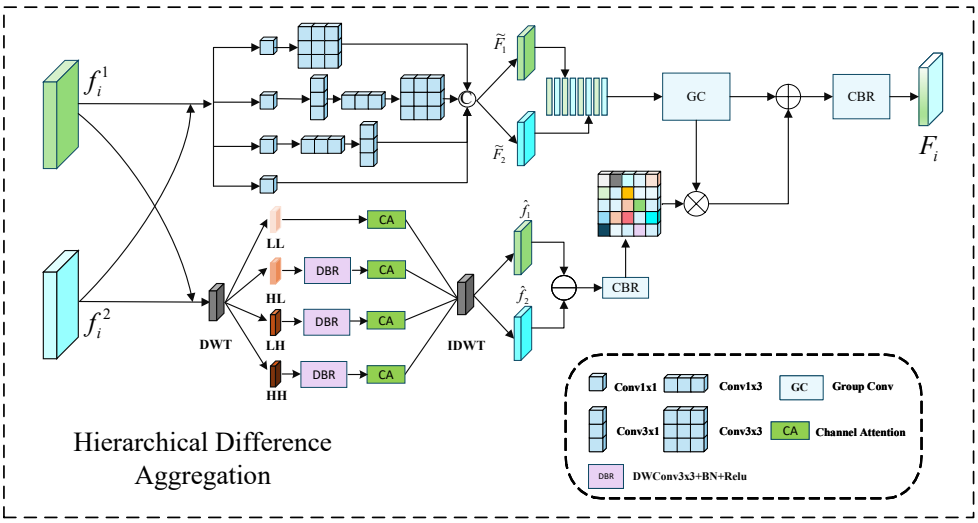


Fig. 3: The structure of the proposed HDA model.

where Cat$(\cdot)$ denotes the concatenation operation, $\widetilde{f}_L$ denotes the high-frequency subbands enhanced by convolution, $\widetilde{f}_{LL}$ denotes the low-frequency subband, and $\hat{f}_i$ denotes the frequency-enhanced feature.

After obtaining the frequency-enhanced difference feature D_i, independent hierarchical feature enhancement is first performed for each temporal feature to enrich the scale diversity and contextual aggregation within each temporal representation, providing a more robust basis for subsequent cross-temporal discrepancy modeling. Specifically, the feature f_i is fed into three parallel branches. The first branch is a local-context branch, which applies a 1×1 convolution for channel squeezing followed by a 3×3 convolution to capture local neighborhood information. The second branch is a horizontally oriented large-receptive-field branch, which combines asymmetric 1×3 and 3×1 convolutions to progressively expand the receptive field along the horizontal and vertical directions, and further adopts a dilated convolution at the end to enlarge the receptive field. The third branch is a vertically oriented large-receptive-field branch, which swaps the order of the 3×1 and 1×3 convolutions to strengthen sensitivity to vertical structures. The mathematical formulations of the three branches are given as follows:

$$f_i^1 = Conv_{3\times3}\big(Conv_{1\times1}(f_i)\big) \tag{25}$$

$$f_i^2 = Conv_{3\times3}\big(Conv_{1\times3}\big(Conv_{3\times1}\big(Conv_{1\times1}(f_i)\big)\big)\big) \tag{26}$$

$$f_i^3 = Conv_{3\times3}\big(Conv_{1\times3}\big(Conv_{3\times1}\big(Conv_{1\times1}(f_i)\big)\big)\big) \tag{27}$$

Subsequently, the features from the three branches are concatenated along the channel dimension and then compressed by a 1×1 convolution. To prevent the enhancement process from excessively altering the original semantics, residual fusion with a scaling factor γ is adopted, formulated as follows:

$$f_s = Conv_{1\times1}\big([f_i^1, f_i^2, f_i^3]\big) \tag{28}$$

$$\widetilde{F}_i = \mathrm{ReLU}\big(\gamma f_s + Conv_{1\times1}(f_i)\big) \tag{29}$$

For the bi-temporal features enhanced by the three-branch module, the two streams are first interleaved along the channel dimension, where each pair of adjacent channels corresponds to semantic information from the two temporal images. Grouped convolution is then applied to obtain the interacted feature F_i^{MC}. This design keeps the parameter count and computational cost controllable, while explicitly learning a "local spatial + temporal interaction" filter for each corresponding channel pair. Consequently, it enhances fine-grained change expression between bi-temporal features without introducing additional cross-semantic channel mixing. The mathematical formulation is given as follows:

$$F_i^{MC} = \mathrm{ReLU}\big(BN\big(Conv_{group\ 3\times3}\big(\widetilde{F}_1, \widetilde{F}_2\big)\big)\big) \tag{30}$$

To explicitly emphasize true changed regions in the spatial domain, the difference feature D_i computed from f_i^1 and f_i^2 is used to perform pixel-wise modulation on the mixed feature F_i^{MC}, yielding the discrepancy-aware feature F_i that integrates discrepancy information. The formulation is given as follows:

$$W_i = \mathrm{ReLU}\big(BN(D_i)\big) \tag{31}$$

$$F_i = \mathrm{ReLU}\big(BN\big(Conv_{3\times3}\big(F_i^{MC}\cdot W_i + F_i^{MC}\big)\big)\big) \tag{32}$$

3.4 Decoder

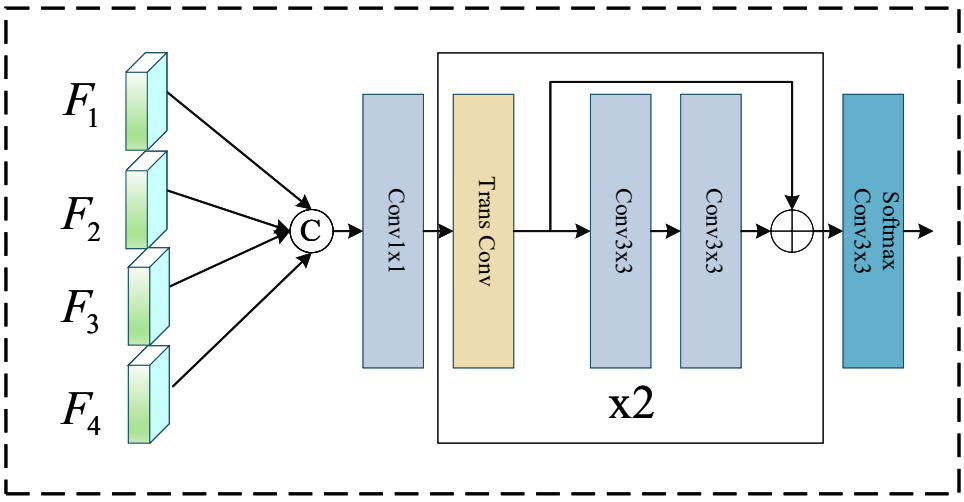


Fig. 4: The structure of the decoder model.

To obtain the final change detection result, a lightweight decoder is adopted, as shown in Figure 4. Specifically, the four-scale features (F_1, F_2, F_3, F_4) are first concatenated along the channel dimension, followed by a 1×1 convolution for feature fusion and channel reduction. Then, a transposed convolution and two stacked 3×3 convolutional layers are used to recover the spatial resolution and refine local details. A residual connection is further introduced to enhance feature propagation. Finally, a 3×3 convolution followed by Softmax is applied to generate the pixel-wise change prediction map.

4 Experiments

4.1 Dataset and Evaluation Metrics

Dataset description: To validate the effectiveness of the proposed network, we conduct experiments on three public remote sensing change detection datasets, namely CLCD, SYSU-CD, and WHU-CD.

CLCD [44] is a high-resolution dataset for cropland change detection, consisting of 600 co-registered bi-temporal remote sensing image pairs. The images were acquired by the Gaofen-2 satellite over Guangdong Province, China, in 2017 and 2019, with spatial resolutions ranging from 0.5 to 2.0 m per pixel. Each sample contains two 512×512 images and a corresponding binary annotation indicating cropland change. Following the official data split, CLCD is divided into 360 training pairs, 120 validation pairs, and 120 test pairs. In this work, CLCD is mainly used to evaluate the model's ability to delineate fine cropland boundaries and detect small parcel-level changes.

SYSU-CD [45] is known for its complex scenes and diverse change types, containing 20,000 pairs of high-resolution aerial images acquired between 2007 and 2014. Each image pair has a spatial size of 256×256 pixels with a consistent resolution of 0.5 m/pixel. The change types include urban expansion, road upgrading, vegetation variation, and coastal construction projects, among others. For training and evaluation, we follow the default split provided by the dataset, with 12,000/4,000/4,000 pairs for training/validation/testing, respectively.

WHU-CD [46] is specifically created for building change detection. A distinctive property of this dataset is that it contains only one pair of ultra-high-resolution aerial images covering a large area. The original image size is 32507×15354 pixels with a resolution of 0.075 m/pixel. Since no official train/test split is provided, we follow a commonly used protocol in the literature by cropping the original images into non-overlapping 256×256 patches. These patches are then split into 5,947 training pairs, 744 validation pairs, and 744 test pairs.

To quantitatively evaluate the performance of DINO-CD, this study employ five metrics to measure the similarity between the predicted results and the ground-truth change regions, including the F1 score (F1), precision (Pre.), recall (Rec.), intersection over union (IoU), and overall accuracy (OA). The definitions of these metrics are as follows:

$$\text{Pr.e.} = \frac{TP}{TP + FP} \tag{33}$$

$$\text{Rec.} = \frac{TP}{TP + FN} \tag{34}$$

$$F1 = \frac{2 \cdot \text{Pr.e.} \cdot \text{Rec.}}{\text{Pr.e.} + \text{Rec.}} \tag{35}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{36}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{37}$$

where TP, FP, TN, and FN denote the numbers of true-positive, false-positive, true-negative, and false-negative pixels, respectively.

4.2 Implementation Details

A pretrained tiny DINOv3 model is adopted, and features are extracted from layers 3, 5, 7, and 11, where each feature map has a size of $16 \times 16 \times 384$ (height $\times$ width $\times$ channels). During training, data augmentation strategies including random flipping, random scale cropping, color jittering, and Gaussian blur are applied to increase training diversity. For fair comparison, all competing methods are trained and evaluated on the same hardware. All experiments are conducted on an NVIDIA GeForce RTX 4090, and all network weights are randomly initialized. The initial learning rate is set to 3.1×10^{-4}, and the numbers of training epochs are 50, 50, and 150 for CLCD, SYSU-CD, and WHU-CD, respectively. The learning rate is empirically selected and linearly decayed to the final epoch during training. AdamW is used as the optimizer with a weight decay of 0.01 and β parameters set to $(0.9, 0.999)$.

Loss function: During network training, we supervise and optimize the model predictions using a pixel-wise cross-entropy loss. For the binary change detection task, the cross-entropy loss can be formulated as:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{38}$$

Where y denotes the ground-truth class label,$\hat{y}$ denotes the predicted class label, and N is the total number of pixels in each sample.

4.3 Comparison with State-of-the-Art Methods

To comprehensively evaluate the proposed DINO-CD for remote sensing change detection, it is compared with a diverse set of representative methods. FC-EF and FC-Diff, as early fully convolutional Siamese frameworks, perform change representation via feature concatenation and feature differencing, respectively. SNUNet employs dense cross-layer connections to enable multi-scale information fusion and detail feedback. BiT and ChangeFormer introduce attention-based Transformer mechanisms to model global bi-temporal interactions and discrepancy relations. USSFC [47] focuses on lightweight spatial–spectral feature learning. ELGC-Net [48] and STFF-GA [49] enhance detection robustness from the perspectives of efficient context representation and pseudo-change suppression, respectively. In addition, two change detection methods that leverage vision foundation model priors are included: SAM-CD [50] and

Meta-CD [51]. SAM-CD exploits the semantic priors of a general segmentation model and incorporates FastSAM to improve few-shot adaptability, while Meta-CD further introduces side adapters on top of this design and cooperates with FastSAM to improve transferability for cross-domain change detection.

4.3.1 Quantitative comparison

Table 1 reports the quantitative comparison between our method and a variety of representative RSCD models on three public datasets, i.e., SYSU-CD, WHU-CD, and CLCD, in terms of Pre., Rec., IoU, F1-score, and OA. Overall, DINO-CD achieves the best or near-best comprehensive performance across all three datasets, exhibiting stable and consistent results under different scene categories and change types.

On the SYSU-CD dataset, DINO-CD significantly outperforms all compared methods in terms of IoU (71.46%) and F1 (83.35%), achieving gains of approximately 3%-4% over STFF-GA, ELGC-Net, and SAM-CD. Meanwhile, DINO-CD attains a higher Recall (79.70%) while maintaining a high Precision (87.36%), indicating that it can better cover true changed regions while effectively suppressing false alarms.

On the WHU-CD dataset, although most methods already achieve strong overall performance, DINO-CD still attains the best results on key metrics, including IoU (89.76%), F1 (94.67%), and OA (99.58%). Moreover, it achieves a Recall of 93.66%, surpassing several Transformer-based methods, which suggests that DINO-CD provides more stable coverage of changed regions in large-scale, structurally regular urban scenes.

On the more challenging CLCD dataset, the performance of most methods drops noticeably, whereas DINO-CD still achieves the best results with an IoU of 66.65% and an F1 score of 79.99%. It also attains a better balance between Precision (84.94%) and Recall (75.58%), demonstrating strong capability in perceiving true changes under complex backgrounds and subtle-change conditions.

Overall, DINO-CD demonstrates stable and consistent advantages across different scenes and change patterns, validating the effectiveness of the proposed DBCR and HDA strategies for remote sensing change detection.

4.3.2 Visual analysis

To visually assess the change detection performance of different methods, we present qualitative comparisons of their predictions on the SYSU-CD, WHU-CD, and CLCD datasets, as shown in Figure 5. Overall, the compared methods exhibit clear differences in terms of change-region completeness, boundary accuracy, and pseudo-change suppression.

On the SYSU-CD dataset, as shown in Figure 5(d-f), most competing methods are easily affected by illumination and texture disturbances under complex backgrounds, producing evident pseudo changes in unchanged areas or yielding breaks and holes within true changed regions. In contrast, DINO-CD exhibits more uniform and continuous responses inside the changed areas, while maintaining good spatial consistency along boundaries. This suggests that, by jointly leveraging the global semantic priors provided by DINOv3 and the local structural constraints from the CNN branch, the model can effectively distinguish pseudo changes caused by appearance perturbations

Table 1: Quantitative comparison of different methods on the SYSU-CD, WHU-CD, and CLCD datasets. The evaluation metrics include Precision (Pre.), Recall (Rec.), Intersection over Union (IoU), F1-score (F1), and Overall Accuracy (OA). The best and second-best results are highlighted in red and blue, respectively.

	SYSU-CD					WHU-CD					CLCD				
Method	Pre.	Rec.	IoU	F1	OA	Pre.	Rec.	IoU	F1	OA	Pre.	Rec.	IoU	F1	OA
FC-EF	74.32	75.84	60.09	75.07	86.02	80.64	71.56	61.07	75.83	98.19	62.92	56.19	42.21	59.37	94.27
FC-Diff	79.13	61.21	56.96	72.57	82.11	47.33	77.66	41.66	58.81	95.63	53.87	57.83	38.68	55.78	93.17
SUN-Net	81.65	72.10	62.05	76.58	89.60	84.29	86.25	74.31	85.26	98.81	59.65	60.39	42.88	60.02	94.01
ChangeF.	81.15	71.65	61.43	76.11	89.39	93.95	88.54	83.76	91.16	99.54	72.47	45.42	38.74	55.84	94.65
BIT	81.54	79.11	65.41	78.33	90.42	93.34	89.39	84.03	91.32	99.31	60.85	47.78	36.54	53.52	93.82
USSFC	82.18	74.49	64.13	78.15	90.18	93.34	89.39	84.03	91.32	99.32	65.33	63.02	47.23	64.16	94.76
ELGC-Net	81.42	77.69	65.99	79.51	90.55	92.72	92.57	86.30	92.64	99.06	69.27	64.12	49.92	66.59	95.21
STFF-GA	86.31	73.03	65.45	79.12	90.91	94.31	92.61	87.71	93.45	99.49	74.09	73.99	58.78	74.04	96.14
SAM-CD	85.56	73.22	65.17	78.91	90.77	91.27	91.34	84.01	91.31	99.31	77.05	61.28	51.82	68.27	95.76
Meta-CD	87.97	69.52	63.48	77.66	90.57	94.58	93.54	86.78	92.80	99.29	64.03	52.78	40.71	57.86	94.28
Ours	87.36	79.70	71.46	83.35	92.49	95.70	93.66	89.76	94.67	99.58	84.94	75.58	66.65	79.99	97.18

Table 2: Ablation study on the WHU-CD and CLCD datasets. DBCR and HDA denote different components. ✓ indicates that the component is included, while × indicates that it is removed. Pre., IoU, and F1 represent Precision, Intersection over Union, and F1-score, respectively.

	Components		WHU-CD			CLCD		
Method	DBCR	HDA	Pre.	IoU	F1	Pre.	IoU	F1
S1	×	×	93.13	85.28	92.06	75.59	59.07	74.27
S2	✓	×	95.25	88.24	93.75	81.43	63.23	77.47
S3	×	✓	94.64	88.06	93.65	81.20	62.29	76.77
S4	✓	✓	95.69	89.76	94.60	84.94	66.65	79.99

from genuine land-cover changes, thereby suppressing false alarms at the semantic level.

On the WHU-CD dataset, as observed in Figure 5(a-c), some methods can roughly localize the changed regions, yet still suffer from blurred building boundaries and local missed detections, especially along narrow edges and adjacent structures where jagged fractures are prone to occur. In contrast, DINO-CD produces change masks with clearer contours, smoother boundaries, and more complete responses inside buildings. This indicates that the HDA module explicitly models bi-temporal discrepancies across multiple scales, which helps preserve global consistency while finely delineating boundary details of regular geometric structures.

On the CLCD dataset, as shown in Figure 5(g-i), most methods exhibit noticeable performance degradation: small-scale changes are easily missed, while spectral or textural differences in unchanged areas are often misclassified as true changes. In contrast, DINO-CD is able to capture fragmented and scattered changed regions more completely and substantially reduce false alarms in the background. This suggests that the

Table 3: Comparison of Parameters, FLOPs, and F1 Scores of Different Methods on the CLCD Dataset.

Method	Param (M)	FLOPs (G)	F1 (%)
FC-EF	1.54	3.57	59.37
FC-Diff	1.66	5.09	55.78
SNUNet	12.03	54.83	60.02
ChangeF	11.94	17.50	55.84
BIT	3.50	17.50	53.32
USSFC	1.52	9.72	64.16
ELGC-Net	10.56	186.96	66.59
STTF-GA	18.93	7.64	74.04
SAM-CD	70.59	17.19	68.27
Meta-CD	47.74	46.11	57.86
DINO-CD	8.03	51.40	79.99

Table 4: Hyperparameter Evaluation Results on the SYSU-CD Dataset.

Layers	Pre.	IoU	F1 (%)
1,2,3,4	80.36	68.70	81.45
7,8,9,10	87.29	69.51	82.01
11,11,11,11	85.81	70.51	82.70
3,6,9,11	87.36	71.46	83.35

DINOv3 semantic branch plays a crucial role in suppressing pseudo changes induced by seasonal appearance variations, while the CNN branch and the difference aggregation mechanism further enhance the model's sensitivity to fine-grained structural changes.

The qualitative results across the three datasets consistently demonstrate that DINO-CD achieves clear and stable advantages in terms of change-region completeness, boundary fidelity, and pseudo-change suppression. Compared with methods that rely solely on local convolutional features or a single global representation strategy, the proposed DBCR and HDA mechanisms enable the model to simultaneously preserve global semantic consistency and capture fine-grained local structures in complex remote sensing scenarios, thereby producing more reliable and robust change detection results.

4.4 Ablation Experiments

To evaluate the contribution of each module to the overall performance, we conduct ablation studies on the WHU-CD and CLCD datasets, and the results are reported in Table 2. Four configurations are considered. S1 is the baseline, which simply interpolates the DINOv3 output features without introducing DBCR or HDA. S2 adds DBCR

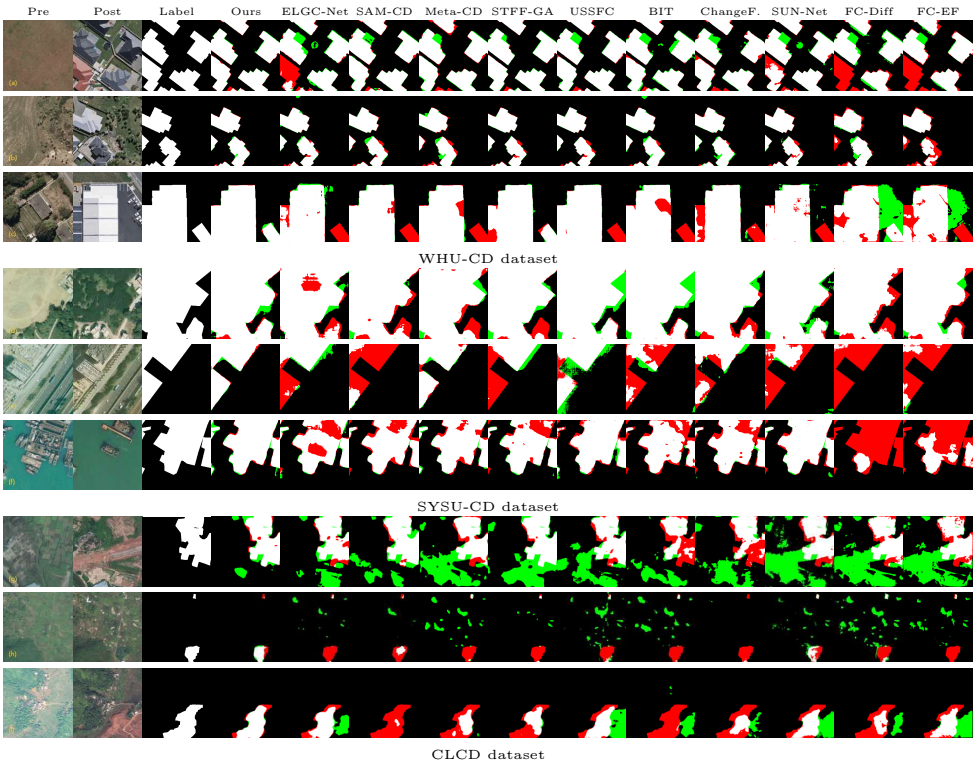


Fig. 5: Qualitative visualization results. (a)-(c), (d)-(f), and (g)-(i) correspond to examples from the WHU-CD, SYSU-CD, and CLCD datasets, respectively. Green denotes false detection pixels, whereas red denotes missed detection pixels.

on top of S1 but does not use HDA. S3 employs HDA for discrepancy representation but removes DBCR. S4 enables both DBCR and HDA, forming the complete model.

As shown in Table 2, S1 yields limited performance on both datasets, achieving IoU and F1 of 85.28% and 92.06% on WHU-CD, and only 59.07% and 74.27% on CLCD. This indicates that relying solely on interpolated single-scale DINOv3 features is insufficient to handle fine-grained changes and complex background disturbances. After introducing DBCR in S2, the performance improves substantially, with clear gains in both IoU and F1-score on WHU-CD and CLCD, demonstrating that DBCR effectively integrates global semantics with local structural details, enhancing feature consistency and boundary delineation. Using HDA alone in S3 also brings consistent improvements, suggesting that multi-scale difference aggregation helps strengthen responses to true changes. However, without feature complementarity, its performance is still constrained by the quality of the input representations. When DBCR and HDA are jointly enabled in S4, the model achieves the best results on both datasets, with more pronounced improvements on CLCD, where change patterns are complex and scales are highly irregular. These results suggest a strong synergy between DBCR

and HDA, where DBCR provides complementary feature foundations and HDA further aggregates hierarchical discrepancy cues, jointly highlighting real changes while suppressing false changes in challenging scenarios.

5 Discussion

5.1 Efficiency Analysis

To evaluate the trade-off between computational efficiency and detection performance of the proposed DINO-CD, we compare the number of parameters, FLOPs, and the F1 score on the CLCD dataset against several representative methods, as reported in Table 3. In terms of model size, conventional CNN-based Siamese architectures, including FC-EF, FC-Diff, and USSFC, contain around 1.5M parameters and incur relatively low computational cost, but their performance is limited in complex scenes and fine-grained change identification. Methods that incorporate multi-scale designs or Transformers, such as SNUNet, ChangeFormer, ELGC-Net, and STFF-GA, typically require more parameters, where STFF-GA and ELGC-Net reach 18.93M and 10.56M parameters, respectively. Foundation-model-based approaches, including SAM-CD and Meta-CD, further increase the parameter count, with SAM-CD as high as 70.59M, leading to a higher deployment cost.

Regarding computational complexity, ELGC-Net requires 186.96G FLOPs, imposing substantial demands on computation and GPU memory. In contrast, DINO-CD requires 51.4G FLOPs, which is moderate and clearly lower than high-complexity models. Considering both efficiency and accuracy, DINO-CD achieves the best F1 score of 79.99% with 8.03M parameters and 51.4G FLOPs, significantly outperforming SAM-CD at 68.27% and Meta-CD at 57.86%. These results indicate that DINO-CD does not rely on excessively large models or heavy computation to gain performance. Instead, it efficiently exploits the robust global semantic priors provided by DINOv3 and the local structural constraints from the CNN branch in a complementary manner, delivering a substantial accuracy improvement with a moderate increase in computational cost, which suggests strong engineering practicality and application potential.

5.2 Hyperparameter Analysis

To analyze the impact of DINOv3 layer selection on performance, a hyperparameter comparison is conducted on the SYSU-CD dataset. With all other training settings kept identical, only the layer combinations used to extract DINOv3 features are varied in the implementation. The results in Table 4 show that different layer selections significantly affect the balance between semantic discrimination and fine-detail preservation. Specifically, selecting shallow layers [1,2,3,4] emphasizes local textures and edge details, but provides relatively weak semantic discriminability, which limits Precision and IoU. Using middle layers [7,8,9,10] strengthens global semantic information and improves Precision, yet still falls short in precise localization for complex change regions. Relying only on deep features [11,11,11,11] further improves IoU and F1, but deep representations are more abstract and may still lose information for fine-grained

structures. Overall, the cross-layer mixed selection [3,6,9,11] achieves the best Precision, IoU, and F1, indicating that combining shallow spatial details with deep semantic representations helps build a more robust multi-scale feature representation, thereby improving both boundary delineation and overall change discrimination. Based on these findings, [3,6,9,11] is adopted as the default DINOv3 layer selection strategy in all experiments.

5.3 Multi-stage Feature Visualization Analysis

As shown in Figure 6, feature responses at key intermediate stages of the network are visualized using heatmaps to analyze the role of each module in change localization and false-change suppression. Specifically, Pre and Post the bi-temporal input images at T0 and T1, and Label is the corresponding change annotation. (a) and (b) are the output features of the DINOv3 branch for T0 and T1, respectively, and (c) and (d) are the complementary reconstructed features produced by DBCR. (e) denotes the difference features obtained by naive subtraction, (f) denotes the aggregated difference features output by HDA, and (g) is the final change probability map produced by the decoder. The visualization is conducted on the CLCD dataset. In the heatmaps, red indicates higher attention or more salient changes, while blue indicates lower attention or less evident changes.

From (a) and (b) in Figure 6, it can be observed that the frozen DINOv3 features can already capture the overall semantic structure of the scene and yield stable responses along major object contours. However, the activations are often biased toward large-scale regions, showing insufficient sensitivity to fine-grained boundaries, slender structures, and small targets. After DBCR, (c) and (d) exhibit more concentrated and spatially continuous high responses at change-related locations, with clearer edge details. This suggests that the semantic priors and convolutional spatial details are effectively aligned and complemented, providing a more robust and discriminative feature basis for the subsequent HDA module. Furthermore, although the naive difference (e) can highlight some change cues, it still contains considerable scattered noise and false-change diffusion. In contrast, the HDA output (f) effectively suppresses background disturbances via hierarchical difference aggregation, making responses over true changed regions more coherent and their boundaries more regular. As a result, the final change map (g) is more consistent with Label, validating the effectiveness of DBCR and HDA in improving change localization accuracy and robustness under complex scenarios.

6 Conclusion

This paper proposes DINO-CD, Marrying DINOv3 with Dual-Branch Complementary Reconstruction for Remote Sensing Change Detection. DINO-CD consists of two key components, namely the DBCR module and the HDA module. Specifically, DBCR aligns and complements DINOv3 global semantic features with convolutional spatial prior features across multiple scales, effectively alleviating the deficiency of fine-grained structures and boundary information caused by the low-resolution and single-scale

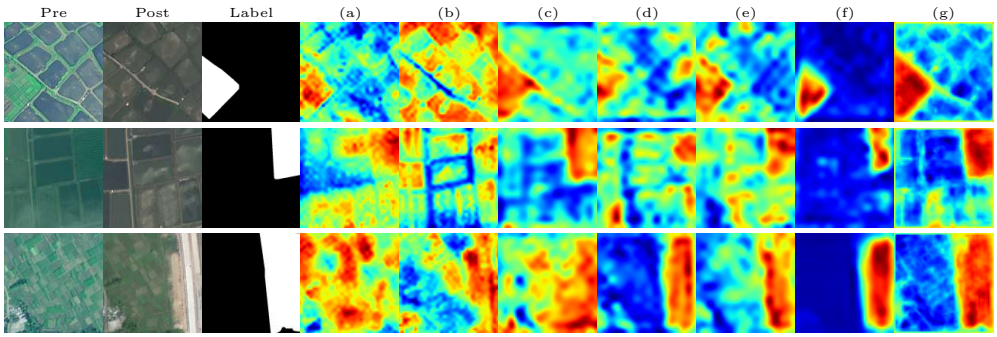


Fig. 6: Visualization of feature representations at different stages of DINO-CD.

nature of the original DINOv3 representations. HDA targets the lack of explicit temporal discrepancy modeling in DINOv3, and strengthens true change cues through multi-scale and multi-directional difference aggregation, thereby reducing false alarms and missed detections. Extensive experiments on multiple public RSCD benchmarks demonstrate that DINO-CD outperforms a range of mainstream RSCD methods, and exhibits stronger robustness, especially under complex backgrounds and subtle-change scenarios. In future work, we will incorporate multimodal information and jointly enhance change cues using complementary modalities, aiming to improve the separability of changed regions and further suppress pseudo-changes induced by variations in illumination, seasons, and imaging conditions.

Acknowledgement This work was supported by National Natural Science Foundation of China under Grant 62201452 and Natural Science Basic Research Program of Shaanxi (Program No. 2024JC-YBQN-0643)

Declarations

Conflict of Interest The authors have no conflict of interest.

Data Availability Data Availability The datasets used in this study are publicly available: WHU-CD for building change detection https://gpcv.whu.edu.cn/data/building_dataset.html, CLCD for fine-grained cropland change detection https://github.com/liumency/CropLand-CD, and SYSU-CD for high-resolution aerial image change detection https://github.com/liumency/SYSU-CD.

Author contributions Bingwen Dou was responsible for the experiments and writing. Xiaopeng Cao, Hailong Ning, and Yizhuo Dong provided guidance on the research direction and reviewed the paper. Qi He contributed to the code programming and polishing of the article.

Ethics approval This study did not require ethics approval.

References

1. Cheng, G., Huang, Y., Li, X., Lyu, S., Xu, Z., Zhao, H., Zhao, Q., Xiang, S.: Change detection methods for remote sensing in the last decade: A comprehensive review. Remote Sensing **16**(13), 2355 (2024) https://doi.org/10.3390/rs16132355

2. Zhu, Z., Qiu, S., Ye, S.: Remote sensing of land change: A multifaceted perspective. Remote Sensing of Environment **282**, 113266 (2022) https://doi.org/10.1016/j.rse.2022.113266

3. Yin, J., Dong, J., Hamm, N.A., Li, Z., Wang, J., Xing, H., Fu, P.: Integrating remote sensing and geospatial big data for urban land use mapping: A review. International Journal of Applied Earth Observation and Geoinformation **103**, 102514 (2021) https://doi.org/10.1016/j.jag.2021.102514

4. Huang, D., Zhou, Z., Zhang, Z., Dai, Q., Lu, H., Li, Y., Huang, Y.: Land use/land cover remote sensing classification in complex subtropical karst environments: Challenges, methodological review, and research frontiers. Applied Sciences **15**(17), 9641 (2025) https://doi.org/10.3390/app15179641

5. Sun, Q., Qiao, R., Jiao, Q., Xing, H., Wang, C., Zhu, X., Huang, W., Zhang, B.: Response of ecological quality to land use/cover change during rapid urbanization of xiong'an new area. Land **13**(12), 2167 (2024) https://doi.org/10.3390/land13122167

6. Kucharczyk, M., Hugenholtz, C.H.: Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. Remote Sensing of Environment **264**, 112577 (2021) https://doi.org/10.1016/j.rse.2021.112577

7. Yang, B., Mao, Y., Liu, L., Liu, X., Ma, Y., Li, J.: From trained to untrained: A novel change detection framework using randomly initialized models with spatial–channel augmentation for hyperspectral images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023) https://doi.org/10.1109/TGRS.2023.3262928

8. Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X.: Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geoscience and Remote Sensing Letters **18**(5), 811–815 (2020) https://doi.org/10.1109/LGRS.2020.2988032

9. Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021) https://doi.org/10.1109/LGRS.2021.3056416

10. Chen, Y., Zhang, J., Shao, Z., Huang, X., Ding, Q., Li, X., Huang, Y.: A siamese

multiscale attention decoding network for building change detection on high-resolution remote sensing images. Remote Sensing **15**(21), 5127 (2023) https://doi.org/10.3390/rs15215127

11. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022) https://doi.org/10.1145/3505244

12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) https://doi.org/arXiv:2010.11929,2020.

13. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision, pp. 280–296 (2022). https://doi.org/10.1007/978-3-031-20077-9_17 . Springer

14. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021) https://doi.org/10.48550/arXiv.2105.15203

15. Yao, J., Wang, X., Yang, S., Wang, B.: Vitmatte: Boosting image matting with pre-trained plain vision transformers. Information Fusion **103**, 102091 (2024) https://doi.org/10.1016/j.inffus.2023.102091

16. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in neural information processing systems **35**, 38571–38584 (2022) https://doi.org/10.48550/arXiv.2204.12484

17. Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2021) https://doi.org/10.1109/TGRS.2021.3095166

18. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 207–210 (2022). https://doi.org/10.1109/IGARSS46834.2022.9883686 . IEEE

19. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021). https://doi.org/10.1109/ICCV48922.2021.00951

20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., *et al.*: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) https:

//doi.org/10.48550/arXiv.2304.07193

21. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., *et al.*: Dinov3. arXiv preprint arXiv:2508.10104 (2025) https://doi.org/10.48550/arXiv.2508.10104

22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023). https://doi.org/10.1109/ICCV51070.2023.00371

23. Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., *et al.*: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) https://doi.org/10.48550/arXiv.2408.00714

24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). https://doi.org/10.48550/arXiv.2103.00020 . PmLR

25. Daudt, R.C., Le Saux, B., Boulch, A.: Fully convolutional siamese networks for change detection. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4063–4067 (2018). https://doi.org/10.1109/ICIP.2018.8451652 . IEEE

26. Peng, D., Bruzzone, L., Zhang, Y., Guan, H., He, P.: Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. International Journal of Applied Earth Observation and Geoinformation **103**, 102465 (2021) https://doi.org/10.1016/j.jag.2021.102465

27. Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021) https://doi.org/10.1109/LGRS.2021.3056416

28. Fang, S., Li, K., Li, Z.: Changer: Feature interaction is what you need for change detection. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–11 (2023) https://doi.org/10.1109/TGRS.2023.3277496

29. Ma, C., Yin, H., Weng, L., Xia, M., Lin, H.: Dafnet: A novel change-detection model for high-resolution remote-sensing imagery based on feature difference and attention mechanism. Remote Sensing **15**(15), 3896 (2023) https://doi.org/10.3390/rs15153896

30. Zhang, R., Zhang, H., Ning, X., Huang, X., Wang, J., Cui, W.: Global-aware siamese network for change detection on remote sensing images. ISPRS journal of photogrammetry and remote sensing **199**, 61–72 (2023) https://doi.org/10.1016/

j.isprsjprs.2023.04.001

31. Huang, Z., Fu, Z., Song, J., Yuan, G., Li, J.: Mfds-net: Multiscale feature depth-supervised network for remote sensing change detection with global semantic and detail information. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024) https://doi.org/10.1109/LGRS.2024.3461957

32. Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2021) https://doi.org/10.1109/TGRS.2021.3095166

33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

34. Tang, Y., Cao, Z., Guo, N., Jiang, M.: A siamese swin-unet for image change detection. Scientific Reports **14**(1), 4577 (2024) https://doi.org/10.1038/s41598-024-54096-8

35. Wang, D., Chen, X., Guo, N., Yi, H., Li, Y.: Stcd: efficient siamese transformers-based change detection method for remote sensing images. Geo-Spatial Information Science **27**(4), 1192–1211 (2024) https://doi.org/10.1080/10095020.2022.2157762

36. Noman, M., Fiaz, M., Cholakkal, H., Narayan, S., Anwer, R.M., Khan, S., Khan, F.S.: Remote sensing change detection with transformers trained from scratch. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024) https://doi.org/10.1109/TGRS.2024.3383800

37. Lu, K., Huang, X., Xia, R., Zhang, P., Shen, J.: Cross attention is all you need: Relational remote sensing change detection with transformer. GIScience & Remote Sensing **61**(1), 2380126 (2024) https://doi.org/10.1080/15481603.2024.2380126

38. Wang, X., Zeng, Z., Li, L.: Dest: difference enhanced-swin transformer for remote sensing change detection. Remote Sensing Letters **15**(12), 1229–1238 (2024) https://doi.org/10.1080/2150704X.2024.2424503

39. Chen, H., Xu, T., Chen, Z., Liu, P., Bai, H., Li, J.: Multi-scale change-aware transformer for remote sensing image change detection. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 2992–3000 (2024). https://doi.org/10.1145/3664647.3680965

40. Zhang, D., Wang, F., Ning, L., Zhao, Z., Gao, J., Li, X.: Integrating sam with feature interaction for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–11 (2024) https://doi.org/10.1109/TGRS.2024.3483775

41. Ning, H., He, Q., Lei, T., Cao, X., Zhang, W., Chen, Y., Nandi, A.K.: Da 2-net: Integrating sam2 with domain adaption and difference aggregation for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing (2025) https://doi.org/10.1109/TGRS.2025.3617980

42. Zhao, Y., Zhang, Y., Dong, Y., Du, B.: Adapting vision transformer for efficient change detection. arXiv preprint arXiv:2312.04869 (2023) https://doi.org/10.48550/arXiv.2312.04869

43. Zhu, D., Huang, X., Huang, H., Cheng, Q., Huang, Z., Shao, Z.: Changevit: Unleashing plain vision transformers for change detection in remote sensing images. Pattern Recognition, 112539 (2025) https://doi.org/10.1016/j.patcog.2025.112539

44. Liu, M., Chai, Z., Deng, H., Liu, R.: A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 4297–4306 (2022) https://doi.org/10.1109/JSTARS.2022.3177235

45. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on geoscience and remote sensing **57**(1), 574–586 (2018) https://doi.org/10.1109/TGRS.2018.2858817

46. Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L.: A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE transactions on geoscience and remote sensing **60**, 1–16 (2021) https://doi.org/10.1109/TGRS.2021.3085870

47. Lei, T., Geng, X., Ning, H., Lv, Z., Gong, M., Jin, Y., Nandi, A.K.: Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023) https://doi.org/10.1109/TGRS.2023.3261273

48. Noman, M., Fiaz, M., Cholakkal, H., Khan, S., Khan, F.S.: Elgc-net: Efficient local–global context aggregation for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–11 (2024) https://doi.org/10.1109/TGRS.2024.3362914

49. Wei, H., Wang, N., Liu, Y., Ma, P., Pang, D., Sui, X., Chen, Q.: Spatio-temporal feature fusion and guide aggregation network for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024) https://doi.org/10.1109/TGRS.2024.3470314

50. Ding, L., Zhu, K., Peng, D., Tang, H., Yang, K., Bruzzone, L.: Adapting segment anything model for change detection in vhr remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–11 (2024) https://doi.org/10.

1109/TGRS.2024.3368168

51. Gao, J., Zhang, D., Wang, F., Ning, L., Zhao, Z., Li, X.: Combining sam with limited data for change detection in remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2025) https://doi.org/10.1109/TGRS.2025.3545040