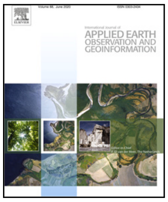
Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Robust change detection for remote sensing images based on temporospatial interactive attention module

Jinjiang Wei, Kaimin Sun, Wenzhuo Li *, Wangbin Li, Song Gao, Shunxia Miao, Qinhui Zhou, Junyi Liu

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430072, China

ARTICLE INFO

Dataset link: https://github.com/wjj282439449/CDNeXt

Keywords:
Change detection
Attention mechanism
Feature interaction
Remote sensing

ABSTRACT

Change Detection (CD) is a vital monitoring method in Earth observation, especially pertinent for land-use analysis, city management, and disaster damage assessment. However, in the era of constellation interconnection and air-sky collaboration, the changes in the Regions Of Interest (ROI) cause many false detections due to geometric perspective rotation and temporal style difference. In response to these challenges, we introduce CDNeXt, this framework elucidates a robust and efficient method for combining Siamese networks based on the pre-trained backbone with the innovative Temporospatial Interactive Attention Module (TIAM) for remote sensing imagery. The CDNeXt can be categorized into four primary components: Encoder, Interactor, Decoder, and Detector. Notably, the Interactor, powered by TIAM, queries and rebuilds spatial perspective dependencies and temporal style correlations from binary temporal features extracted by the Encoder to enlarge the difference of ROI change. Culminating the process, the Detector integrates the hierarchical features generated by the Decoder, subsequently producing a binary change mask. We have achieved new State-Of-The-Art (SOTA) performance in change detection, with our method surpassing existing techniques on four benchmark datasets: an F1 score of 82.63% on SYSU-CD, 87.14% on LEVIR-CD+, 66.71% on S2Looking, and 71.11% BANDON. To further validate the effectiveness of the TIAM, we compared it to other attention modules in both interactive and non-interactive modes. Our code is available on GitHub: https://github.com/wjj282439449/CDNeXt.

1. Introduction

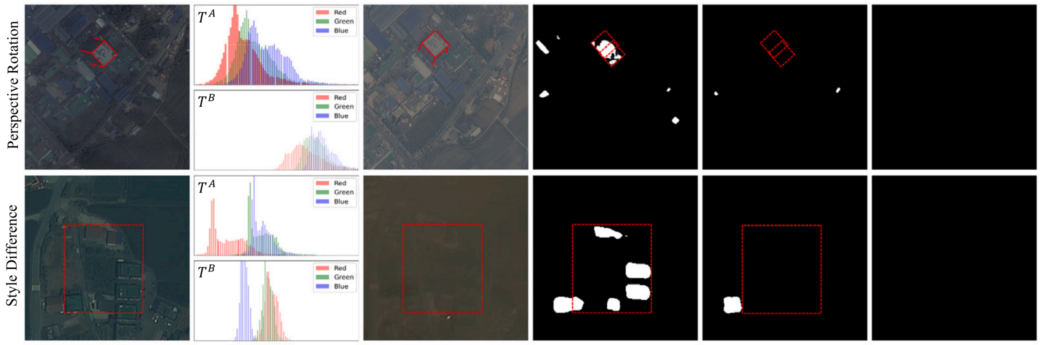
Change Detection (CD) is a pivotal technique in Earth observation, focusing on discerning differences within the Region Of Interest (ROI). The process entails segmenting dual temporal remote sensing images, consequently eliminating any unchanged and superfluous background elements (Pan et al., 2022). This task proves invaluable in monitoring terrestrial alterations at varying spatial resolution scales, spanning from vegetation cover and water body area to urban expansion and disaster damage assessment (Li et al., 2022a). Given the critical importance of overseeing ecological conservation and sociological activities (Li et al., 2023), the pursuit of fully automated, high-precision change detection algorithms becomes paramount.

In its essence, change detection seeks to segment neighboring local regions and compare their similarity in terms of texture and semantics. Here, similarity indicates non-change, while dissimilarity signals change. Traditional methods (Chen and Shi, 2020), which are predicated on spectral pixel comparisons and object-oriented analysis, harbor stringent stipulations regarding sensors, observational targets, and viewing angles. Therefore, this necessitates extensive pre-processing for these remote sensing images, encompassing geometric correction, radiometric correction, spatial alignment (Shen et al., 2021), cloud and shadow removal (Wei et al., 2019). In contrast to machine learning and object-oriented methods that have difficulty dealing with nonlinear and semantic change, utilizing pre-trained classification deep learning models as backbone networks has become a popular strategy, especially when harnessed in Siamese network architectures for CD (Feng et al., 2022). Currently, many methods have sought to improve performance by exploring how features interact within Siamese networks, including techniques such as feature concatenation, difference, and various attention mechanisms — such as channel (Daudt et al., 2018), spatial (Pan et al., 2022), or global attention (Chen et al., 2022a). Effective feature interaction mechanisms and their suitability for remote sensing change detection scenarios are currently at the forefront of research.

In the era of constellation interconnection and air-sky collaboration, the presence of differences in stereo perspective rotation and temporal

* Corresponding author.
E-mail addresses: weijinjiang@whu.edu.cn (J. Wei), sunkm@whu.edu.cn (K. Sun), alvinlee@whu.edu.cn (W. Li), lwb1996@whu.edu.cn (W. Li), gaosong@whu.edu.cn (S. Gao), shunxiamiao@whu.edu.cn (S. Miao), 2018302130135@whu.edu.cn (Q. Zhou), liujunyi_ljy@163.com (J. Liu).

https://doi.org/10.1016/j.jag.2024.103767
Received 21 December 2023; Received in revised form 28 February 2024; Accepted 10 March 2024
Available online 16 March 2024



(a) Image T^A (b) RGB Histogram (c) Image T^B (d) Others (e) Ours (f) Label

Fig. 1. Change detection results from our CDNeXt and other methods are presented in two rows. The extreme images highlight false alarms: the first row due to geometric perspective rotation and the second row due to temporal style difference.

style differences in images of the same region has become an increasingly significant issue. Stereo perspective rotation can create misleading interferences when the same area is viewed from different angles, leading to semantic misinterpretations due to apparent visual geometric changes. In Fig. 1's first row, differing viewing angles cause variations in the building's orientation, resulting in spatial misalignment and semantic differences within the combined roof and facade regions of the bi-temporal images. Such differences lead to false alarms in the CD process. Meanwhile, temporal style differences arise from variations in lighting, weather, and season across images taken at different times. These differences can result in pseudo-changes that intensify with the visual contrasts between the images, rather than actual ROI changes. In Fig. 1's last row, buildings partly veiled by a sandstorm are still partially visible, yet they prompt false detections by the algorithm.

These factors present a challenge for methods based on pre-trained siamese network models in extracting changes, as it becomes difficult to compare the semantic similarity of differ-temporal features under significant visual style and viewpoint differences. As demonstrated in Fig. 1, although histogram distributions exhibit similarities despite variations in viewing angles and lighting conditions, other factors such as weather and seasons lead to entirely distinct RGB distributions within the histograms. At present, mainstream methods address these challenges by not only constructing specific datasets to simulate various conditions during training, such as visual style (Chen and Shi, 2020; Zhang et al., 2020) and viewing angles (Shen et al., 2021; Pang et al., 2023) but also improving the CD framework (Daudt et al., 2018; Fang et al., 2022; Papadomanolaki et al., 2021) and feature attention mechanism (Shi et al., 2022; Chen et al., 2022a; Feng et al., 2022) to enhance the expression of change features. Despite these advancements, existing attention mechanisms often overlook the inherent self-similarity of features within the current temporal phase when applied to concatenated features from bi-temporal images.

Motivated by the abovementioned challenges, we introduce the CDNeXt equipped with the Temporospatial Interactive Attention Module (TIAM). It establishes a solid foundational CD framework for the TIAM module, specifically designed to mitigate interferences related to perspective rotation and temporal style. Our methodology is structured into four principal segments: Encoder, Interactor, Decoder, and Detector. Initially, the network leverages a pre-trained network as its backbone and extracts pyramid features from bi-temporal images while sharing weights across various depths. Thereafter, the Interactor, equipped with the TIAM, queries the global interactive relevance, such as spatial perspective dependencies and temporal style correlations, and reconstructs the unchanged semantic features between same-scale features in both time and space. Subsequently, the decoder, incorporating the Feature Squeeze Residual (FSR) block, upsamples the low-level features and skip-connects them with the Interactor features, squeezing the feature dimensions and extracting the residuals of change features. In the concluding phase, the detector, encompassing the FSR block, fuses the hierarchical features, spanning from low-level object contours and textures to high-level scene semantics, ultimately yielding a binary change detection mask. This streamlined, CD framework is both robust and efficient, foregoing the need for specialized training techniques or hyper-parameter fine-tuning, and can output change detection results end-to-end. This paper contributes the following:

- We propose the CDNeXt framework, a novel change detection paradigm tailored for remote sensing imagery. The framework features a four-tier architecture, namely Encoder, Interactor, Decoder, and Detector, capable of leveraging the pre-trained backbone synergizes effectively with the innovative TIAM.
- Introducing the Temporospatial Interactive Attention Module (TIAM) module, our easily embeddable attention mechanism addresses the challenges posed by geometric perspective rotation and temporal style difference in a singular computational process. The module is adept at querying and rebuilding spatial perspective dependencies and temporal style correlations.
- We have achieved unparalleled performance in the domain of change detection on four renowned datasets: SYSU-CD, LEVIR-CD+, S2Looking, and BANDON. Our approach surpasses nine different papers, demonstrating superior robustness and a significant reduction in false alarms. The code for our method is publicly available on GitHub.

The remainder of this paper is organized as follows: Section 2 reviews related work in the field. Section 3 presents our proposed methodology. Section 4 evaluates and discusses the effectiveness of our approach through extensive experiments. Section 5 summarizes our work and gives a brief discussion of future work.

2. Related work

2.1. *Robust change detection*

Most change detection networks perform feature extraction and segmentation result output through encoder–decoder structures (Fang et al., 2022). The encoder usually uses weight-shared pre-trained backbone networks to extract image features from different periods (Basavaraju et al., 2022). Improvements in method robustness can be achieved through various techniques, such as joint multi-task learning (Gao et al., 2022), additional training constraint (Shi et al., 2022), and deep feature supervision (Pan et al., 2022). The Image Fusion Network (IFNet) (Zhang et al., 2020) supervises the output fusion of each decoder layer to obtain segmentation results. Moreover, Shi et al. (2022) supervised attention metrics based on the Siamese structure and proposed targeted datasets to mitigate pseudo-change effects. Wang et al. (2023) proposed MSFF-CDNet, a multiscale feature fusion CD

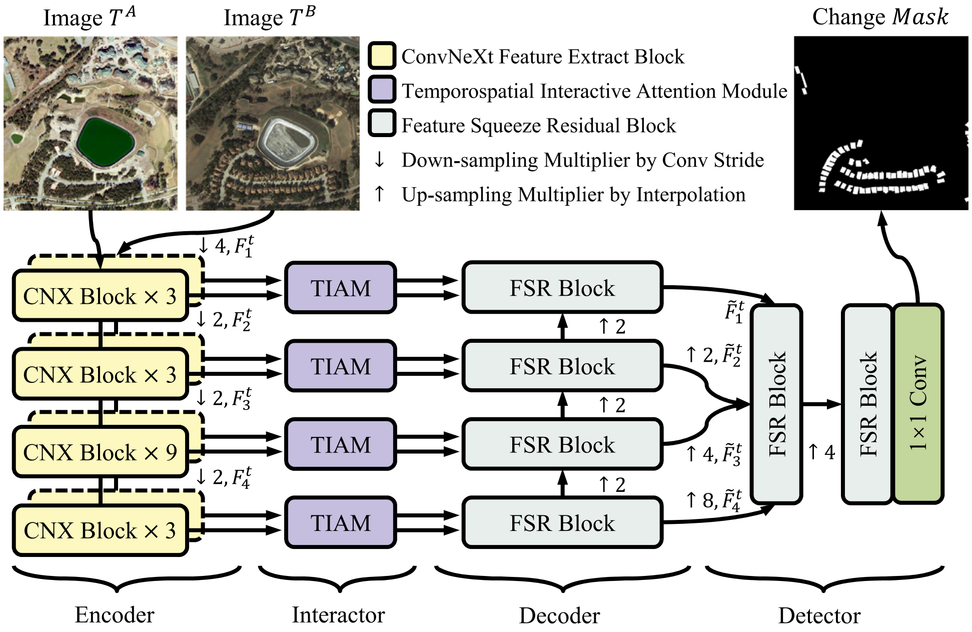


Fig. 2. The schematic illustration of our proposed CDNeXt. This framework with four sub-modules: Encoder, Interactor, decoder, and detector.

network that addresses the issue of insufficient extraction and utilization of deep features, leading to unstable performance in capturing multiscale changes. Lin et al. (2023) proposed P2V-CD, a pair-to-video CD framework that explicitly models time and addresses the issues of incomplete temporal modeling and space–time coupling in deep learning-based approaches. Huang et al. (2024) proposed SEIFNet to reduce pseudo changes and scale variations by effectively capturing global and local information, integrating interlevel features, and enhancing boundary details and internal integrity. Current Siamese CD frameworks lack systematized structures, exhibit poor robustness when facing interfering factors, and do not sufficiently utilize advanced structures.

2.2. Attention in change detection

The attention method is a popular feature enhancement mechanism in change detection, such as the early channel and spatial attention (Zhang et al., 2022), global attention (Chen et al., 2022b; Li et al., 2022b), and the Transformer with multi-head attention (Chen et al., 2022a), but interactive attention between features is less discussed (Song et al., 2022). The attention mechanism originated from the field of natural language processing but is not exclusive to it (Vaswani et al., 2017). Woo et al. (2018) proposed the Convolutional Block Attention Module (CBAM) to enhance channel and spatial information sequentially beyond the SENet (Hu et al., 2020). Wang et al. (2018) propose a global attentive module that is easily embeddable for computer vision tasks. Fu et al. (2019) performed a weighted summation of spatial and channel non-local for self-attentive features. Feature comparison tasks using the correlation between feature interactive attention are widespread, such as feature matching (Sun et al., 2021). But Current attention mechanisms for CD tasks, need to ensure features fully interact in the temporospatial domain, but they exhibit computational redundancies and intra-temporal similarities. In the field of CD, Chen et al. (2022b) both used non-local attention and fused different levels of attention features to refine the prediction results. Differently, Feng et al. (2022) proposed a network that utilizes Transformer and Convolutional Neural Network (CNN) extracted features for self- and inter-attention features aggregation and alignment. However, earlier methods in feature interaction overlooked the integration of temporal–spatial dimensions and the impact of intra-temporal self-similarity on viewing angles and visual style issues.

3. Methodology

3.1. Overview network architecture

We propose CDNeXt, a systematized framework tailored for bi-temporal change detection. Harnessing the strength of the TIAM and the pre-trained backbone network, CDNeXt provides a comprehensive framework adept at countering multi-interfering factors. For given matched image pairs $\{T^A, T^B\} \in \mathbb{R}^{C\times H\times W}$, the CDNeXt processes these through its robust pipeline, culminating in the extraction of a change $Mask$. The CD process can be concisely represented as:

$$Mask = \text{CDNeXt}(T^A, T^B) \tag{1}$$

This framework, depicted in Fig. 2, consists of four primary modules: Encoder, Interactor, Decoder, and Detector.

Encoder: It leverages a backbone network to derive a set of hierarchical features, $\{F_1, \dots, F_n\}$, via top-down downsampling. It can be described as:

$$\{F_1^t, \dots, F_n^t\} = \text{Encoder}(T^t),\ t \in \{A, B\} \tag{2}$$

Interactor: This module queries and rebuilds the spatial perspective dependencies and temporal style difference in same-level feature pairs F_l^A and F_l^B. The lth layer features have the corresponding lth $\text{Encoder}_{(l)}$ parameters. It can be described as:

$$\{\hat{F}_l^A, \hat{F}_l^B\} = \text{Interactor}_{(l)}(F_l^A, F_l^B), l \in \{1, \dots, n\} \tag{3}$$

Decoder: It ingests the lth Interactor's outputs $\{\hat{F}_l^A, \hat{F}_l^B\}$ and concatenates them with interpolated features from the preceding $\text{Decoder}_{(l+1)}$ output. It can be described as:

$$\tilde{F}_l = \begin{cases} \text{Decoder}_{(l)}(\hat{F}_l^A, \hat{F}_l^B, F_{l+1}^D) & l \in \{1, \dots, n-1\} \\ \text{Decoder}_{(l)}(\hat{F}_l^A, \hat{F}_l^B) & l = n \end{cases} \tag{4}$$

Detector: It consolidates the decoded hierarchical features$\{\tilde{F}_1, \dots, \tilde{F}_l\}$ to generate the change $Mask$, as follows:.

$$Mask = \text{Detector}(\tilde{F}_1, \dots, \tilde{F}_l) \tag{5}$$

The essence of CDNeXt lies in the Interactor, which helps the parameters θ^* to approximate the $Label$. The model's optimization relies on

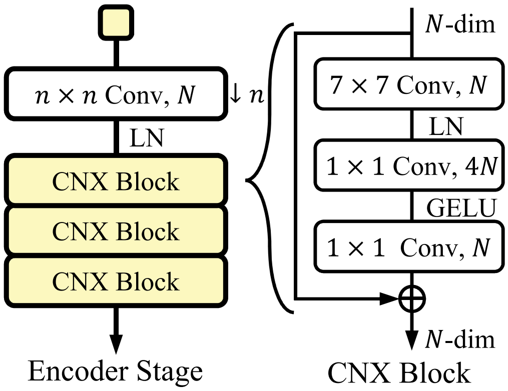


Fig. 3. The details of the CNX block in the Encoder.

the Cross-Entropy (CE) loss ($\mathcal{L}_{CE}$) for the labels and predictions at each pixel position, denoted as $\hat{y}_n$ and y_n, respectively. We define training optimization as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log(\hat{y}_n) + (1 - y_n)\log(1 - \hat{y}_n)] \tag{6}$$

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{CE}(\text{CDNeXt}(T^A, T^B, \theta),\ Label) \tag{7}$$

3.2. Encoder based on CNX block

The encoding process from the top-down approach is effective in capturing the local features and alleviating the complexity of fitting the pre-trained backbone network. Thus, the Siamese structure is an exquisite design for the CD task (Daudt et al., 2018). In our framework, the Encoder uses a hierarchical architecture with a pre-trained ConvNeXt (CNX) network to extract multi-level features $\{F_1, F_2, F_3, F_4\}$ from different temporal image pairs $\{T^A, T^B\} \in \mathbb{R}^{C\times H\times W}$, such as defined Eq. (2).

As shown in Fig. 3, Each level feature from the backbone must include a down-sampling layer and multi-base blocks. For example, the ConvNeXt (Liu et al., 2022) down-sample multiplier is $n \in \{4, 2, 2, 2\}$. We can expand the down-sampling as follows:

$$F_l' = \text{LN}(\text{Conv}^{(\text{n,n})})(F_l),\ \ l \in \{1, 2, 3, 4\} \tag{8}$$

where $F_l' \in \mathbb{R}^{C\times\frac{H}{n}\times\frac{W}{n}}$, Conv superscript "$^{(n,n)}$" first position is the number of convolution (Conv) core size, and second is the number of the Conv stride. LN denotes Layer Normalization (LN). Then, CNX block is computed as:

$$\text{CNX}(F) = \text{Conv}(\text{GELU}(\text{Conv}(\text{LN}(\text{Conv}^{(7)}(F))))) \oplus F \tag{9}$$

$$F_{l+1} = \begin{cases} \text{CNX}_{(m)}(F_l') & m = 1 \\ \text{CNX}_{(m)}(\text{CNX}_{(m-1)}) & m \in \{2, \dots, M\} \end{cases} \tag{10}$$

where m represents the mth repetition of CNX block computation within each stage. The input features at the F_{l+1} stage result from downsampling the features of F_l' and undergoing $M \in \{3, 3, 9, 3\}$ iterations of CNX block computations. Superscript $^{(7)}$ denotes 7×7 Conv, without superscript denotes 1×1 Conv, "$\oplus$" denotes matrix element plus. These two 1×1 Conv expand the channel 4 times and then recover. The GELU (Hendrycks and Gimpel, 2017), a smoother variant than ReLU, is utilized in the most advanced Transformers.

3.3. Temporospatial interactive attention module

The correlation of feature representations across perspectives and styles is crucial for change detection in the presence of factors such as geometric perspective rotation and temporal style difference. As illustrated in Fig. 4, we propose a Temporospatial Interactive Attention Module (TIAM) as a key solution to mitigate these interfering factors.

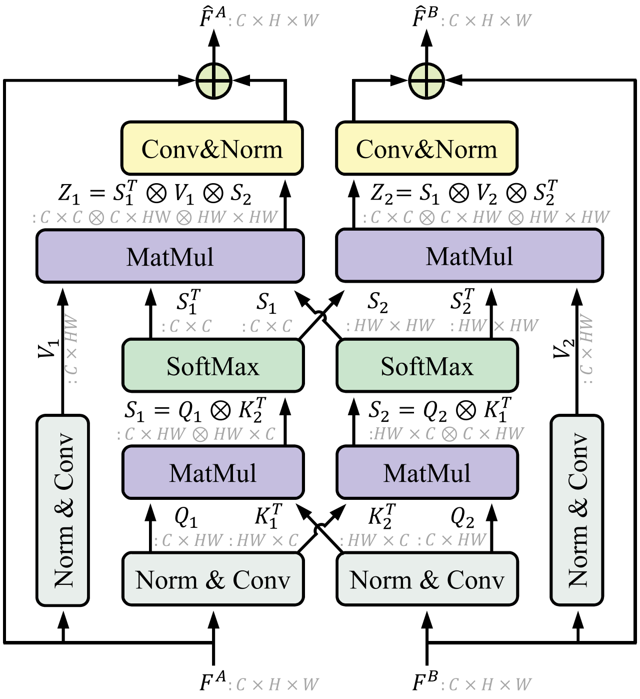

$\hat{F}^A$: $C \times H \times W$
$\hat{F}^B$: $C \times H \times W$
Conv&Norm
Conv&Norm
$Z_1 = S_1^T \otimes V_1 \otimes S_2$
: $C \times C \otimes C \times HW \otimes HW \times HW$
$Z_2 = S_1 \otimes V_2 \otimes S_2^T$
: $C \times C \otimes C \times HW \otimes HW \times HW$
MatMul
MatMul
S_1^T
: $C \times C$
S_1
: $C \times C$
S_2
: $HW \times HW$
S_2^T
: $HW \times HW$
V_1
: $C \times HW$
V_2
: $C \times HW$
SoftMax
SoftMax
$S_1 = Q_1 \otimes K_2^T$
: $C \times HW \otimes HW \times C$
$S_2 = Q_2 \otimes K_1^T$
: $HW \times C \otimes C \times HW$
MatMul
MatMul
Q_1
K_1^T
: $C \times HW$
: $HW \times C$
K_2^T
Q_2
: $HW \times C$
: $C \times HW$
Norm & Conv
Norm & Conv
Norm & Conv
Norm & Conv
F^A: $C \times H \times W$
F^B: $C \times H \times W$

Fig. 4. The details of the Temporospatial Interactive Attentive Module (TIAM). The bi-features are shown as the shape of their tensor, "F^t: $C_i{\times}H_i{\times}W_i$" denotes tth ($t \in \{A, B\}$) time and subscript i denotes different feature or dimension. The gray font represents the process of dimensionality changes in the feature matrix.

The theoretical basis of the TIAM is constructing Gram matrices between two temporal features to query the temporal and spatial attention scores. These attention scores respectively represent the spatial perspective dependencies S_2 and the temporal style correlations S_1 of the two temporal features. The queries in spatial perspective dependencies ensure the modeling of semantic invariance in the target, reducing false alarms. Similarly, the queries in temporal style correlations characterize the visual style distances between different temporal features, reducing differences caused by invariant and background features. This module can be embedded into visual tasks requiring the representation of similarity between two features.

Modeling the TIAM. As defined in Eq. (3), given each pair of features $\{F^A, F^B\} \in \mathbb{R}^{C\times H\times W}$ by the backbone of each level. First, We restructure the feature shape from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{C\times HW}$, followed by the application of Batch Normalization (BN) and 1×1 convolution (Conv$^{(1)}$). This sequence avoids gradient explosion from the backbone. V_1 and V_2 share parameters, while Q_1 and Q_2 do not, as follows:

$$\begin{cases} V_i = \text{Conv}^{(1)}(\text{BN}(F^t)) \\ Q_i = \text{Conv}^{(1)}_{(i)}(\text{BN}_{(i)}(F^t)) \\ K_i = Q_i^T \end{cases} \tag{11}$$

where $(i, t) \in \{(1, A), (2, B)\}$, subscript i denotes different feature and came from. Then, the time- and space-relevance interactive attention by Embedded Gaussian function (Wang et al., 2018) as follows:

$$S_i = \text{Softmax}(Q_i \otimes K_j^T) \tag{12}$$

where $(i, j) \in \{(1, 2), (2, 1)\}$, the dimension close to V_k will be Softmax in Eq. (13), "$\otimes$" denotes matrix multiplication. S_1 represents the similarity between each channel of the current temporal feature and each channel of another temporal feature, representing temporal style correlations. S_2 represents the similarity between each spatial position

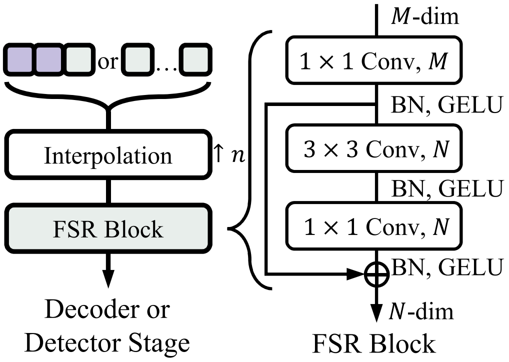


Fig. 5. The details of the FSR block in the decoder and detector.

of the current temporal feature and each spatial position of another temporal feature, representing spatial perspective dependencies.

Then, the matrix of the temporal style correlations S_1 and the matrix of the spatial perspective dependencies S_2 to weighted rebuild the V_i in time and space as follows:

$$\begin{cases} Z_1 = S_1^T \otimes V_k \otimes S_2 \\ Z_2 = S_1 \otimes V_k \otimes S_2^T \end{cases} \tag{13}$$

where one of $\{S_1, S_2\}$ will be transpose according to different Z_k. Lastly, Z_k will recover the same size with F through Conv and BN as follows:

$$\hat{F}^t = \mathrm{BN}(\mathrm{Conv}^{(1)}(Z_k)) \oplus F^t \tag{14}$$

where $(k,t) \in \{(1,A),(2,B)\}$ is the same as Eq. (11), and "⊕" denotes element-wise sum. "$\hat{F}^t$" denotes the residual sum by the rebuilt result "Z_k" and input "F^t", facilitating the network's training.

Comparing mainstream attention mechanisms. We abstract and simplify the core concepts of traditional attention mechanisms into three distinct modules: Self-Attention Module (SAM) (Wang et al., 2018), Dual-Attention Module (DAM) (Fu et al., 2019), and Cross-Attention Module (CAM) (Sun et al., 2021), as shown in the equations below. By comparing them, we demonstrate that our module represents a novel attention mechanism designed specifically to address perspective and style issues in change detection tasks.

SAM:

$$\mathrm{SAM}(K,Q,V) = Q \otimes K^T \otimes V \tag{15}$$

DAM:

$$\mathrm{DAM}(K,Q,V) = Q \otimes K^T \otimes V + Q^T \otimes K \otimes V \tag{16}$$

CAM:

$$\mathrm{CAM}(K,Q,V) = Q_i \otimes K_j^T \otimes V_i \tag{17}$$

Our TIAM:

$$\mathrm{TIAM}(K,Q,V) = (Q_i \otimes K_j^T) \otimes V_i \otimes (Q_i^T \otimes K_j) \tag{18}$$

where $(i,j) \in \{(1,2),(2,1)\}$ denotes two temporal features.

3.4. Decoder and detector

We have designed a feature Decoder and Detector that adeptly compresses the feature dimensions of multi-level representations and integrates the detection outcomes derived from the hierarchical feature outputs. This architecture is proficient in fusing local textures with high-dimensional semantics. Therefore, each Decoder layer obtains temporospatial rebuilt features with discriminable representation from the Interactor and up-sample feature from the last Feature Squeeze Residual (FSR) block, as shown in Fig. 5. This module also establishes a skip connection with the Interactor, thereby bolstering gradient propagation.

Decoder based on FSR block. The FSR block is a distinctive module that encompasses two primary components: the FSRSqueeze (elaborated in Eq. (21)) and the FSRResidual (outlined in Eq. (22)). Both components are constructed upon the FSRBase, as depicted in Eq. (20), but they handle different input feature dimensions. The FSRSqueeze is specifically designed to compress feature dimensions, while the FSRResidual leverages a residual structure, playing a pivotal role in bolstering feature learning and mitigating the risks of overfitting, as captured in Eq. (19). The definitions for the FSR and the FSRBase are articulated in the subsequent mathematical formulations:

$$\mathrm{FSR}(\dots) = \mathrm{FSRResidual}(\mathrm{FSRSqueeze}(\dots)) \tag{19}$$

$$\mathrm{FSRBase}(\dots) = \mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}(\dots))) \tag{20}$$

The FSRSqueeze and the FSRResidual can be defined as follows:

$$\hat{F}_l = \mathrm{FSRSqueeze}(\dots) = \begin{cases} \mathrm{FSRBase}_{(l)}(\hat{F}_l^A, \hat{F}_l^B, F_{l+1}^D) & l \in \{1,\dots,n-1\} \\ \mathrm{FSRBase}_{(l)}(\hat{F}_l^A, \hat{F}_l^B) & l = n \end{cases} \tag{21}$$

$$\tilde{F}_l = \mathrm{FSRResidual}(\dots) = \mathrm{FSRBase}_{(l)}(\mathrm{FSRBase}_{(l)}(\hat{F}_l)) \oplus \hat{F}_l \tag{22}$$

where Eq. (21) using 1 × 1 Conv and squeeze channel. Eq. (22) using 3 × 3 and 1 × 1 Conv, sequentially.

Detector with hierarchical fusion. The unique aspect of the Detector lies in its operation of the FSR Squeeze, which receives $\tilde{F}_1, \dots, \tilde{F}_4$ as inputs, as delineated in Eq. (23). As depicted in Fig. 2 the feature interpolate to input size and do Eq. (19) twice:

$$\hat{F} = \mathrm{FSR}(\tilde{F}_1, \dots, \tilde{F}_4) \tag{23}$$

where $\tilde{F}$ denotes different level layers from the Decoder. Lastly, 1 × 1 Conv to output the change $Mask$, where Softmax works in each space position:

$$Mask = Softmax(\mathrm{Conv}(\mathrm{FSR}(\hat{F}))) \tag{24}$$

4. Experiments and analysis

4.1. Datasets and evaluation metrics

We assess the performance of our proposed methods using four widely-used datasets: SYSU-CD (Shi et al., 2022), LEVIR-CD+ (Chen and Shi, 2020), S2Looking (Shen et al., 2021), and BANDON (Pang et al., 2023). We divided the aforementioned datasets into two categories: temporal style datasets (SYSU-CD, LEVIR-CD+) and perspective rotation datasets (BANDON, S2Looking). This division was conducted to validate the effective improvements of CDNeXt in addressing two specific issues.

SYSU-CD. This dataset comprises 20 000 pairs of aerial image patches measuring 256 × 256 with a resolution of 0.5 m. The ratio between the training and test sets is 6:2. It offers diverse change types in complex scenarios, including urban, vegetation, road, and sea construction.

LEVIR-CD+. This CD dataset contains more than 985 pairs (0.5 m/pixel) of bitemporal building images. The ratio between the training and test sets is 2:1., and a valid set was created by randomly selecting 25% of the training set. Each image has dimensions of 1024 pixels in width and height.

BANDON. This dataset comprises a vast collection of off-nadir aerial images with a resolution of 0.6 m, spanning six major cities in China. The collection has been partitioned into train (1689 images), valid (202 images), and test (392 images), each of 2048 × 2048 pixels in dimensions. This dataset captures building offsets, matching, and change labels.

Table 1
Quantitative assessment of CDNeXt's performance relative to other methods, analyzed using P(%), R(%), F1(%), and IoU(%) metrics across the SYSU-CD and LEVIR-CD+ datasets. Color convention: best, 2nd-best, and **3rd-best** for change detection models. The CDNeXt-Base without the TIAM.

Method	SYSU-CD				LEVIR-CD+			
	P	R	F1	IoU	P	R	F1	IoU
FC-EF	72.28	72.85	72.57	56.94	70.60	73.64	72.09	56.36
FC-Diff	84.94	51.45	64.08	47.15	79.94	72.55	76.06	61.37
FC-Conc	83.03	71.62	76.89	62.45	78.07	68.22	72.82	57.25
STANet	70.76	**85.33**	77.37	63.09	69.74	**83.92**	76.17	61.52
IFNet	**87.33**	70.17	77.82	63.69	86.01	82.09	**84.08**	**72.42**
DSAMNet	74.81	**81.86**	78.18	64.18	60.68	**89.04**	72.18	56.47
SNUNet	83.49	76.37	79.77	66.35	85.32	80.18	82.67	70.46
L-UNet	80.09	**80.27**	80.18	66.91	83.31	79.45	81.34	68.55
BIT	83.13	73.67	78.12	64.09	84.29	83.19	83.84	72.02
ICIF-Net	83.37	78.51	**80.74**	**68.12**	**87.79**	80.88	83.65	71.89
P2V	80.38	76.82	78.56	64.69	78.00	80.94	79.44	65.90
CDNeXt-Base	**88.70**	75.39	**81.51**	**68.78**	**88.80**	82.04	**85.29**	**74.35**
CDNeXt	**89.72**	76.57	**82.63**	**70.39**	**89.68**	**84.73**	**87.14**	**77.21**

Table 2
Quantitative assessment of CDNeXt's performance relative to other methods, analyzed using P(%), R(%), F1(%), and IoU(%) metrics across the BANDON and S2Looking datasets.

Method	BANDON				S2Looking			
	P	R	F1	IoU	P	R	F1	IoU
FC-EF	**73.89**	36.13	48.53	32.04	62.55	42.47	50.59	33.86
FC-Diff	72.83	56.80	63.82	46.87	61.64	61.25	61.45	44.35
FC-Conc	**73.71**	55.39	63.25	46.25	68.67	53.88	60.38	43.24
STANet	35.37	**88.50**	50.54	33.82	18.06	**88.60**	30.00	17.65
IFNet	64.28	54.93	59.24	42.08	68.81	46.00	56.02	38.07
DSAMNet	35.06	**88.48**	50.23	33.53	16.85	**89.74**	28.37	16.53
SNUNet	65.95	61.96	63.90	46.95	69.54	53.82	60.68	43.55
L-UNet	73.41	58.21	64.93	48.07	**72.41**	51.85	60.43	43.30
BIT	71.95	64.25	67.88	51.38	68.84	57.59	62.71	45.68
ICIF-Net	72.77	61.72	66.79	50.14	**72.67**	50.77	59.78	42.63
P2V	72.46	56.18	63.29	46.29	68.74	51.61	58.96	41.80
FC-Conc+TIAM	70.78	59.97	64.93	48.07	71.76	52.78	60.82	43.70
IFNet+TIAM	71.83	66.63	**69.13**	**52.83**	**72.63**	58.89	**65.04**	**48.19**
CDNeXt-Base	70.05	**69.99**	**70.02**	**53.87**	71.20	60.08	**65.17**	**48.33**
CDNeXt	**74.46**	68.06	**71.11**	**55.18**	70.78	**63.08**	**66.71**	**50.05**

S2Looking. This dataset is a collection of side-looking satellite images captured over 3 years. It comprises 5000 pairs of high-resolution images with 1024 × 1024 pixels. This dataset is split into three parts: train (3500 images), valid (500 images), and test (1000 images). It includes large perspective changes and illumination variances in building changes.

Evaluation metric. Following recent works (Pang et al., 2023) on bi-temporal CD, we compare all methods and ablation studies using four metrics, where higher values indicate better performance, namely Precision (P), Recall (R), F1-measure score (F1), and Intersection over Union (IoU). These metrics consider the detection quality of change regions to evaluate the CD performance quantitatively. The F1 score combines precision and recall, providing a comprehensive evaluation. Moreover, the IoU metric quantifies the pixel-level accuracy between predicted and labeled regions.

4.2. Implementation details

We trained the CDNeXt and mainstream methods on the SYSU-CD, LEVIR-CD+, S2Looking, and BANDON datasets. For testing, we selected models that were fitted on the validation set. The images of each dataset are cropped without overlap to 256 × 256 pixel size and kept the original training set and test set divisions. For the CDNeXt model, the compact version of the backbone network initializes its parameters through pre-training weights. During the model optimization process, we employed the AdamW (Loshchilov and Hutter, 2019) optimizer with a fixed learning rate of 4×10^{-5} and no decay strategy. We use the gradient accumulation to avoid 32 batch sizes resulting in insufficient memory on a single 4090 and to ensure that the batch size is consistent across ablation experiments. During model training, image pairs and labels are randomly augmented using photometric and geometric distortions. We implemented our model with PyTorch, while the others were obtained from the official implementation.

4.3. Comparison on temporal style datasets

We compare our model CDNeXt with eleven methods (from nine papers) for change detection on temporal style datasets: FC-EF (Daudt et al., 2018), FC-Diff (Daudt et al., 2018), FC-Conc (Daudt et al., 2018), STANet (Chen and Shi, 2020), IFNet (Zhang et al., 2020), DSAMNet (Shi et al., 2022), SNUNet (Fang et al., 2022), L-UNet (Papadomanolaki et al., 2021), BIT (Chen et al., 2022a), ICIF-Net (Feng et al., 2022) and P2V (Lin et al., 2023). And visual comparisons of FC-Diff, IFNet, SNUNet, L-UNet, BIT, and ICIF-Net.

Quantitative comparisons. Table 1 shows the superiority of our CDNeXt on bi-temporal change detection. The CD performance on the visual style dataset achieves current SOTA results in all aspects except for the recall rate. And we can see outperforms for comprehensive evaluation metrics F1 and IoU. The CDNeXt improvements achieve 1.89% F1 and 2.27% IoU in the SYSU-CD dataset, 3.06% F1 and 4.79% IoU in the LEVIR-CD+ dataset, compared with mainstream methods. Our approach obtains more significant accuracy gains on temporal style datasets.

Qualitative comparisons. We visually compare the change mask produced by our method and SOTA methods on the SYSU-CD and LEVIR-CD+ in the context of visual style problems. It is worth noting that other models trained on visual style datasets still yield false alarms due to the issue of perspective rotation, but our CDNeXt does not. The visual results of Fig. 6 illustrate the general change detection performance across different types of objects in the SYSU-CD dataset. Although this dataset does not address viewpoint rotation issues, we can still observe the prevalence of viewpoint problems in change detection tasks (from line three to five). Furthermore, changes in lighting and shadow conditions and vegetation phenology can lead to incorrect representations of changes by mainstream methods, resulting in decreased boundary accuracy and false alarm issues (from line six to last). Regarding the CD visualizations in Fig. 7, our method exhibits fewer false alarms and improved boundary accuracy. This is primarily due to significant visual disparities between the two temporal images of the dataset, including differences in ground resolution, lighting conditions, and shadows (except for the first line). In comparison to BIT and ICIF-Net, which also incorporate spatial interactive attention mechanisms, this demonstrates the effectiveness of TIAM's temporal interactivity in query and reconstruction.

4.4. Comparison on perspective rotation datasets

We compare our model CDNeXt with nine methods for change detection on perspective rotation datasets. And visualization comparison of FC-Diff, IFNet, SNUNet, L-UNet, BIT, ICIF-Net, and P2V. We also compare the accuracy of FC-Conc and IFNet with the addition of the TIAM module.

Quantitative comparisons. Table 2 presents the comparative performance of CDNeXt and TIAM on two viewpoint rotation datasets in terms of CD performance. Our method not only achieves SOTA performance but also improves the accuracy to varying degrees when other mainstream methods (FC-Conc+TIAM and IFNet+TIAM) incorporate our TIAM, indicating the effectiveness of the TIAM. Our proposed method outperforms the SOTA approach, BIT, in terms of accuracy by 3.23% on the BANDON dataset and 4.00% on the S2Looking dataset.

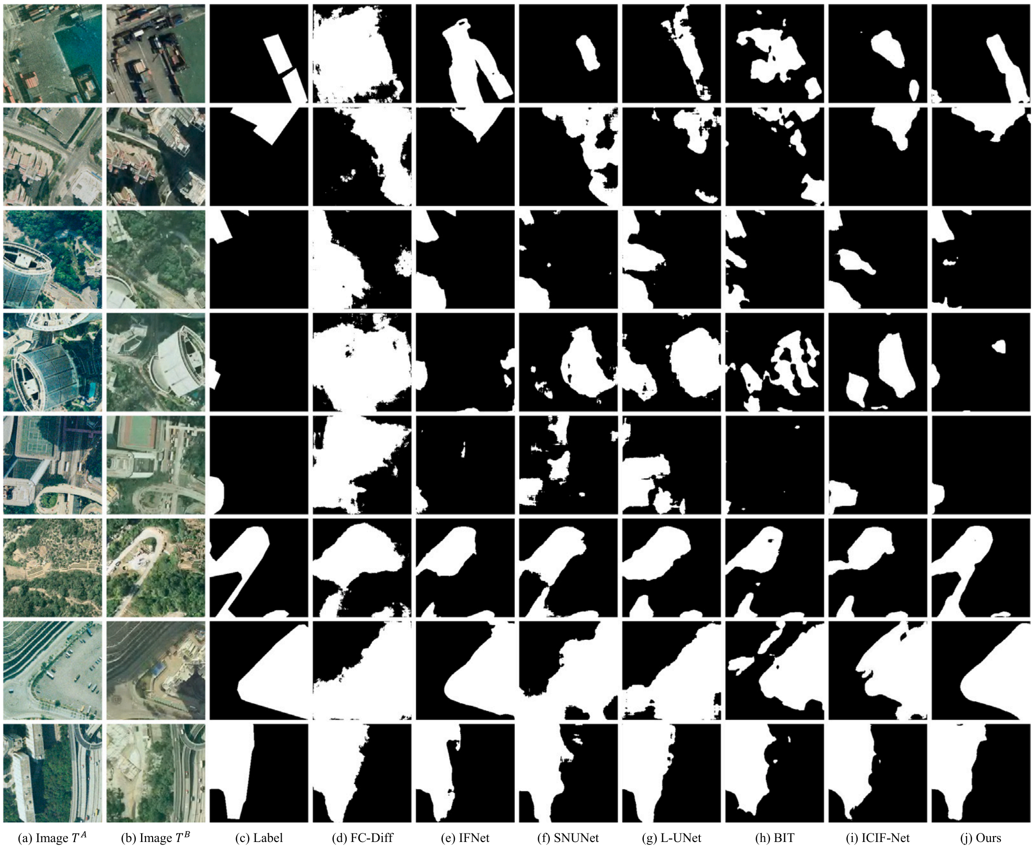


Fig. 6. Visualization results of different methods for SYSU-CD dataset.

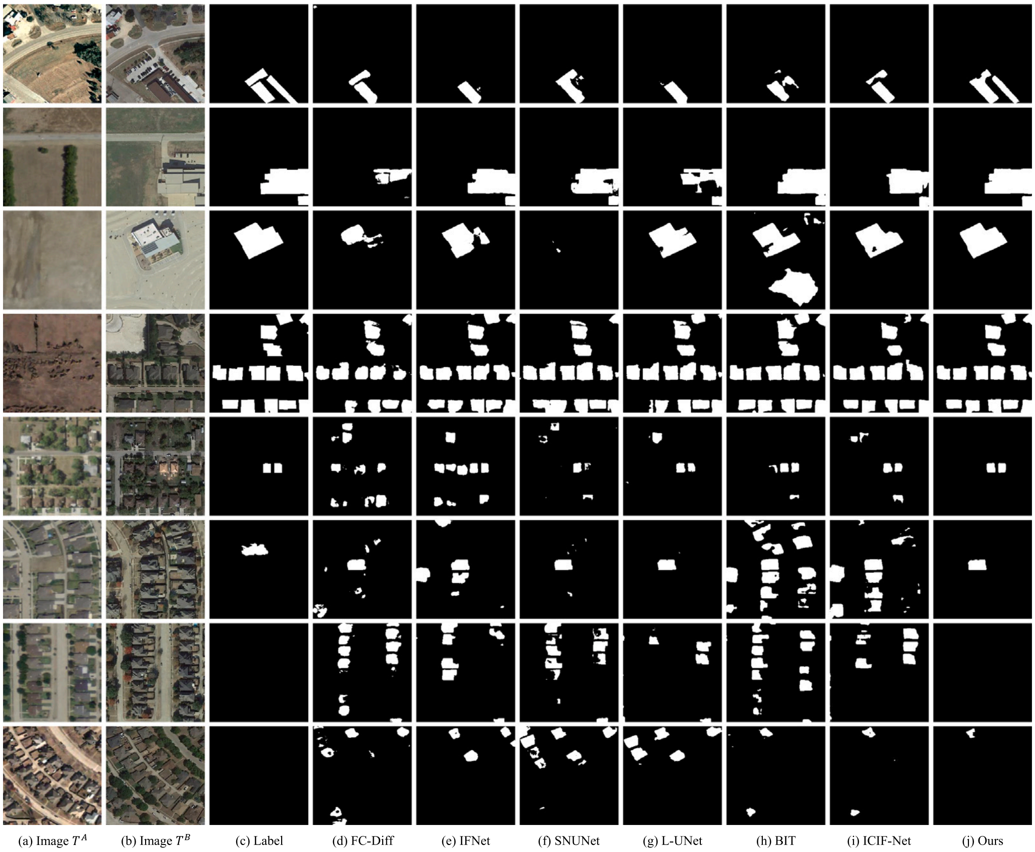


Fig. 7. Visualization results of different methods for LEVIR-CD+ dataset.

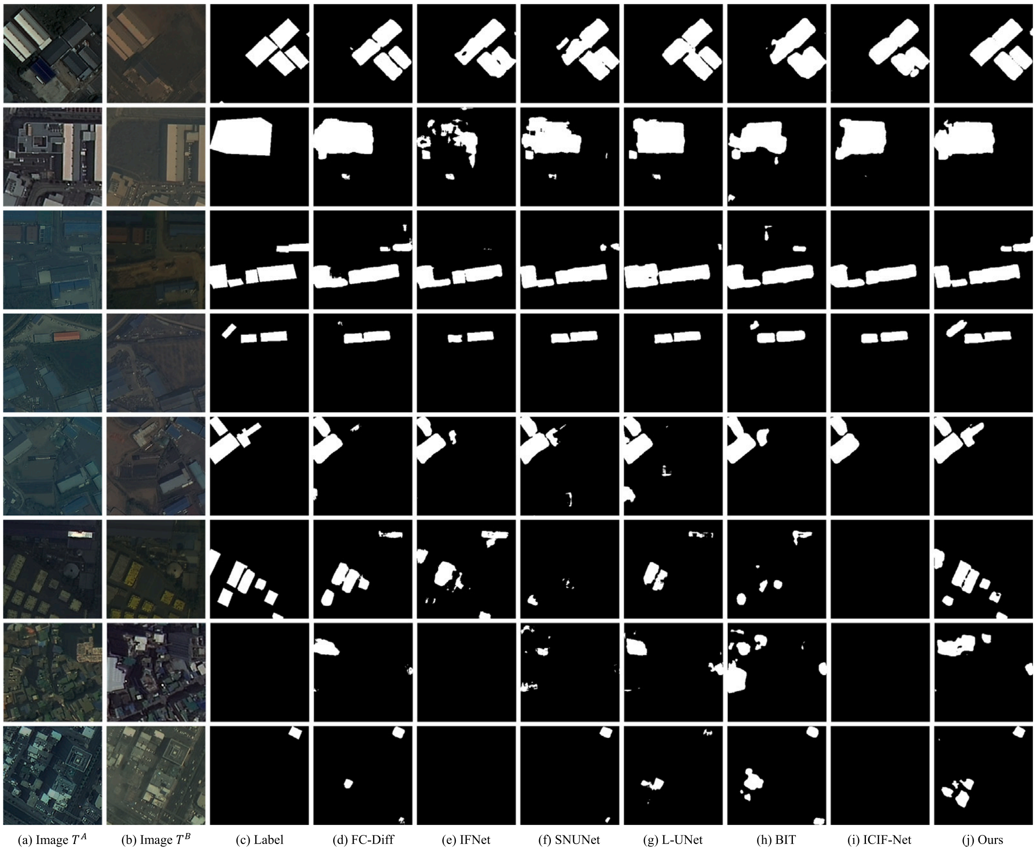

(a) Image T^A (b) Image T^B (c) Label (d) FC-Diff (e) IFNet (f) SNUNet (g) L-UNet (h) BIT (i) ICIF-Net (j) Ours

Fig. 8. Visualization results of different methods for S2Looking dataset.

While STANet and DSAMNet exhibit high recall accuracy, CDNeXt maintains an overall superior performance.

Qualitative comparisons. We visually compare the change mask produced by our method and SOTA methods on the S2Looking dataset in Fig. 8. It is noteworthy that other models trained on viewpoint rotation datasets resulted in false alarms and missed detections due to visual style problems, but our CDNeXt does not. In Fig. 8, our method exhibits lower missed detections and demonstrates precise boundary prediction for change targets (first four rows). Additionally, the S2Looking dataset inherently possesses significant visual style differences between the two periods, which result in missed detections by other mainstream methods. However, CDNeXt effectively mitigates this interference and provides stable change detection results. In urban areas with tall buildings, our method generates fewer false alarms (last four rows). For changes with missing annotations(the last two rows), our method also understands the semantics of the change object from different viewpoints and with fewer false detections.

Generalization ability validation. We visualize the large-scale scenarios in the in-domain test set of the BANDON dataset to validate the generalization capability of CDNeXt. The in-domain test set is from the same city as the training set but represents different regions. In Fig. 9, we present two cases, Region A and Region B, which represent sparse and dense change scenarios, respectively. Comparing the change masks generated by different methods, we observe that our method exhibits fewer false positives and false alarm areas. This observation aligns with the evaluation metrics presented in Table 2. Additionally, CDNeXt demonstrates higher precision in boundary prediction and better integrity in individual building predictions, as evident in Region B.

4.5. Ablation study on mainstream attention mechanisms

The ablation experiments on mainstream attention mechanisms validate the effectiveness of the TIAM on SYSU-CD and S2Looking datasets.

Table 3
Comparisons of CDNeXt with other attention modules with interactive attention or not. The TIAM-ws means the V_i has independent parameters for Conv and BN, and the same for $\hat{F}^t$ in Section 3.3. The IA (Independent Attention) ✓ refers to the attention module that operates independently on separate features, F^A and F^B, while IA ✗ denotes concatenation of the two features in channel dimension. The CAM lacks the "$S_1\otimes$" component as described in Eq. (13).

Attention module	IA	SYSU-CD				S2Looking			
		P	R	F1	IoU	P	R	F1	IoU
Base	✗	88.70	75.39	81.51	68.78	71.20	60.08	65.17	48.33
SEM	✗	**89.77**	75.35	81.93	69.39	72.52	59.59	65.42	48.61
CBAM	✗	**90.13**	74.10	81.33	68.54	67.61	**62.77**	65.10	48.26
CBAM	✓	88.17	76.25	81.78	69.17	68.90	**62.58**	65.59	48.80
SAM	✗	88.65	75.34	81.45	68.71	**76.40**	57.13	65.37	48.56
SAM	✓	**90.52**	75.41	82.27	69.88	72.34	60.86	66.10	49.37
DAM	✗	89.25	74.97	81.49	68.76	**75.13**	58.08	65.51	48.71
DAM	✓	89.54	**76.36**	**82.43**	**70.11**	70.31	61.52	65.62	48.83
CAM	✓	89.06	75.65	81.81	69.21	**72.91**	60.72	**66.26**	**49.54**
TIAM-ws	✓	88.45	**78.34**	**83.09**	**71.07**	70.86	62.03	**66.15**	**49.42**
TIAM	✓	89.72	**76.57**	**82.63**	**70.39**	70.78	**63.08**	**66.71**	**50.05**

Our TIAM compares four embeddable attention mechanisms, SEM (Squeeze-and-Excitation Module) (Hu et al., 2020), CBAM (Convolutional Block Attention Module) (Woo et al., 2018), SAM (Self-Attention Module with the Non-local version) (Wang et al., 2018), DAM (Dual-Attention Module) (Fu et al., 2019), and CAM. Additionally, temporal Independent Attention (IA) and concatenate attention are individually validated to examine the impact of intra-temporal self-similarity on change detection.

Table 3 show the results of different attention modules on the same CDNeXt, the TIAM is better than other attention modules in the change detection task. We have the following observations:

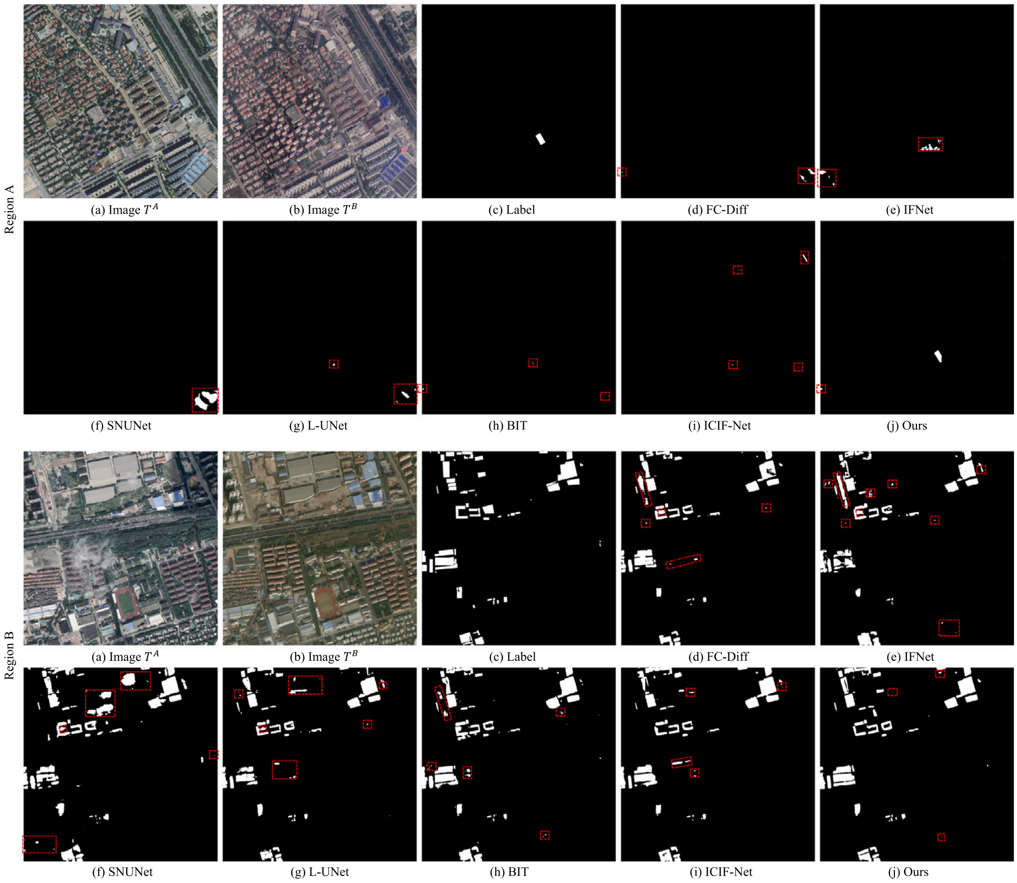


Fig. 9. Visualization results of different methods for BANDON dataset. The red boxes indicate false alarms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- A comparison between IA (✗) and IA (✓) demonstrates the generally negative impact of intra-temporal self-similarity on change detection accuracy in the SYSU-CD and S2Looking datasets. Furthermore, this validates the necessity of temporal feature interaction for change detection tasks.
- The DAM performs feature self-attention on space and channel, limiting its ability to fully capture spatio-temporal information due to intra-temporal self-similarity. The CAM interacts only with geometric perspective dependencies, yielding performance comparable to the SAM and the DAM but lower than the TIAM by at least 0.6%. Other attention mechanisms provide incremental improvements over the base model but fall short of top performance.
- The TIAM-ws performs well on the SYSU-CD dataset with a consistent temporal visual style, while standard TIAM is suitable for most datasets. In the last two columns, TIAM outperforms TIAM-ws by 0.56% on the S2Looking dataset. The performance of the CAM is lower than the TIAM by 0.34% to 1.28%, indicating the significance of temporal style information interaction in CD tasks.

As shown in the red box in Fig. 10, we have also included a qualitative comparison of the effects with and without TIAM. This figure provides a clear and intuitive demonstration of the alleviating effect of TIAM on geometric perspective rotation and temporal style difference problems. The first two rows of images demonstrate false alarms caused by varying viewpoints of tall buildings. The last two rows of images show false alarms and missed detections resulting from temporal styles such as lighting and shadows. Therefore, with the inclusion of TIAM in CDNeXt, robustness is greatly enhanced.

Table 4
Ablation results on CDNeXt framework with the different backbone networks. In the base column, Res18 denotes the backbone ResNet18, and CNX-t means ConvNeXt tiny version. FS ✓denotes boot the hierarchical features fusion. TIAM ✓represents the Interactor turn-on.

CDNeXt with			LEVIR-CD+		S2Looking	
Base	FS	TIAM	F1	IoU	F1	IoU
Res18	✗	✗	82.72	70.53	62.62	45.58
Res18	✗	✓	83.17	71.19	63.56	46.58
Res18	✓	✗	83.51	71.69	63.98	47.04
Res18	✓	✓	**84.45**	**73.08**	**64.40**	**47.50**
CNX-t	✓	✗	85.27	74.35	65.17	48.33
CNX-t	✓	✓	87.14	77.21	66.71	50.05

4.6. Ablation study on CDNeXt framework components

The ablation experiments on the CDNeXt framework validate the necessity of the framework component. We start the ablation study on three important modules, the Encoder with different backbone networks, the Detector with the hierarchical Features fuSion (FS), and Interactor with the TIAM. And we utilized Grad-CAM (Selvaraju et al., 2017) to visualize the mechanism of each component in CDNeXt on feature extraction.

Components in CDNeXt. Table 4 summarizes the F1 and IoU values of our network on two datasets. The quantitative comparisons demonstrate the performance and rationality capability of our framework and TIAM. The CDNeXt with ResNet18, the hierarchical Features fuSion (FS), and the TIAM can also improve performance on different datasets. On the one hand, our well-designed architecture and TIAM module have improved the change detection accuracy by 1.73% for CDNeXt

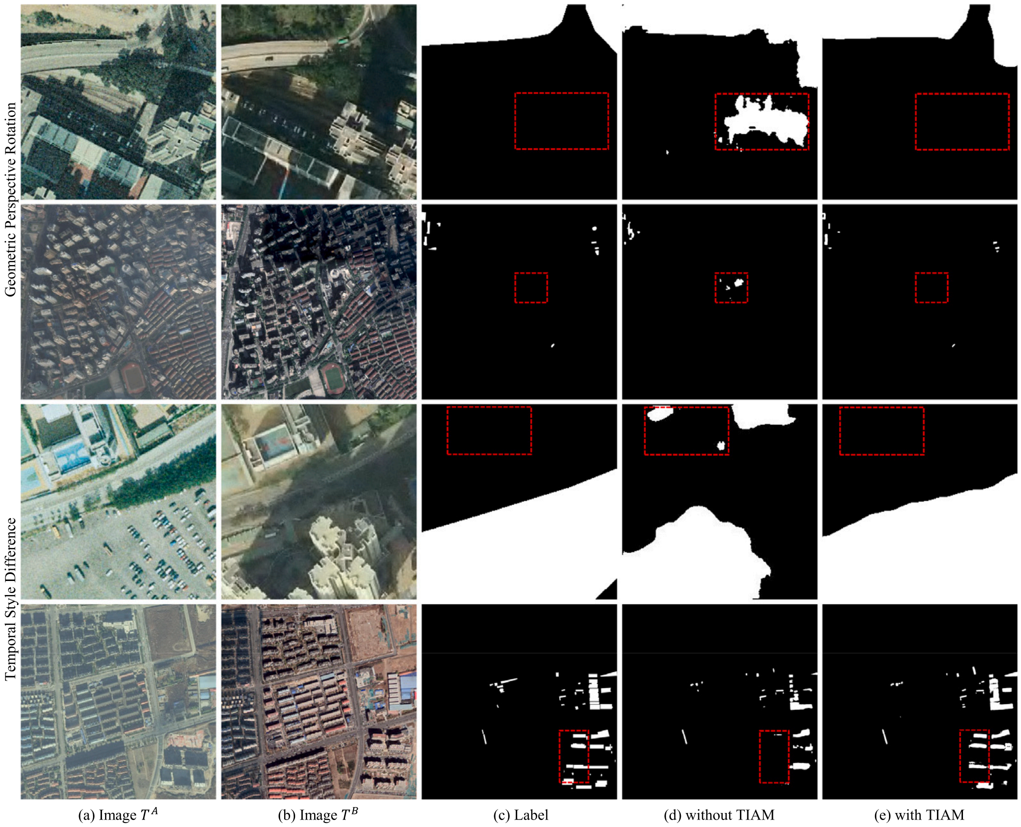


Fig. 10. Visualization results of the effects with and without TIAM. The red boxes indicate false detections by geometric perspective rotation and temporal style difference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with ResNet18 as the backbone network. On the other hand, for better backbone networks ranging from Res18 to CNX-t, our proposed TIAM module can significantly enhance the performance of CDNeXt, with accuracy improvements of 0.94% and 1.87%, respectively.

Visualized in CDNeXt with or without TIAM. As shown in Fig. 11, we can clearly observe the feature transformation process and its contribution to the results in CDNeXt. We notice that CDNeXt with TIAM exhibits fewer false alarms compared to CDNeXt without TIAM when dealing with tall buildings captured from different viewpoints. This indicates that the TIAM module accurately recognizes spatial perspective dependencies under different viewing angles. For example, the high-heat regions in $\tilde{F}^B$ indicate high similarity between objects in this area, while dissimilar objects are represented by high-heat regions in the Decoder. This demonstrates that TIAM can capture the identities of objects from different perspectives, enhancing the extraction of similar targets. At the Detector stage, we observe that CDNeXt with TIAM shows consistent background features even in the presence of large visual style differences. This reflects the effectiveness of the temporal style correlations extracted by TIAM, aligning with our theoretical basis discussed in Section 3.3.

4.7. Ablation study on loss functions

As shown in Table 5, the ablation experiments on mainstream loss functions validate the effectiveness on LEVIR-CD+ and S2Looking datasets. Our CDNeXt compares three kinds of loss functions, Cross-Entropy (CE) loss, Dice loss (Milletari et al., 2016), and Focal loss (Lin et al., 2020). The gamma parameter in the Focal loss is set to 2 to enhance hard example mining, resulting in further improvements on the S2Looking with more complex scene variations. In addition, Dice loss is commonly used in combination with CE loss and achieves the highest F1 on the LEVIR-CD+ that demands stricter edge requirements. However, Dice loss requires tuning the hyperparameters of weighting coefficients. CE loss, on the other hand, demonstrates a more balanced performance and is suitable for framework structure and module improvement studies.

Table 5
Ablation results in different loss functions with different weighting coefficients. “–” denotes the absence of the corresponding loss function.

Loss coefficients			LEVIR-CD+		S2Looking	
CE	Dice	Focal	F1	IoU	F1	IoU
1.0	–	–	87.14	77.21	66.71	50.05
1.0	0.1	–	87.48	77.75	66.86	50.22
1.0	0.5	–	87.39	77.60	66.91	50.27
1.0	1.0	–	**87.83**	**78.30**	66.64	49.97
–	–	1.0	86.91	76.86	**67.12**	**50.52**

4.8. Complexity analysis

In this section, a comparison is made between two CDNeXt methods, two TIAM-embedded methods, and eleven mainstream methods in terms of parameter count (Params), floating-point operations (FLOPs), and F1 accuracy. Params and FLOPs are calculated based on input image dimensions of 3 × 256 × 256 pixels. According to the data presented in Table 6, these results indicate that our CDNeXt effectively balances computational cost and accuracy, achieving significant improvements in F1 accuracy with relatively lower FLOPs. Furthermore, both IFNet and FC-Conc exhibit substantial accuracy improvements after incorporating TIAM, with IFNet achieving at least a 9% increase in accuracy on both datasets without significantly increasing Params and FLOPs. This is attributed to the fact that TIAM operates on quarter-sized image features, allowing for efficient computation of interactive attention at that scale. Additionally, FC-Diff and BIT achieve competitive

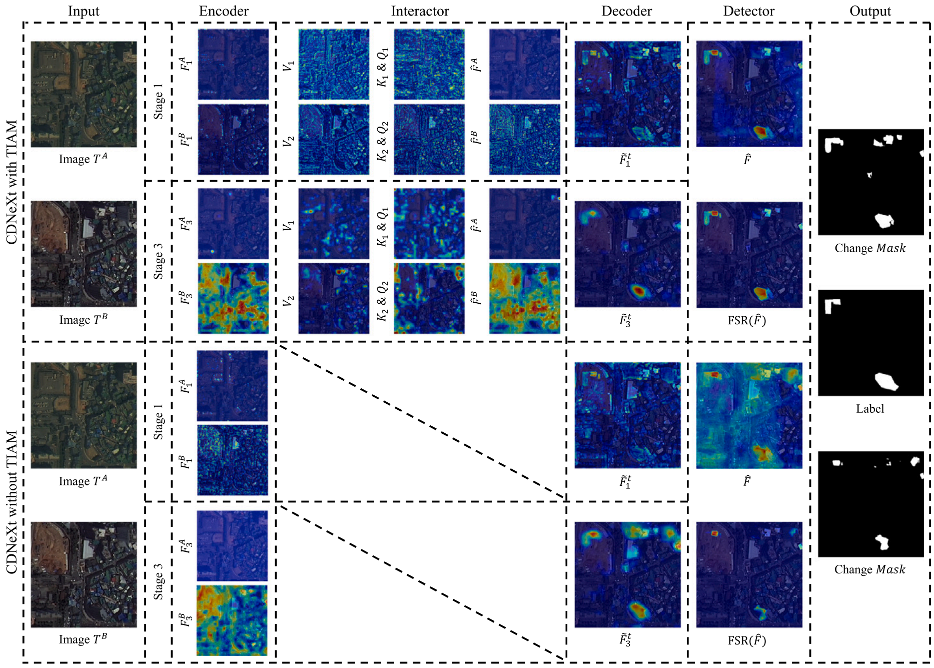


Fig. 11. Example of CDNeXt network in S2Looing dataset, visualized by Grad-CAM. The term "CDNeXt with TIAM" refers to the CDNeXt framework that includes the Interactor module with the TIAM component during training. The header row of the figure is structured similarly to our proposed CDNeXt architecture, and the symbols next to each subfigure indicate their corresponding feature outputs in the CDNeXt computation process, as detailed in Section 3.

Table 6
Quantitative comparison on the number of Parameters (Params), FLoating point OPerations (FLOPs), and F1 accuracy among different methods, showcasing the performance of TIAM when embedded within IFNet and FC-Conc. Lower Params and FLOPs are preferable for better performance.

Method	BANDON F1	S2Looking F1	Params(M)	FLOPs(G)
FC-EF	63.90	50.59	1.35	3.58
FC-Diff	63.82	61.45	1.35	4.73
FC-Conc	63.25	60.38	1.54	5.33
STANet	50.54	30.00	12.21	12.56
IFNet	59.24	56.02	35.73	82.26
DSAMNet	50.23	28.37	12.23	65.68
SNUNet	63.90	60.68	10.20	44.38
L-UNet	64.93	60.43	8.45	17.33
BIT	67.88	62.71	3.04	8.75
ICIF-Net	66.79	59.78	23.84	25.41
P2V	63.29	58.96	5.42	32.96
FC-Conc+TIAM	64.93	60.82	1.55	5.49
IFNet+TIAM	**69.13**	**65.04**	36.92	84.22
CDNeXt-Base	70.02	65.17	37.84	15.18
CDNeXt	71.44	66.71	39.42	15.76

accuracy levels with relatively fewer Params and FLOPs compared to our methods. While the performance of other mainstream methods is acceptable, they have higher requirements for computational resources and currently do not achieve top performance.

5. Conclusion

Addressing the prevalent challenges of geometric perspective rotation and temporal style differences in CD task, this paper presented CDNeXt with the Temporospatial Interactive Attention Module (TIAM), a novel framework for embedding the attention module and the pretrained network. Both the theoretical formulation and experimental results of the TIAM demonstrate its effectiveness in mitigating the aforementioned issues. By comparing our method with mainstream CD methods on both the viewpoint rotation dataset and the visual style dataset, our approach demonstrates superior effectiveness and robustness, and TIAM has shown the potential to enhance the performance of mainstream methods. These experiments also confirm that the targeted datasets designed for respective problems cannot eliminate other interfering factors; Instead, they can be alleviated through the reasonable design of interactive modules. The ablation experiment of the TIAM demonstrated the superiority over mainstream attention mechanisms and confirmed the detrimental effects of intra-temporal self-similarity on CD and the necessity of incorporating interactive attention.

In future work, we plan to improve the performance of our framework for the bi-image similarity task with the TIAM, *e.g.*image matching, multi-modal fusion, and semantic change detection. Moreover, we plan to improve the structure of the TIAM to reduce the amount of computation and improve efficiency.

CRediT authorship contribution statement

Jinjiang Wei: Writing – original draft, Software, Methodology, Conceptualization. **Kaimin Sun:** Writing – review & editing, Supervision, Conceptualization. **Wenzhuo Li:** Writing – review & editing, Conceptualization. **Wangbin Li:** Writing – review & editing, Methodology. **Song Gao:** Validation. **Shunxia Miao:** Visualization. **Qinhui Zhou:** Software. **Junyi Liu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Our code is available on GitHub: https://github.com/wjj282439449/CDNeXt.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2022YFB3902900), the Natural Science Foundation of China Projects (NSFC) (No: 42192583, 41801344).

References

Basavaraju, K.S., Sravya, N., Lal, S., Nalini, J., Reddy, C.S., Dell'Acqua, F., 2022. UCDNet: A deep learning model for urban change detection from bi-temporal multispectral sentinel-2 satellite images. IEEE Trans. Geosci. Remote Sens. 60, 1–10. http://dx.doi.org/10.1109/TGRS.2022.3161337.

Chen, H., Qi, Z., Shi, Z., 2022a. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–14. http://dx.doi.org/10.1109/TGRS.2021.3095166.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sens. 12 (10), 1662. http://dx.doi.org/10.3390/rs12101662.

Chen, P., Zhang, B., Hong, D., Chen, Z., Yang, X., Li, B., 2022b. FCCDN: Feature constraint network for VHR image change detection. ISPRS J. Photogramm. Remote Sens. 187, 101–119. http://dx.doi.org/10.1016/j.isprsjprs.2022.02.021.

Daudt, R.C., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. In: IEEE Int. Conf. Image Process.. ICIP, IEEE, pp. 4063–4067. http://dx.doi.org/10.1109/ICIP.2018.8451652.

Fang, S., Li, K., Shao, J., Li, Z., 2022. SNUNet-CD: A densely connected siamese network for change detection of VHR images. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2021.3056416.

Feng, Y., Xu, H., Jiang, J., Liu, H., Zheng, J., 2022. ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2022.3168331.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 3141–3149. http://dx.doi.org/10.1109/CVPR.2019.00326.

Gao, S., Li, W., Sun, K., Wei, J., Chen, Y., Wang, X., 2022. Built-up area change detection using multi-task network with object-level refinement. Remote Sens. 14 (4), 957. http://dx.doi.org/10.3390/rs14040957.

Hendrycks, D., Gimpel, K., 2017. Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. In: Int. Conf. Learn. Represent.. ICLR, URL https://openreview.net/forum?id=Bk0MRI5lg.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2011–2023. http://dx.doi.org/10.1109/TPAMI.2019.2913372.

Huang, Y., Li, X., Du, Z., Shen, H., 2024. Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection. IEEE Trans. Geosci. Remote Sens. 62, 1–14. http://dx.doi.org/10.1109/TGRS.2024.3360516.

Li, W., Sun, K., Li, W., Wei, J., Miao, S., Gao, S., Zhou, Q., 2023. Aligning semantic distribution in fusing optical and SAR images for land use classification. ISPRS J. Photogramm. Remote Sens. 199, 272–288. http://dx.doi.org/10.1016/j.isprsjprs.2023.04.008.

Li, W., Sun, K., Zhao, H., Li, W., Wei, J., Gao, S., 2022a. Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment. Int. J. Appl. Earth Obs. Geoinf. 113, 102970. http://dx.doi.org/10.1016/j.jag.2022.102970.

Li, Z., Tang, C., Wang, L., Zomaya, A.Y., 2022b. Remote sensing change detection via temporal feature interaction and guided refinement. IEEE Trans. Geosci. Remote Sens. 60, 1–11. http://dx.doi.org/10.1109/TGRS.2022.3199502.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327. http://dx.doi.org/10.1109/TPAMI.2018.2858826.

Lin, M., Yang, G., Zhang, H., 2023. Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images. IEEE Trans. Image Process. 32, 57–71. http://dx.doi.org/10.1109/TIP.2022.3226418.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 11966–11976. http://dx.doi.org/10.1109/CVPR52688.2022.01167.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: Int. Conf. Learn. Represent.. ICLR, URL https://openreview.net/forum?id=Bkg6RiCqY7.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. 4th Int. Conf. 3D Vis.. Ieee, pp. 565–571. http://dx.doi.org/10.1109/3DV.2016.79.

Pan, J., Cui, W., An, X., Huang, X., Zhang, H., Zhang, S., Zhang, R., Li, X., Cheng, W., Hu, Y., 2022. MapsNet: Multi-level feature constraint and fusion network for change detection. Int. J. Appl. Earth Obs. Geoinf. 108, 102676. http://dx.doi.org/10.1016/j.jag.2022.102676.

Pang, C., Wu, J., Ding, J., Song, C., Xia, G.-S., 2023. Detecting building changes with off-nadir aerial images. Sci. China Inf. Sci. 66 (4), 140306. http://dx.doi.org/10.1007/s11432-022-3691-4.

Papadomanolaki, M., Vakalopoulou, M., Karantzalos, K., 2021. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. IEEE Trans. Geosci. Remote Sens. 59 (9), 7651–7668. http://dx.doi.org/10.1109/TGRS.2021.3055584.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE Int. Conf. Comput. Vis.. ICCV, pp. 618–626. http://dx.doi.org/10.1109/ICCV.2017.74.

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 Sens. 13 (24), 5094. http://dx.doi.org/10.3390/rs13245094.

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 Trans. Geosci. Remote Sens. 60, 1–16. http://dx.doi.org/10.1109/TGRS.2021.3085870.

Song, X., Hua, Z., Li, J., 2022. Remote sensing image change detection transformer network based on dual-feature mixed attention. IEEE Trans. Geosci. Remote Sens. 60, 1–16. http://dx.doi.org/10.1109/TGRS.2022.3209972.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 8918–8927. http://dx.doi.org/10.1109/CVPR46437.2021.00881.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need. In: Adv. Neural Inf. Process. Syst. (NeurIPS), Vol. 30, Curran Associates Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. CVPR, pp. 7794–7803. http://dx.doi.org/10.1109/CVPR.2018.00813.

Wang, L., Li, Y., Zhang, M., Shen, X., Peng, W., Shi, W., 2023. MSFF-CDNet: A multiscale feature fusion change detection network for bi-temporal high-resolution remote sensing image. IEEE Geosci. Remote Sens. Lett. 20, 1–5. http://dx.doi.org/10.1109/LGRS.2023.3305623.

Wei, J., Long, C., Zou, H., Xiao, C., 2019. Shadow inpainting and removal using generative adversarial networks with slice convolutions. Comput. Graph. Forum 38 (7), 381–392. http://dx.doi.org/10.1111/cgf.13845.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proc. Eur. Conf. Comput. Vis.. ECCV, Vol. 11211, Springer International Publishing, pp. 3–19. http://dx.doi.org/10.1007/978-3-030-01234-2_1.

Zhang, H., Ma, G., Zhang, Y., 2022. Intelligent-BCD: A novel knowledge-transfer building change detection framework for high-resolution remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 5065–5075. http://dx.doi.org/10.1109/JSTARS.2022.3184298.

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 J. Photogramm. Remote Sens. 166, 183–200. http://dx.doi.org/10.1016/j.isprsjprs.2020.06.003.