

# ChangeDINO: DINOv3-Driven Building Change Detection in Optical Remote Sensing Imagery

Ching-Heng Cheng<sup>1</sup>, Chih-Chung Hsu<sup>2</sup>

<sup>1</sup> Inst. of Data Science, National Cheng Kung University,  
Tainan, Taiwan - re6131016@gs.ncku.edu.tw

<sup>2</sup> Inst. of Intelligent Systems College of Artificial Intelligence,  
National Yang Ming Chiao Tung University, Hsinchu, Taiwan - chihchung@nycu.edu.tw

**Keywords:** Change Detection, Optical Image, Deep Learning, Foundation Model, Transformer, Morphology.

## Abstract

Remote sensing change detection (RSCD) aims to identify pixel-wise surface changes from co-registered bi-temporal images. However, many deep learning-based RSCD methods rely solely on change-map annotations and underuse the semantic information in non-changing regions, which limits robustness under illumination variation, off-nadir views, and scarce labels. This paper presents ChangeDINO, an end-to-end multiscale Siamese framework for optical building change detection. The model fuses a lightweight backbone stream with features transferred from a frozen DINOv3, yielding semantic- and context-rich pyramids even on small datasets. A spatial-spectral differential transformer decoder then exploits multi-scale absolute differences as change priors to highlight true building changes and suppress irrelevant responses. Finally, a learnable morphology module refines the upsampled logits to recover clean boundaries. Experiments on four public benchmarks demonstrate that ChangeDINO achieves strong accuracy and robustness under cross-temporal appearance variations, yielding cleaner building boundaries with improved data efficiency.

## 1. Introduction

Remote sensing change detection (RSCD) in multi-temporal remote sensing imagery is central to Earth observation. By comparing satellite or aerial images across time, RSCD reveals land-cover dynamics from natural and human activities (Asokan and Anitha, 2019). Buildings are an important focus area because changes inform urban planning, regulatory compliance, and risk assessment. With high-resolution aerial, drone, and satellite imagery, deep learning enables large-scale monitoring, detection of subtle structural changes, and data-driven support for infrastructure management and sustainable development (Peng et al., 2025).

RSCD remains challenging due to cross-temporal and cross-domain variability. Image pairs may come from different sensors such as optical, multispectral, or SAR, or they may differ in illumination, seasonality, and viewing geometry even within one modality (Chen et al., 2024b). These factors introduce spectral and geometric inconsistencies unrelated to true changes (Hussain et al., 2013). Traditional hand-crafted methods struggle with such variability, while deep learning has become the prevailing approach by learning hierarchical, task-specific representations with stronger robustness and generalization (Daudt et al., 2018).

However, current models face limited data scale and architectural constraints. Many RSCD datasets are small, geographically narrow, or task-specific labeled, encouraging overfitting and limiting access to global semantic context (Ding et al., 2025). Mainstream Siamese pipelines (Daudt et al., 2018) fuse multiscale features with convolutional or transformer decoders, yet often miss fine pixel-level differences and are influenced by irrelevant context. Downsampling for efficiency and later up-sampling also blurs boundaries.

To address these challenges, we propose **ChangeDINO**, a multiscale Siamese framework. It leverages the pretrained

DINOv3 foundation model as the encoder, introduces a differential transformer-based decoder, originally proposed in the large-language-model domain (Ye et al., 2024), to reason over cross-temporal context and suppress noise, and adopts a learnable morphological module for final mask refinement. The main contributions of ChangeDINO include:

- Leverage DINOv3 pretrained in the encoder to inject semantically rich features without requiring task-specific semantic labels.
- Propose a differential transformer-based decoder that strengthens attention to relevant cross-temporal context for precise, pixel-level change modeling while suppressing distractors.
- Introduce a learnable, morphology-based refinement head with trainable structuring kernels that denoise predictions and sharpen subtle-change boundaries in end-to-end training.

## 2. Related Work

### 2.1 Conventional Methods

Classical RSCD methods mainly rely on pixel-wise algebra or statistical analysis of bi-temporal imagery (Chen et al., 2024b, Asokan and Anitha, 2019). Representative examples include image differencing, ratio-based transforms, and feature transformations for emphasizing change-related variation (Coppin and Bauer, 1996, Liu et al., 2017, Jimenez-Sierra et al., 2022). Other common techniques include Change Vector Analysis (CVA) (He et al., 2014), PCA-like transforms (Celik, 2009), and post-classification comparison (PCC), which produces class-to-class transition maps but may propagate classification errors (Qi et al., 2015). To improve spatial consistency,

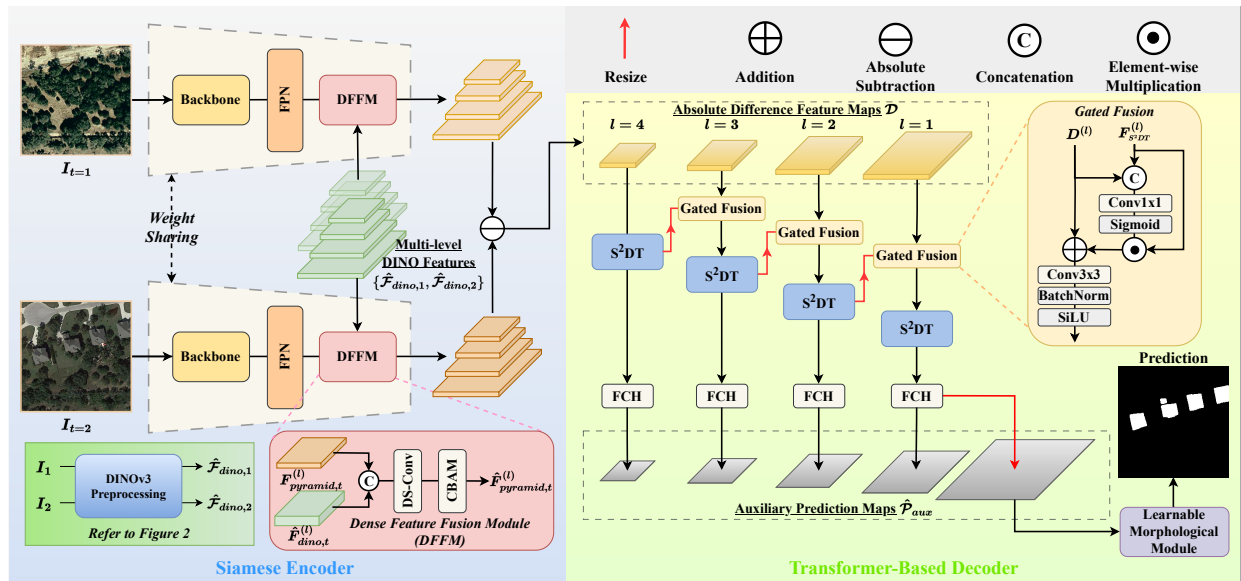


Figure 1. Overall architecture of ChangeDINO. A multiscale encoder–decoder for optical building change detection.

object-based image analysis (OBIA) segments images into geo-objects prior to comparison (Johansen et al., 2010), while morphological operators are often used as post-processing to denoise and regularize boundaries (Dalla Mura et al., 2008).

## 2.2 Deep Learning-Based Methods

Deep learning has reformulated RSCD as a supervised segmentation problem on image pairs. Early CNN-based methods include Fully Convolutional Early Fusion (FC-EF) and Fully Convolutional Siamese (FC-Siam) (Daudt et al., 2018). Building on this paradigm, later Siamese architectures improved multi-scale alignment and feature interaction, such as SNUNet-CD (Fang et al., 2021) and IFNet (Zhang et al., 2020). Existing deep RSCD models can be broadly grouped into *difference-based* methods, which explicitly inject image or feature differences (Li et al., 2023, Wang et al., 2023), and *fusion-based* methods, which combine multi-scale bi-temporal features through concatenation or attention (Han et al., 2023, Zhang et al., 2024).

Transformer-based models further enhance long-range context modeling. Bi-temporal Image Transformer (BIT) (Chen et al., 2021) introduces transformer encoding over bi-temporal tokens, while ChangeFormer (Bandara and Patel, 2022) adopts hierarchical vision transformers for stronger cross-temporal interaction. More recently, foundation-model approaches such as ChangeCLIP (Dong et al., 2024) exploit vision-language pre-training, and Mamba-based variants (Zhang et al., 2025, Chen et al., 2024a) provide efficient long-range modeling through state-space designs. In parallel, self-supervised domain-adaptation methods (Lebedev et al., 2018, Niu et al., 2018) and GAN-based augmentation (Knyaz et al., 2024) have also been explored to improve robustness under cross-sensor and cross-seasonal shifts.

Overall, RSCD has evolved from hand-crafted differencing to context-aware neural architectures. However, cross-domain generalization, fine boundary fidelity, and data efficiency remain challenging (Peng et al., 2025, Ding et al., 2025), motivating methods that combine large-scale semantic priors with precise differential reasoning and morphology-aware refinement.

## 3. Methodology

### 3.1 Method Overview

As illustrated in Fig. 1, ChangeDINO tackles optical building change detection under cross-temporal appearance variations and limited annotations by combining semantic priors, change-focused context reasoning, and boundary-aware refinement. A pair of cross-temporal optical images  $\{I_t\}_{t \in \{1,2\}}$  is processed by a Siamese encoder to extract multi-scale features. The encoder fuses a lightweight backbone with an FPN branch and semantic features transferred from a frozen DINOv3, so that the resulting pyramids remain robust to illumination and seasonal shifts while preserving building structures. From the fused pyramids, we compute multi-scale absolute-difference maps  $\mathcal{D}$  as change priors, which explicitly highlight candidate change evidence and reduce ambiguity for subsequent decoding. Guided by  $\mathcal{D}$ , a Spatial–Spectral Differential Transformer decoder performs cross-temporal reasoning with spatial and channel-wise attention, emphasizing truly changed regions and suppressing irrelevant responses. At each pyramid level, fully convolutional heads provide auxiliary predictions for deep supervision, and the final high-resolution logit is produced by a gated residual upsampling head to avoid oversmoothing. Finally, since downsampling–upsampling pipelines can still yield fragmented interiors or jagged boundaries, we append a learnable morphology module with trainable structuring kernels to denoise predictions and sharpen building boundaries, yielding the final change mask.

### 3.2 Semantic and Context-Rich Multi-Scale Features via Siamese Encoder with DINOv3 Pretraining

Given a cross-temporal optical pair  $\{I_t\}_{t \in \{1,2\}}$ , we adopt a lightweight convolutional backbone (MobileNet (Sandler et al., 2019)) followed by an FPN (Wang et al., 2023) to construct multi-scale representations. Let the backbone and FPN be  $\Phi_{\text{backbone}}$  and  $\Phi_{\text{fpn}}$ , respectively. For each time  $t$ , the backbone extracts per-level features that are aggregated by the FPN into a top–down pyramid:

$$\mathcal{F}_{\text{pyramid},t} = \{F_{\text{pyramid},t}^{(l)}\}_{l=1}^4 = \Phi_{\text{fpn}}(\Phi_{\text{backbone}}(I_t)), \quad (1)$$

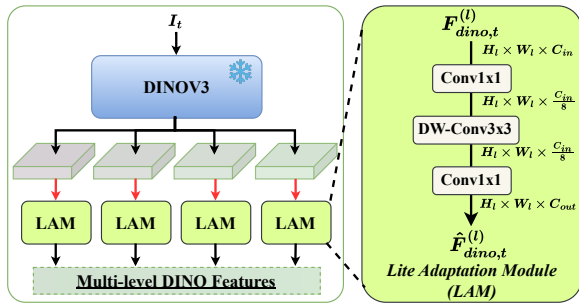


Figure 2. Lightweight feature adapter for aligning DINOv3 features with our backbone. The snowflake icon indicates that the DINOv3 weights are frozen.

where  $l$  indexes the pyramid levels.

When trained on limited data without strong semantic supervision, the pyramid features  $\mathcal{F}_{\text{pyramid},t}$  tend to lack context. To inject semantic priors, we incorporate the large-scale pre-trained foundation model DINOv3 (Siméoni et al., 2025). To avoid catastrophic forgetting, DINOv3 is frozen (Fig. 2) and four intermediate, semantics-rich feature maps are tapped, resized to the corresponding pyramid scales, and passed through a *Lite Adaptation Module* (LAM) to align channels and distill semantics:

$$\mathcal{F}_{\text{dino},t} = \text{DINOv3}(I_t) = \{F_{\text{dino},t}^{(l)}\}_{l=1}^4, \quad (2)$$

$$\hat{\mathcal{F}}_{\text{dino},t} = \{\hat{F}_{\text{dino},t}^{(l)}\}_{l=1}^4 = \{\Phi_{\text{LAM}}^{(l)}(\text{Resize}(F_{\text{dino},t}^{(l)}))\}_{l=1}^4, \quad (3)$$

where  $\text{DINOv3}(\cdot)$  denotes the frozen DINOv3 model,  $\Phi_{\text{LAM}}^{(l)}(\cdot)$  the  $l$ -th lightweight adapter, and  $\text{Resize}(\cdot)$  denotes bilinear interpolation resizing.

To fuse the adapted DINO features with the backbone-FPN pyramid, we employ a *Dense Feature Fusion Module* (DFFM) that concatenates the two streams, applies a depthwise-separable convolution (Chollet, 2017), and uses convolutional block attention modules (CBAM) (Woo et al., 2018) to produce a semantic- and context-rich pyramid:

$$\hat{\mathcal{F}}_{\text{pyramid},t}^{(l)} = \Phi_{\text{DFFM}}^{(l)}(F_{\text{pyramid},t}^{(l)}, \hat{F}_{\text{dino},t}^{(l)}), \quad l = 1, \dots, 4, \quad (4)$$

where  $F_{\text{pyramid},t}^{(l)}$  and  $\hat{F}_{\text{dino},t}^{(l)}$  are defined in Eqs. (1) and (3), respectively, and  $l$  indexes pyramid levels, with larger  $l$  corresponding to deeper, lower-resolution features.

At the end of the encoder, we compute multi-scale element-wise absolute differences  $\mathcal{D} = \{D^{(l)}\}_{l=1}^4$  between the bi-temporal pyramids to obtain the decoder inputs:

$$D^{(l)} = \left| \hat{\mathcal{F}}_{\text{pyramid},1}^{(l)} - \hat{\mathcal{F}}_{\text{pyramid},2}^{(l)} \right|. \quad (5)$$

Overall during training, the backbone+FPN branch adapts to task-specific details and local structures, while the frozen DINOv3 branch supplies robust semantic context; their fusion yields multi-scale representations that are both fine-grained and semantically consistent, from which we compute per-level absolute differences as a multi-scale change prior to drive the decoder.

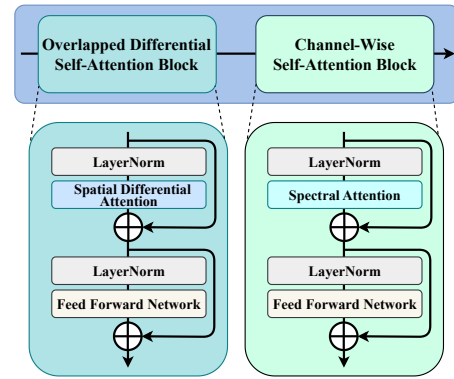


Figure 3. **Spatial-spectral Differential Transformer (S<sup>2</sup>DT) block.** Incorporates a differential transformer into overlapped-window spatial self-attention and pairs it with channel-wise self-attention to refine feature intensities.

### 3.3 Building-Aware Decoder via Spatial and Spectral Self-Attention Decoder

The decoder upsamples and transforms the fused bi-temporal features, specifically the multi-scale absolute-difference maps  $\mathcal{D}$  (Eq. 5), into a high-resolution change map. It reconstructs spatial detail and produces a pixel-wise prediction that separates changed from unchanged regions.

Guided by  $\mathcal{D}$  as a change prior, we design a *Spatial-Spectral Differential Transformer* (S<sup>2</sup>DT), illustrated in Fig. 3, a transformer module that fuses spatial and spectral (channel-wise) self-attention and instantiates the differential transformer (Ye et al., 2024) as the core attention mechanism. Differential transformers, originally validated in large-language-model settings for focusing on informative tokens while suppressing noise, are well suited here for filtering illumination-induced or misregistration artifacts and other irrelevant responses.

The spatial differential attention in the differential self-attention block of S<sup>2</sup>DT is summarized as follows. Given a feature map  $X \in \mathbb{R}^{C \times H \times W}$ , we obtain queries, keys, and values via  $1 \times 1$  convolutions and reshape them into  $h$  heads with token length  $N=HW$  and head width  $d=C/h$ :

$$\begin{aligned} Q, K, V &\in \mathbb{R}^{h \times N \times 2d}, \\ Q &= [Q_1; Q_2], \quad K = [K_1; K_2], \end{aligned} \quad (6)$$

where  $[\cdot; \cdot]$  denotes channel splitting. Two spatial attentions are computed on the halves:

$$A_1 = \text{softmax}\left(\frac{Q_1 K_1^\top}{\sqrt{d}}\right), \quad A_2 = \text{softmax}\left(\frac{Q_2 K_2^\top}{\sqrt{d}}\right). \quad (7)$$

They are combined in a multihead differential form:

$$\begin{aligned} \tilde{X} &= (A_1 - \Lambda A_2) V \in \mathbb{R}^{h \times N \times 2d}, \\ \tilde{X} &= (\tilde{X}^{(i)})_{i=1}^h, \quad \tilde{X}^{(i)} \in \mathbb{R}^{N \times 2d}, \end{aligned} \quad (8)$$

where  $\Lambda = \text{diag}(\lambda^{(1)}, \dots, \lambda^{(h)})$  holds per-head, positive, learnable coefficients. The per-head outputs are normalized with RMSNorm (Zhang and Sennrich, 2019), concatenated, and projected back to the spatial tensor. The spatial differential at-

tention operator  $\Phi_{\text{SDA}}$  is defined as:

$$\Phi_{\text{SDA}}(X) = \mathbf{W}_{\text{prj}} \text{Concat}_{i=1}^h(\text{RMSNorm}(\tilde{X}^{(i)})) \in \mathbb{R}^{C \times H \times W}, \quad (9)$$

where  $\mathbf{W}_{\text{prj}}$  denotes the projection weights. This differential attention ( $A_1 - \Lambda A_2$ ) emphasizes informative spatial correspondences while attenuating distractors, yielding sharper and cleaner responses for change localization.

Consequently,  $S^2\text{DT}$  targets pixel-level change discrimination across pyramid levels. To coordinate information across scales, we adopt a gated-fusion operator  $G(\cdot)$ , illustrated in Fig. 1, that adaptively controls cross-level contributions. For level  $l$ ,

$$F_{S^2\text{DT}}^{(l)} = \begin{cases} G(D^{(l)}, F_{S^2\text{DT}}^{(l+1)}), & l = 1, 2, 3, \\ \Phi_{S^2\text{DT}}^{(l)}(D^{(l)}), & l = 4, \end{cases} \quad (10)$$

where  $\Phi_{S^2\text{DT}}^{(l)}(\cdot)$  denotes the  $S^2\text{DT}$  block at level  $l$ . Fully convolutional heads (FCHs) then produce auxiliary predictions for deep supervision:

$$\hat{P}_{\text{aux}}^{(l)} = \text{Upsample}\left(H^{(l)}\left(F_{S^2\text{DT}}^{(l)}\right)\right), \quad l = 1, \dots, 4, \quad (11)$$

where  $H^{(l)}(\cdot)$  denotes the per-level FCH that maps  $S^2\text{DT}$  features to a binary change logit map, and  $\text{Upsample}(\cdot)$  denotes bilinear interpolation.

### 3.4 Refining Binary Prediction Using Learnable Morphology

Direct upsampling via interpolation can produce a reasonably structured change logit, but it may still contain fragmented interiors and spurious protrusions. Classical morphology helps alleviate this issue, but fixed structuring elements often over-smooth edges or remove fine details. We therefore propose a Learnable Morphological Module (LMM), shown in Fig. 4, which adopts *learnable* structuring elements (fixed window size, trainable weights) and fuses the refined result with the original logits in an end-to-end manner. The final prediction  $\hat{P}$  is

$$\hat{P} = \alpha \sigma^{-1}\left(\text{Closing}\left(\text{Opening}\left(\sigma(\hat{P}_{\text{aux}}^{(1)}), \Omega_1\right), \Omega_2\right)\right) + (1 - \alpha) \hat{P}_{\text{aux}}^{(1)}, \quad (12)$$

where  $\alpha \in [0, 1]$  is a learnable mixing weight;  $\sigma(\cdot)$  denotes the sigmoid function and  $\sigma^{-1}(\cdot)$  its inverse (logit), and  $\Omega_1, \Omega_2$  are learnable structuring kernels with window sizes of 3 and 5, respectively. We define opening and closing as

$$\text{Opening}(M, \Omega) = \text{Dil}(\text{Ero}(M, \Omega), \Omega), \quad (13)$$

$$\text{Closing}(M, \Omega) = \text{Ero}(\text{Dil}(M, \Omega), \Omega), \quad (14)$$

with  $\text{Ero}(\cdot, \cdot)$  and  $\text{Dil}(\cdot, \cdot)$  denoting differentiable erosion and dilation implemented via softmin/softmax approximations. This morphology head removes noise while preserving building shapes and connectivity.

Finally, we obtain the binary change map  $\hat{P}_{\text{bin}}$  by thresholding the logit prediction  $\hat{P}$ .

### 3.5 Objective Function

Let  $Y \in \{0, 1\}^{H \times W}$  denote the binary ground truth,  $\hat{P} \in \mathbb{R}^{H \times W}$  the final prediction logits, and  $\{\hat{P}_{\text{aux}}^{(l)}\}_{l=1}^4$  the auxil-

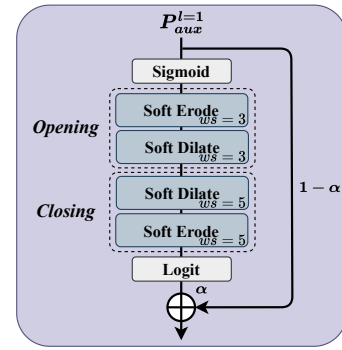


Figure 4. **Learnable Morphological Module (LMM)**. Classical opening and closing with learnable structuring elements further refine the prediction.

ary prediction logits. To address the foreground–background class imbalance, we use two popular losses, Focal loss (Lin et al., 2018)  $\mathcal{L}_{\text{focal}}$  and Dice loss (Sudre et al., 2017)  $\mathcal{L}_{\text{dice}}$ , and weight them with a hyperparameter  $\beta$ . We also apply deep supervision to the auxiliary maps:

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{focal}}(\hat{P}, Y) + \beta \mathcal{L}_{\text{dice}}(\hat{P}, Y), \quad (15)$$

$$\mathcal{L}_{\text{aux}} = \sum_{l=1}^4 \left[ \mathcal{L}_{\text{focal}}(\hat{P}_{\text{aux}}^{(l)}, Y) + \beta \mathcal{L}_{\text{dice}}(\hat{P}_{\text{aux}}^{(l)}, Y) \right], \quad (16)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + 0.5 \mathcal{L}_{\text{aux}}. \quad (17)$$

We set  $\beta = 0.5$  for all experiments.

## 4. Experiments

### 4.1 Experiment Details

**Benchmark Datasets.** We evaluate on four public change detection datasets: LEVIR-CD (Chen and Shi, 2020), WHU-CD (Ji et al., 2019), S2Looking-CD (Shen et al., 2021), and SYSU-CD (Shi et al., 2022), the cross-temporal image pairs in all datasets are well registered. We use the same data and official splits as in prior work (Han et al., 2023) to ensure a fair apples-to-apples comparison. LEVIR-CD, WHU-CD, and S2Looking-CD are object-specific, focusing on building change detection (BCD), whereas SYSU-CD is category-agnostic. LEVIR-CD contains Google Earth image-pair patches collected across 20 regions with 5–14 year intervals. WHU-CD comprises aerial images of the same area before and after an earthquake. S2Looking-CD consists of large side-looking satellite pairs captured at different off-nadir angles, with significant illumination variation and extensive rural scenes. SYSU-CD includes diverse change types such as road expansion, newly built urban structures, vegetation changes, suburban sprawl, and ground-work prior to construction, and we include it to evaluate the generalization ability of our method beyond building-specific changes. The numbers of training/validation/testing pairs are 7,120/1,024/2,048 (LEVIR-CD), 4,536/504/2,760 (WHU-CD), 56,000/8,000/16,000 (S2Looking-CD), and 12,000/4,000/4,000 (SYSU-CD). Following common practice, we tile image pairs into  $256 \times 256$  patches without overlap and apply no additional curation beyond the official partitions.

**Evaluation Metrics.** For the binary change detection performance evaluation, we use the Intersection-over-Union (IoU), F1 score, precision (Pre.) and recall (Rec.) for the change class.

Table 1. Quantitative comparisons in terms of IoU, F1, Precision (Pre.), and Recall (Rec.) on four RSCD benchmarks.

Methods	LEVIR-CD				WHU-CD				S2Looking-CD				SYSU-CD				Model Efficiency	
	IoU	F1	Pre.	Rec.	IoU	F1	Pre.	Rec.	IoU	F1	Pre.	Rec.	IoU	F1	Pre.	Rec.	Params (M)	Time (ms)
FC-Diff	73.23	84.55	89.18	80.37	63.63	77.78	89.29	68.89	24.61	39.50	<b>80.20</b>	26.10	42.03	59.18	<b>90.31</b>	44.01	1.35	2.47
IFNet	84.51	91.60	<b>93.63</b>	89.66	81.52	89.82	87.47	<u>92.30</u>	44.76	61.84	69.06	55.98	66.02	79.53	87.30	73.04	35.73	6.90
BIT	81.72	89.94	90.33	89.56	68.02	80.97	74.01	89.37	45.62	62.65	70.26	56.53	57.88	73.32	75.15	71.58	3.50	6.32
ChangeFormer	81.69	89.92	91.70	88.20	78.51	87.96	91.07	85.06	46.69	63.65	69.51	58.71	64.29	78.26	78.17	78.36	41.03	15.00
A2Net	84.30	91.48	92.10	90.87	86.66	92.85	95.19	90.62	48.39	65.61	69.21	61.66	71.37	83.29	86.54	80.28	3.78	9.28
CGNet	85.21	92.01	93.15	90.90	86.21	92.59	94.47	90.79	46.78	63.74	70.72	58.02	66.55	79.92	86.37	74.37	33.68	4.67
CLAFA	<u>85.31</u>	<u>92.07</u>	<u>93.26</u>	90.91	<u>87.80</u>	<u>93.50</u>	<u>95.51</u>	91.58	48.75	65.55	71.09	60.81	70.10	82.43	84.38	80.56	5.06	14.62
BiFA	82.65	90.50	91.56	89.46	86.75	92.91	94.41	91.45	45.70	62.73	65.15	60.49	<u>71.73</u>	<u>83.53</u>	87.05	80.29	5.58	2.77
WS-Net++	83.86	91.22	92.65	89.84	86.96	93.03	94.82	91.30	<u>49.54</u>	<u>66.26</u>	69.50	<b>63.30</b>	70.13	82.44	87.05	78.29	237.54	276.13
ChangeCLIP*	85.26	92.04	92.62	91.47	81.91	90.05	92.59	87.65	-	-	-	-	70.46	82.67	84.89	80.56	63.82	24.76
ChangeRD	78.85	88.18	90.36	86.10	75.22	85.86	91.44	80.92	29.73	45.84	62.10	36.33	59.38	74.52	79.87	69.83	29.47	14.54
CDMamba	82.38	90.34	91.57	89.15	83.71	91.13	94.93	87.63	44.85	61.92	65.44	58.77	65.72	79.32	81.01	77.69	8.88	39.09
<b>Ours</b>	<b>85.72</b>	<b>92.31</b>	92.47	<b>92.15</b>	<b>89.00</b>	<b>94.18</b>	<b>95.69</b>	<b>92.72</b>	<b>50.52</b>	<b>67.13</b>	<u>71.63</u>	<u>63.17</u>	<b>73.46</b>	<b>84.70</b>	<u>87.87</u>	<b>81.75</b>	234.03	73.62

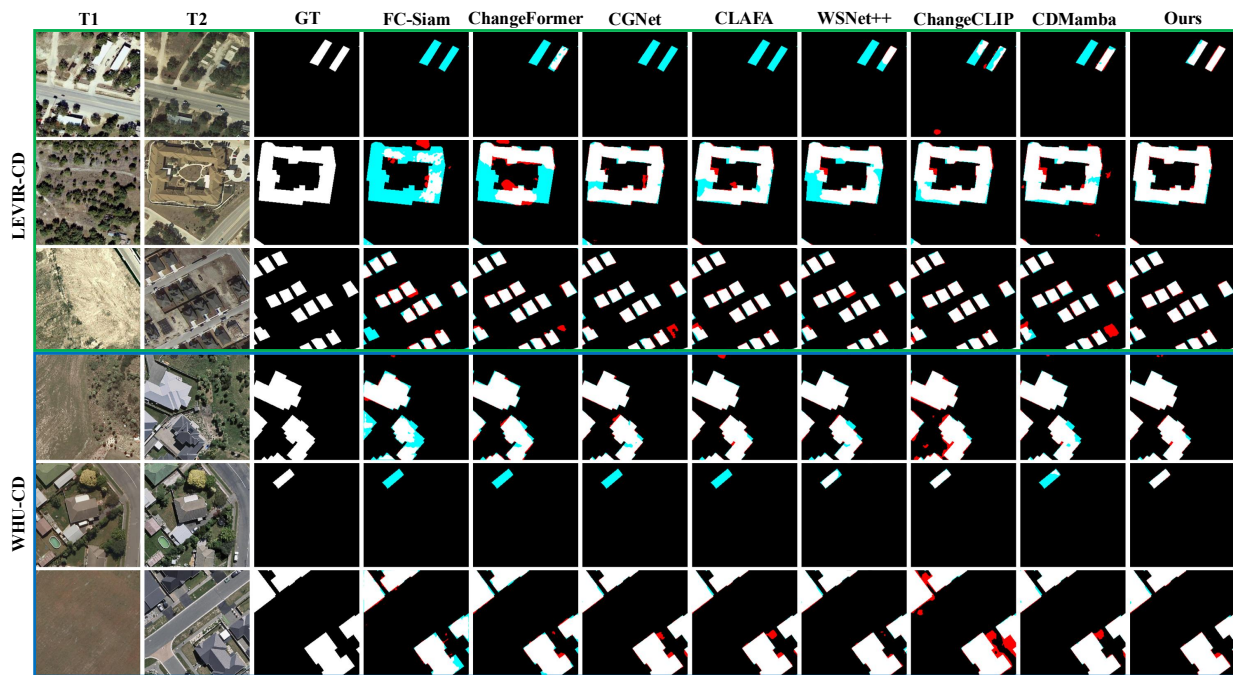


Figure 5. Qualitative experimental results on LEVIR-CD and WHU-CD.

**Compared Methods.** To evaluate the proposed ChangeDINO, we compare it against a broad set of state-of-the-art RSCD methods. As noted in *Related Work* section, recent RSCD research is dominated by deep learning, primarily along two lines: CNN-based and Transformer-based models. Accordingly, we benchmark against multiple representative approaches: five CNN-based (FC-Diff (Daudt et al., 2018), IFNet (Zhang et al., 2020), A2Net (Li et al., 2023), CGNet (Han et al., 2023), CLAFA (Wang et al., 2023)) and four Transformer-based (BIT (Chen et al., 2021), ChangeFormer (Bandara and Patel, 2022), BiFA (Zhang et al., 2024), ChangeRD (Jing et al., 2025)). In addition, we include three recent frameworks that reflect emerging trends: a multi-model-based method (ChangeCLIP (Dong et al., 2024)), a domain-adaptation method (WS-Net++ (Xiong et al., 2024)), and a Mamba-based method (CDMamba (Zhang et al., 2025)).

**Implementation Details.** The proposed ChangeDINO is implemented in PyTorch (Paszke et al., 2019) and trained/evaluated on a single NVIDIA RTX 3090 GPU. During training, input pairs are cropped to  $256 \times 256$  with a batch size of 16. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of  $5 \times 10^{-4}$  (set to  $1 \times 10^{-4}$  for WHU-CD), and apply a cosine decay schedule

down to  $1 \times 10^{-7}$ . Models are trained for 100 epochs on LEVIR-CD and WHU-CD, and for 50 epochs on SYSU-CD and S2Looking-CD. Standard data augmentation (random rotations, flips, and crops) is applied during training.

## 4.2 Quantitative and Qualitative Results

In quantitative comparisons, the **best** and **second best** results are highlighted in bold and underline, respectively, for clarity. LEVIR-CD and WHU-CD are comparatively easier benchmarks, whereas S2Looking-CD and SYSU-CD are more challenging: the former involves large side-looking imagery with varying off-nadir angles, and the latter is category-agnostic with diverse change types. Tab. 1 reports the quantitative results of our method against the recent state of the art. An asterisk (\*) on ChangeCLIP indicates the use of officially released pretrained weights only. Across all four datasets, ChangeDINO outperforms all SOTAs and attains the best IoU and F1 scores. CLAFA performs strongly on LEVIR-CD and SYSU-CD but lags on S2Looking-CD, while WS-Net++ excels on WHU-CD and S2Looking-CD yet trails on LEVIR-CD. ChangeCLIP is also highly competitive, reinforcing that foundation/model-pretrained representations are an effective and growing direction for RSCD. It is worth noting

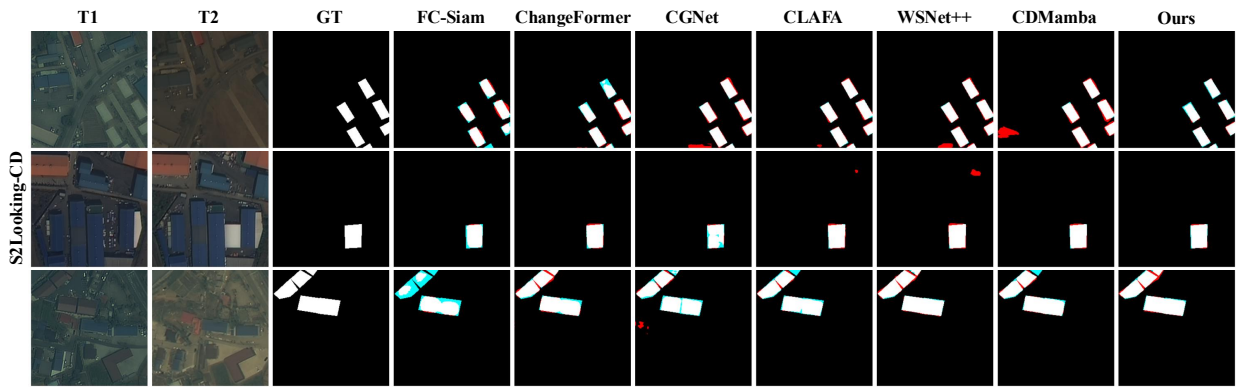


Figure 6. Qualitative experimental results on S2Looking-CD.

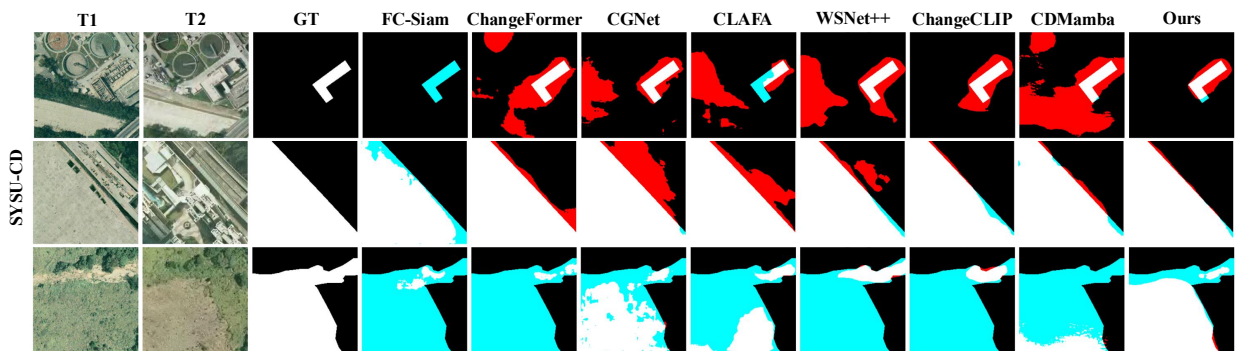


Figure 7. Qualitative experimental results on SYSU-CD

that FC-Diff and IFNet achieve high precision but underperform in other metrics, likely because they favor only the most salient changes and miss subtle or boundary pixels. We also report model efficiency in the rightmost columns of Tab. 1. Inference time is measured as the average latency over 100 runs on  $256 \times 256$  patches. Although ChangeDINO has higher parameter count and latency than most baselines, this largely stems from using a frozen ViT foundation model (DINOv3) to inject semantic priors. Nevertheless, RSCD is commonly performed in offline, ground-based processing pipelines, where strict real-time inference is typically not a primary requirement. As the only category-agnostic benchmark, SYSU-CD is included to assess generalization beyond buildings. The competitive results suggest that ChangeDINO remains effective under heterogeneous land-cover changes, which we attribute in part to DINOv3 semantic priors and the noise-suppressing property of  $S^2DT$ . For qualitative comparison, due to space constraints, we select representative methods spanning different RSCD paradigms to visualize predictions. As shown in Fig. 5, on LEVIR-CD and WHU-CD our method exhibits the most accurate delineation of changes with fewer false positives (FP) and false negatives (FN). Note that true positives (TP) and true negatives (TN) are rendered in white and black, respectively, while FP and FN are rendered in red and cyan, respectively. Figs. 6 and 7 present results on the more challenging benchmarks. ChangeDINO maintains superior visual quality, producing cleaner masks with sharper building boundaries and less spurious noise. To further illustrate the effect of each component, we select six representative scenes from LEVIR-CD and SYSU-CD and visualize (i) the difference features at two pyramid levels ( $D^3$  and  $D^1$ ), (ii) the corresponding  $S^2DT$  features, and (iii) the final prediction after LMM ( $\hat{P}$ ), as shown in Fig. 8. PCA is used for all visualizations. The difference maps show that deep levels already highlight potential change regions, while shallow levels better pre-

serve building structure and suppress irrelevant objects. After the  $S^2DT$  blocks, true changes become more distinct and noise is reduced. The LMM output further sharpens boundaries and produces cleaner masks. In addition, Fig. 9 visualizes level-2 DFFM features of three image pairs (also via PCA). Different land covers such as trees, bare land, and roads are clearly separated, indicating that the DINOv3 branch supplies rich and discriminative semantics beyond the target buildings.

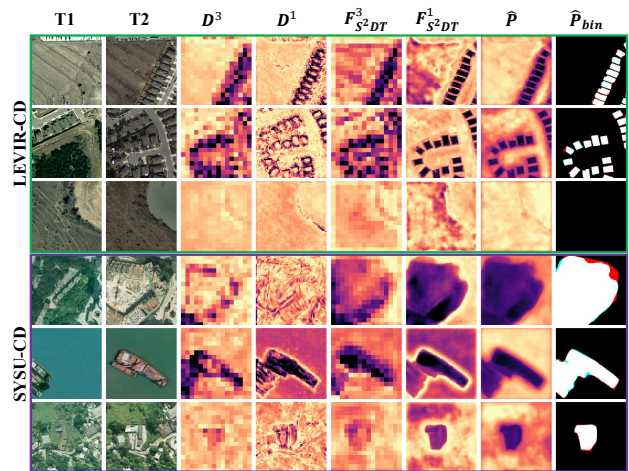


Figure 8. Visualized features of scenes from LEVIR-CD and SYSU-CD. Darker colors indicate stronger activation responses.

### 4.3 Ablation Study

To evaluate the contribution of each component in the proposed architecture, we conduct ablation experiments on the LEVIR-CD and SYSU-CD datasets. As shown in Tab. 2, removing the DFFM causes the performance to drop by about

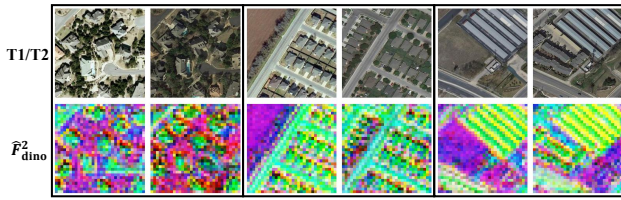


Figure 9. Visualized DFFM features (adapted DINOv3 features) from LEVIR-CD. We use vibrant colors for visualization.

Table 2. Ablation studies of the proposed components on the LEVIR-CD and SYSU-CD datasets. ✓ and ✗ denote "w/" and "w/o" the specific module.

Components			LEVIR-CD		SYSU-CD	
DFFM	S <sup>2</sup> DT	LMM	IoU	F1	IoU	F1
✗	✗	✗	84.23	91.44	70.06	82.40
✓	✓	✓	84.49	91.59	71.61	83.46
✓	✗	✓	84.98	91.88	72.87	84.30
✓	✓	✗	85.65	92.27	72.54	84.09
✓	✓	✓	<b>85.72</b>	<b>92.31</b>	<b>73.46</b>	<b>84.70</b>

1.23 and 1.85 points of IoU on LEVIR-CD and SYSU-CD, respectively. This suggests that DFFM is the most influential of the three modules, and confirms that distilling semantics from a large-scale pretrained model is particularly beneficial when training on relatively small RSCD datasets. For the S<sup>2</sup>DT decoder, the "w/o S<sup>2</sup>DT" variant is implemented by replacing it with a residual convolutional block. The observed degradation further indicates the effectiveness of combining spatial and spectral attention with the differential transformer design.

For the LMM, we find that enabling it improves results on both LEVIR-CD and SYSU-CD. It helps suppress small spurious responses while preserving building shapes, which is useful for both finely annotated data (LEVIR-CD) and coarser, large-area changes (SYSU-CD). Overall, the morphology-based refinement acts as an effective post-prediction regularizer across different annotation granularities. Taken together, these ablations demonstrate that each proposed component in **ChangeDINO** contributes to improving RSCD performance under different dataset characteristics.

## 5. Conclusion

We presented **ChangeDINO**, an end-to-end framework for optical building change detection that combines a Siamese backbone, DINOv3-based multi-scale features, a spatial–spectral differential transformer decoder, and a learnable morphology head. The DINOv3 branch improves semantic robustness, the S<sup>2</sup>DT decoder enhances change-focused reasoning, and the morphology module refines boundaries. Experiments on four public benchmarks show that ChangeDINO consistently outperforms recent CNN-, Transformer-, and foundation-model-based methods in IoU and F1, while ablation studies verify the contribution of each component. Future work will extend the framework to multi- and hyperspectral remote sensing data and related change analysis tasks.

## References

Asokan, A., Anitha, J., 2019. Change detection techniques for remote sensing applications: A survey. *Earth Science Informatics*, 12(2), 143–160.

Bandara, W. G. C., Patel, V. M., 2022. A transformer-based siamese network for change detection. *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 207–210.

Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering. *IEEE geoscience and remote sensing letters*, 6(4), 772–776.

Chen, H., Qi, Z., Shi, Z., 2021. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10), 1662.

Chen, H., Song, J., Han, C., Xia, J., Yokoya, N., 2024a. ChangeMamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–20.

Chen, J., Hou, D., He, C., Liu, Y., Guo, Y., Yang, B., 2024b. Change Detection With Cross-Domain Remote Sensing Images: A Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 11563–11582.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions.

Coppin, P. R., Bauer, M. E., 1996. Digital change detection in forest ecosystems with remote sensing imagery. *Remote sensing reviews*, 13(3-4), 207–234.

Dalla Mura, M., Benediktsson, J. A., Bovolo, F., Bruzzone, L., 2008. An Unsupervised Technique Based on Morphological Filters for Change Detection in Very High Resolution Images. *IEEE Geoscience and Remote Sensing Letters*, 5(3), 433–437.

Daudt, R. C., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. *2018 25th IEEE international conference on image processing (ICIP)*, IEEE, 4063–4067.

Ding, L., Hong, D., Zhao, M., Chen, H., Li, C., Deng, J., Yokoya, N., Bruzzone, L., Chanussot, J., 2025. A Survey of Sample-Efficient Deep Learning for Change Detection in Remote Sensing: Tasks, strategies, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 13(3), 164–189.

Dong, S., Wang, L., Du, B., Meng, X., 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208, 53–69.

Fang, S., Li, K., Shao, J., Li, Z., 2021. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.

Han, C., Wu, C., Guo, H., Hu, M., Li, J., Chen, H., 2023. Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 8395–8407.

He, P., Shi, W., Zhang, H., Hao, M., 2014. A novel dynamic threshold method for unsupervised change detection from remotely sensed images. *Remote sensing letters*, 5(4), 396–403.

- Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80, 91–106.
- Ji, S., Wei, S., Lu, M., 2019. 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.
- Jimenez-Sierra, D. A., Quintero-Olaya, D. A., Alvear-Munoz, J. C., Benitez-Restrepo, H. D., Florez-Ospina, J. F., Chanussot, J., 2022. Graph learning based on signal smoothness representation for homogeneous and heterogeneous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16.
- Jing, W., Chi, K., Li, Q., Wang, Q., 2025. ChangeRD: A registration-integrated change detection framework for unaligned remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220, 64–74.
- Johansen, K., Arroyo, L. A., Phinn, S., Witte, C., 2010. Comparison of geo-object based and pixel-based change detection of riparian environments using high spatial resolution multi-spectral imagery. *Photogrammetric Engineering & Remote Sensing*, 76(2), 123–136.
- Knyaz, V. A., Kniaz, V. V., Zheltov, S. Y., 2024. Improving change detection performance with generative-adversarial augmentation of dataset. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 265–271.
- Lebedev, M., Vizilter, Y. V., Vygolov, O., Knyaz, V. A., Rubis, A. Y., 2018. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 565–571.
- Li, Z., Tang, C., Liu, X., Zhang, W., Dou, J., Wang, L., Zomaya, A. Y., 2023. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–12.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal loss for dense object detection.
- Liu, Z., Li, G., Mercier, G., He, Y., Pan, Q., 2017. Change detection in heterogeneous remote sensing images via homogeneous pixel transformation. *IEEE Transactions on Image Processing*, 27(4), 1822–1834.
- Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization.
- Niu, X., Gong, M., Zhan, T., Yang, Y., 2018. A conditional adversarial network for change detection in heterogeneous images. *IEEE Geoscience and Remote Sensing Letters*, 16(1), 45–49.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library.
- Peng, D., Liu, X., Zhang, Y., Guan, H., Li, Y., Bruzzone, L., 2025. Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges. *International Journal of Applied Earth Observation and Geoinformation*, 136, 104282.
- Qi, Z., Yeh, A. G.-O., Li, X., Zhang, X., 2015. A three-component method for timely detection of land cover changes using polarimetric SAR images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107, 3–21.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2019. Mobilenetv2: Inverted residuals and linear bottlenecks.
- Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., Jiang, B., 2021. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24), 5094.
- Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L., 2022. 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.
- Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P., 2025. Dinov3.
- Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *International Workshop on Deep Learning in Medical Image Analysis*, Springer, 240–248.
- Wang, G., Cheng, G., Zhou, P., Han, J., 2023. Cross-level attentive feature aggregation for change detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7), 6051–6062.
- Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module.
- Xiong, F., Li, T., Yang, Y., Zhou, J., Lu, J., Qian, Y., 2024. Wavelet siamese network with semi-supervised domain adaptation for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*.
- Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F., 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.
- Zhang, B., Sennrich, R., 2019. *Root mean square layer normalization*. Curran Associates Inc., Red Hook, NY, USA.
- Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166, 183–200.
- Zhang, H., Chen, H., Zhou, C., Chen, K., Liu, C., Zou, Z., Shi, Z., 2024. Bifa: Remote sensing image change detection with bitemporal feature alignment. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–17.
- Zhang, H., Chen, K., Liu, C., Chen, H., Zou, Z., Shi, Z., 2025. CDMamba: Incorporating Local Clues Into Mamba for Remote Sensing Image Binary Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1–16.