Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

HeteCD: Feature Consistency Alignment and difference mining for heterogeneous remote sensing image change detection

Wei Jing [a,b], Haichen Bai [b], Binbin Song [b], Weiping Ni [c], Junzheng Wu [c], Qi Wang [b,*]

[a] National Elite Institute of Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[b] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China
[c] Department of Remote Sensing, Northwest Institute of Nuclear Technology, Xi'an 710072, China

ARTICLE INFO

Keywords:
Heterogeneous change detection
Optical remote sensing image
Synthetic aperture radar image
Heterogeneous feature spaces align
3D spatio-temporal attention difference

ABSTRACT

Optical change detection is limited by imaging conditions, hindering real-time applications. Synthetic Aperture Radar (SAR) overcomes these limitations by penetrating clouds and being unaffected by lighting, enabling all-weather monitoring when combined with optical data. However, existing heterogeneous change detection datasets lack complexity, focusing on single-scene targets. To address this gap, we introduce the XiongAn dataset, a novel urban architectural change dataset designed to advance heterogeneous change detection research. Furthermore, we propose HeteCD, a fully supervised heterogeneous change detection framework. HeteCD employs a Siamese Transformer architecture with non-shared weights to effectively model heterogeneous feature spaces and includes a Feature Consistency Alignment (FCA) loss to harmonize distributions and ensure class consistency across bi-temporal images. Additionally, a 3D Spatio-temporal Attention Difference module is incorporated to extract highly discriminative difference information from bi-temporal features. Extensive experiments on the XiongAn dataset demonstrate that HeteCD achieves a superior IoU of 67.50%, outperforming previous state-of-the-art methods by 1.31%. The code will be available at https://github.com/weiAI1996/HeteCD.

1. Introduction

Change detection, a crucial task in remote sensing interpretation, is designed to identify and monitor changes on the Earth's surface by analyzing image data obtained from remote sensing platforms at various times (Chen et al., 2023a). Automated and semi-automated change detection techniques have been applied to various typical scenarios, including urban planning (Wen et al., 2016), disaster assessment (Abuelgasim et al., 1999), and environmental monitoring (Satalino et al., 2018).

Most change detection benchmarks, such as LEVIR-CD (Chen and Shi, 2020), WHU-CD (Ji et al., 2019), and SVCD (Lebedev et al., 2018), focus on homogeneous optical imagery. However, obtaining two cloud-free, observable optical remote sensing images of the same geographical area within a short temporal window presents significant challenges. Cloud cover, atmospheric disturbances, and varying lighting conditions frequently impede the acquisition of clear optical imagery, leading to incomplete or unreliable data for accurate change detection. These limitations are exacerbated in regions with persistent cloud cover or during adverse weather seasons, reducing the temporal frequency and spatial consistency of available optical datasets. Additionally, temporal misalignment between image acquisitions can result in discrepancies unrelated to actual land changes, further complicating the analysis. These obstacles highlight the necessity for complementary sensing modalities, such as SAR, which can penetrate cloud cover and operate independently of daylight conditions (Lv et al., 2022). Integrating SAR with optical data thus offers a robust solution for continuous, all-weather monitoring, enabling more reliable and comprehensive change detection in diverse and challenging environments (Han et al., 2022b). Although some studies have explored synergistic heterogeneous change detection between optical and SAR images, research in this area remains in its early stages due to differences in imaging principles and data scale limitations.

Unlike optical images, SAR images are generated by emitting and receiving radar waves in the microwave band, allowing operation under low light or cloud cover but resulting in significantly different representations of the same surface types compared to heterogeneous (*i.e.* optical) images (Dellinger et al., 2015). Therefore, conventional arithmetic operators like image differencing or ratio/log ratio are ineffective for multi-temporal heterogeneous images. Early heterogeneous

* Corresponding author.
E-mail addresses: wei_adam@mail.nwpu.edu.cn (W. Jing), hcbai@mail.nwpu.edu.cn (H. Bai), songbinbin@mail.nwpu.edu.cn (B. Song), niweiping@nint.ac.cn (W. Ni), wujunzheng@nint.ac.cn (J. Wu), crabwq@nwpu.edu.cn (Q. Wang).

https://doi.org/10.1016/j.isprsjprs.2025.03.008
Received 6 November 2024; Received in revised form 10 March 2025; Accepted 10 March 2025

change detection methods can be categorized into three types: post-classification comparison, similarity measurement, and feature space unification. The post-classification comparison method involves classifying image pixels from two time points and comparing them pixel by pixel to identify changes (Zhou et al., 2008; Wan et al., 2019). The similarity measurement method assumes that unchanged pixels in heterogeneous images will exhibit high similarity in a certain feature space, while changed pixels will not, allowing change detection through similarity assessment (Alberga, 2009; Wan et al., 2018; Sun et al., 2021). The feature space unification method maps heterogeneous images into a unified feature space where identical features represent the same ground objects, facilitating change detection by comparing features across time points (Liu et al., 2018b; Luppino et al., 2019).

Recent advancements in heterogeneous change detection involve deep learning combined with feature space unification (Zhan et al., 2018; Liu et al., 2018a; Wang et al., 2024a). These methods use deep networks to approximate the feature space of bi-temporal images and perform change detection at the pixel or object level. However, due to the scarcity of labeled datasets, most current research employs unsupervised algorithms to process low spatial resolution images, typically single-scene images with hundreds of pixels. The publicly available datasets usually contain change objects that are relatively prominent in the overall scene. There is still a lack of complex urban scene datasets for building change detection, comparable to those available for homogeneous change detection benchmarks, for the research community to utilize.

In response to these challenges, we collected and screened time-series optical and SAR data from Xiong'an, China, and accurately annotated these samples based on optical images from the same time period to establish a heterogeneous change detection sample set. The constructed XiongAn dataset focuses on building transformations in rural and urban areas during urbanization, including demolition and construction. Additionally, we developed a fully supervised heterogeneous change detection framework, HeteCD, to accurately extract changing building instances from complex backgrounds. Specifically, HeteCD is an encoder–decoder deep network that extracts deep semantic features from heterogeneous images, aligns their feature distributions, and models discriminative differential features for building change detection. On the proposed dataset, HeteCD outperformed state-of-the-art transfer supervised homogeneous change detection algorithms. The main contributions of this work are summarized as follows:

(1) A high-quality dataset for advancing heterogeneous change detection research is proposed, containing 22 pairs of optical and SAR images from different times in the Xiong'an area, with precise manual annotations of changing building instances.

(2) To extract highly discriminative differential features from bitemporal data, a 3D Spatio-temporal Attention Difference module is proposed. This module treats the deep features of heterogeneous images as temporal information, uses 3D convolution to extract change information in the time-space domain, and introduces an attention mechanism to direct the model's focus on important spatio-temporal information.

(3) To bridge the gap between heterogeneous feature spaces and mitigate the risk of semantic ambiguity, we devise a Feature Consistency Alignment loss function to supervise the uniformity and semantic consistency of the heterogeneous feature spaces.

(4) Based on the 3D Spatio-temporal Attention Difference module, a non-shared weight Siamese heterogeneous change detection framework, HeteCD, is proposed. It robustly extracts change instances after mapping heterogeneous data into a unified feature space. Extensive experiments on the XiongAn dataset validate the effectiveness of HeteCD.

This paper is systematically organized into six sections. Section 2 critically reviews existing research on change detection, highlighting the identified limitations in current methodologies. Following this, Section 3 elaborates on the construction process of the heterogeneous change detection dataset proposed in this study, with particular emphasis on cross-domain data acquisition and preprocessing strategies. The subsequent section (Section 4) systematically presents the innovative methodology and theoretical framework developed for this research. Section 5 conducts comparative experiments and performs rigorous data analysis to validate the model's effectiveness through empirical results, employing both quantitative metrics and visual interpretation. Finally, Section 6 synthesizes the key research findings and proposes potential directions for future investigations in heterogeneous change detection.

2. Related work

2.1. Homogeneous change detection

Initially, traditional homogeneous change detection methods generated change maps based on spectral differences at the pixel level (Weismiller et al., 1977). The difference method identifies changes by calculating pixel value differences between remote sensing images at two time points, typically through image subtraction followed by thresholding or other criteria to determine which pixels indicate change. The ratio method computes pixel value ratios between images, highlighting spectral changes (Coppin et al., 2004). The regression analysis method studies the relationship between images, using models to fit data and examine residuals (*i.e.* the differences between predicted and actual values) for changes. Koller and Samimi (2011) developed a semi-automated method that quantifies deforestation using pixels, classifying deforestation status into fire and non-fire categories, while Al Rawashdeh (Rawashdeh, 2011) assessed new irrigation areas using pixel-by-pixel difference detection. These traditional methods are effective in simple, low-resolution scenes but fail in complex, high-resolution scenarios due to their neglect of semantic information around pixels (Wang et al., 2024b).

The advent of object-based methods significantly improved the accuracy of change detection. Following the introduction of the image object concept, object-based change detection has rapidly advanced (Desclée et al., 2006). This approach mimics human visual recognition. Comber et al. (2004) classified images to obtain objects, overlaying these on pixel-based classification results from another image to identify true changes. Desclée et al. (2006) used region-merging techniques to segment multiple remote sensing images in a single pass, identifying objects with statistical reflectance differences and detecting changes through outlier analysis. The multi-temporal image object method overlays time-series images for segmentation, fully utilizing image information and is widely used in time-series remote sensing image change detection. However, this method is vulnerable to registration errors and can produce small or blurry boundary objects due to the heterogeneity in object size and shape.

With the development of remote sensing big data (Lecun et al., 2015; Jing et al., 2023; Wang et al., 2023; Lihe et al., 2024), the availability of large volumes of annotated remote sensing data has facilitated the emergence of numerous deep learning-based change detection models. A notable model is the Siamese network, which extracts distinguishing features from image pairs and performs change detection. To enhance the boundary consistency and internal cohesion of objects in the generated change maps, Zhang et al. (2020) designed a deeply supervised difference discrimination network. This network uses attention modules to fuse deep features extracted from the original inputs with difference features from bi-temporal images, thereby reconstructing the change map. Chen et al. (2023a) addressed the challenge of detecting irrelevant changes by leveraging the Siamese network to capture differential representations between foreground and background. To improve the generalization and detection accuracy of the network, Zhang et al. (2023) developed a composite high-order attention network with multiple encoding paths, named MCHA-Net. This

network includes four learning paths: Siamese learning, residual learning, transformer learning, and decoding paths. By integrating these paths, the network achieves stronger feature representation capability, forming a local–global-cross-domain data modeling approach, thus endowing the network with powerful data perception and mining capabilities. Ye et al. (2023) proposed a novel adjacent-level feature fusion network with 3D convolution for deep learning-based remote sensing change detection, effectively addressing the challenge of extracting and fusing bi-temporal features. Wang et al. (2024c) proposed a novel lightweight network termed FFBDNet (Feature-interleaved Fusion and Bistable Decoding Network), which groups fused features during the decoding phase and employs a dual-phase decoding framework to progressively generate precision variation maps. Addressing the issues of model parameter optimization and sample imbalance, Han et al. (2022a) proposed a lightweight fully convolutional change detection network. This network utilizes an artificially padded convolution (APC) module as the convolution unit of the encoder to achieve more detailed information transmission and feature extraction.

Transformers' self-attention mechanism captures correlations between different image regions, modeling global dependencies and outperforming CNN-based methods in classification and detection tasks (Dosovitskiy et al., 2020). Bandara and Patel (2022) integrated a hierarchical Transformer structure within the Siamese network, effectively presenting multi-scale long-range details for precise change detection, achieving state-of-the-art (SOTA) performance on multiple change detection datasets. Huang et al. (2022) proposed the Iterative Difference Enhancement Transformer (IDET) framework, using transformers to extract long-range information and iteratively enhance feature differences.

2.2. Heterogeneous change detection

Heterogeneous remote sensing image change detection leverages multiple data types to overcome the limitations of single sources. Due to the ease of acquisition and all-weather monitoring capabilities of SAR images, current heterogeneous change detection research primarily focuses on SAR and optical data. Research in this field can be classified into three categories: post-classification comparison, similarity measurement, and unified feature space methods (Han et al., 2022b).

The post-classification comparison method involves classifying images from two time points and comparing the classifications pixel-by-pixel. Pixels that differ between classifications are identified as changed. Zhou et al. (2008) detected changes through separate classification of heterogeneous images, which can accumulate errors and affect accuracy. Wan et al. Wan et al. (2019) improved accuracy by combining multi-temporal segmentation and composite classification to reduce salt-and-pepper noise in SAR images. While the post-classification comparison method can determine the type of land cover change, its performance is highly dependent on the classification accuracy.

The similarity measurement method assumes that unchanged pixels in heterogeneous images exhibit high similarity in a certain feature space, while changed pixels show lower similarity. By measuring this similarity, changes can be distinguished from unchanged pixels, thereby extracting change information. Alberga (2009) used Distance to Independence (DTI), Mutual Information, Cluster Reward Algorithm, Woods Criterion, and Robust Woods Criterion to perform similarity measurements in a moving window approach to extract change information from heterogeneous images. Wan et al. (2018) performed metric analysis on sorted grayscale histograms to identify changes. Sun et al. (2021) used the K-nearest neighbor algorithm to construct non-local structure graphs of image patches and detected changes by measuring graph similarity.

Traditional methods map images into each other's feature spaces for direct comparison (Liu et al., 2018b; Luppino et al., 2019). However, the traditional methods fail to accommodate the increased images' spatio-temporal resolution and scene complexity, researchers have introduced machine learning and deep learning algorithms into heterogeneous change detection field, and achieved better performance. Since optical and SAR images have different noise distributions—Gaussian and Gamma, respectively. Based on this prior knowledge, Zhan et al. (2018) developed the LTFL model, using log transformation to align noise distributions for feature extraction. Liu et al. (2018a) proposed the Symmetric Convolutional Coupling Network to unify heterogeneous images into the same feature space for direct comparison. Niu et al. (2019) used a conditional generative adversarial network to map optical images into the SAR feature space, performing change detection by comparing mapped and approximated images.

Despite the significant advances that deep learning has brought to feature space-based heterogeneous change detection, current research focuses on prominent target changes in simple scenes due to the lack of temporal heterogeneous data in complex scenarios. There is an urgent need for large-scale, publicly available datasets and benchmark frameworks for heterogeneous change detection in complex scenarios.

3. Proposed benchmark dataset

Our comprehensive review of previous work reveals a significant gap in the field of heterogeneous change detection: the lack of precisely annotated heterogeneous datasets for complex scenes. In the following sections, we will provide a detailed introduction to the constructed heterogeneous change detection dataset (XiongAn), including data collection and annotation processes.

3.1. Data collection

Against the backdrop of large-scale urban redevelopment in Xiong'an New Area, China, we have constructed a heterogeneous change detection dataset focusing on building change detection. The XiongAn dataset combines multispectral and SAR data to capture ground changes from 2018 to 2023. Specifically, the multispectral data is sourced from China's GaoFen-2 satellite, with a spatial resolution of 4 m, while the SAR data is derived from GaoFen-3's quad-polarization stripmap mode (QPSI), with a spatial resolution of 8 m.

In terms of data processing, we first utilized ENVI software to perform multilooking on the SAR images to reduce speckle noise and enhance their radiometric resolution. To further improve image quality, the Refine Lee filter algorithm was adopted to reduce speckle while preserving edge and detail information as much as possible. Additionally, to ensure precise alignment of the optical and SAR images within the same geographic coordinate system, we conducted geocoding and radiometric calibration. We manually annotated a large number of tie points in the overlapping areas of the bi-temporal images to achieve accurate spatial registration. Ultimately, after cropping and integration, a high-quality heterogeneous dataset comprising 22 pairs of multispectral and SAR data was constructed, providing robust data support for subsequent building change detection research. As shown in the Table 1, various heterogeneous change detection datasets cover regions in the UK, France, and China, using different sensor combinations. The XiongAn dataset stands out with significantly higher total and change pixels, using GaoFen-2 and GaoFen-3 sensors. This makes it ideal for large-scale, high-precision change detection research.

3.2. Annotation

SAR images rely on microwave signals and possess the capabilities to penetrate cloud layers and perform all-weather imaging. However, the texture and grayscale features of SAR images differ significantly from those of optical images. This disparity complicates the direct alignment and comparison of the two types of images. During the imaging process of SAR, various terrains such as buildings, vegetation, and water bodies exhibit distinct scattering patterns of electromagnetic

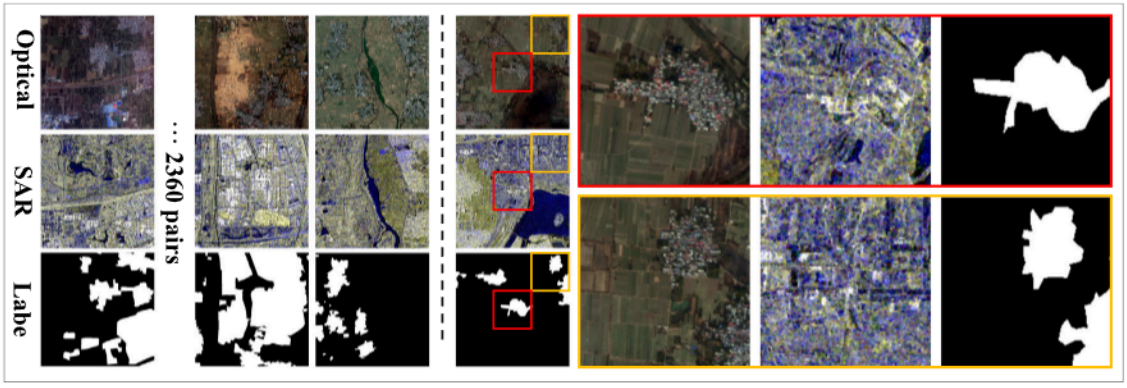

Fig. 1. Annotation samples in the XiongAn dataset. The right side shows close-up of the region with changes in the sample.

Table 1
Comparison of Heterogeneous Change Detection Datasets.

Dataset	Location	Sensors	Total Pixels	Change Pixels
UK Dataset-1	Gloucester, UK	Pre-temporal: TerraSAR-X Post-temporal: QuickBird 02	9.61×10^6	6.17×10^5
France Dataset-1	Toulouse, France	Pre-temporal: TerraSAR-X Post-temporal: Pleiade	1.15×10^7	9.13×10^5
Shuguang Dataset	Shuguang village, China	Pre-temporal: Radarsat-2 Post-temporal: Google Earth	5.46×10^5	8.73×10^3
Island Town Dataset	Island town, China	Pre-temporal: Radarsat-2 Post-temporal: Google Earth	1.67×10^5	6.85×10^3
River Dataset	Yellow River Estuary, China	Pre-temporal: Radarsat-2 Post-temporal: Google Earth	9.98×10^4	4.26×10^3
XiongAn Dataset	**Xiong'an, China**	**Pre-temporal: GaoFen-2 Post-temporal: GaoFen-3**	$\mathbf{2.04 \times 10^8}$	$\mathbf{3.94 \times 10^6}$

waves, imbuing the pixel values in the image with complex physical meanings. Consequently, interpreters require extensive experience to comprehend these patterns.

To address the aforementioned issues, we employ optical images with similar timestamps and historical optical images for annotating heterogeneous change detection data. We have established uniform annotation standards and procedures to ensure consistency and reliability in our annotations. Our approach includes: (1) A clear definition of changes: The XiongAn dataset focuses exclusively on architectural evolution, considering the significant influence of seasonal factors on surface coverage types such as vegetation. (2) A unified definition of the boundaries of change areas: The outermost boundary of the changed buildings in bi-temporal images is used as the boundary for the change instances, regardless of the type of change (construction, demolition, or reconstruction). (3) A definition of annotation precision: Pixel-level annotations are executed, and the boundary error of change instances does not exceed three pixels. (4) Label calibration: Upon completion of annotations, multiple interpretation experts perform cross-validation for more precise label calibration, ensuring the quality of the final annotated results. Some examples of the annotations are shown in Fig. 1.

4. Methodology

4.1. Overall framework of HeteCD

HeteCD is an end-to-end network designed to unify the feature distribution of heterogeneous data and ensure intra-class consistency while exploring intrinsic differences in bi-temporal images. The network includes Siamese encoders with non-shared weights and a Hete-Alignment spatio-temporal decoder. Given the significant differences in feature distribution between optical and SAR images, we use transformers with non-shared weights to model their feature spaces separately. Within the decoding network, we unify these feature spaces using class consistency alignment loss and distribution consistency alignment loss. Instead of the commonly used difference feature modeling strategy in homogeneous change detection, we propose a 3D Spatio-temporal Attention Difference module to simultaneously model highly discriminative difference information in both spatial and temporal dimensions.

As shown in Fig. 2, HeteCD takes heterogeneous images as inputs, using a Siamese transformer backbone to extract spatio-temporal features at different hierarchical levels. These features are aligned in the decoder through the proposed FCA loss. Then, a pyramid strategy is used to fuse the semantic features from different levels represented as follows:

$$X = \mathcal{K}_{1\times1}\left(\mathcal{B}\left(\mathcal{G}\left(\Pi_{i=1}^{4}\mathcal{K}_{1\times1}\left(\mathcal{U}(X_i)\right)\right)\right)\right), \tag{1}$$

where $\mathcal{K}_{1\times1}(\cdot)$ is a linear convolution kernel, $\mathcal{B}$ refers to batch normalization, $\mathcal{G}$ refers to the GELU activation function, Π denotes the operation of concatenation, and $\mathcal{U}(\cdot)$ denotes the operation of upsampling. The differential information is robustly extracted at multiple levels using the 3D Spatio-temporal Attention Difference module. The computation process can be expressed as:

$$\tilde{X}_{diff} = \mathcal{D}\left(\tilde{X}_{t1}, X_{t2}\right), \tag{2}$$

where $\mathcal{D}$ denotes the 3D spatio-temporal attention differencing calculation, X_{t1} and X_{t2} are optical and SAR features respectively. The prediction layer primarily consists of convolution layers:

$$y_p = \mathcal{K}_{1\times1}\left(\mathcal{B}\left(\mathcal{R}\left(\mathcal{K}_{3\times3}\left(\tilde{X}_{diff}\right)\right)\right)\right), \tag{3}$$

where $\mathcal{R}$ refers to the ReLU activation function. This design avoids the need for manually crafted and computationally demanding components typically used in alternative methods.

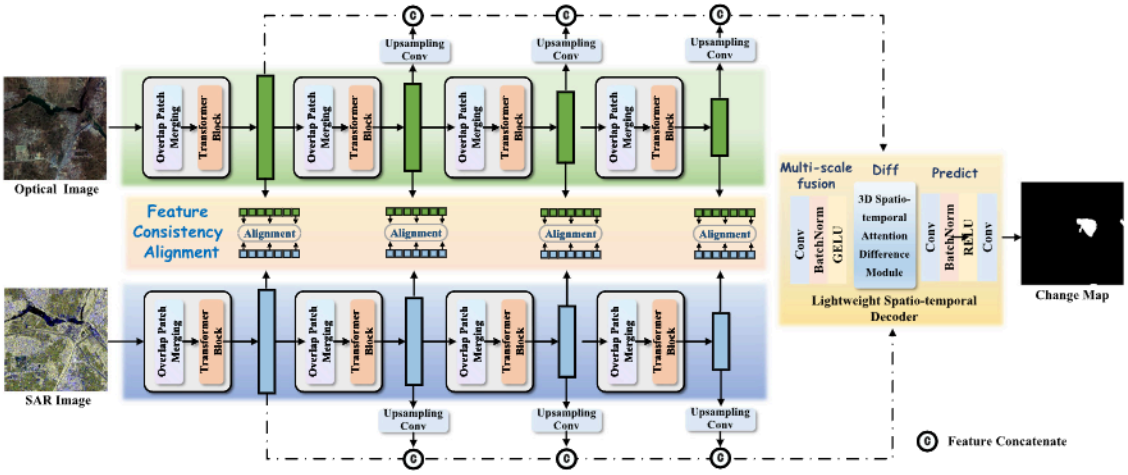

Fig. 2. Overall Framework of HeteCD. The feature consistency alignment loss is used to constrain the heterogeneous feature space, and the 3D Spatio-temporal Attention Difference module is employed to model high-discriminative discrepancy information in both spatial and temporal dimensions.

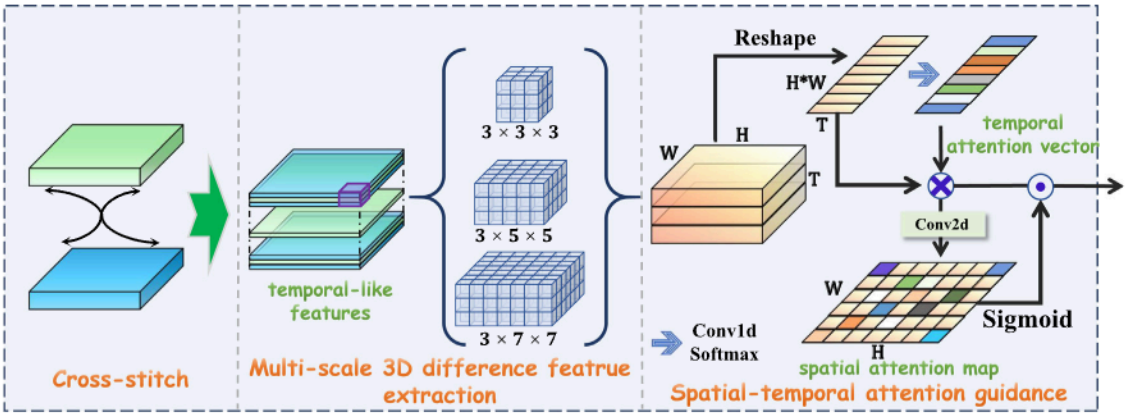

Fig. 3. Structure of the proposed 3D Spatio-temporal Attention Difference module.

4.2. 3D spatio-temporal attention difference module

High-precision change detection relies on accurately modeling and understanding differences between images. However, traditional point-to-point differencing focuses only on local pixel-level variations, neglecting the broader spatial context and thus limiting the model's adaptability and accuracy under complex or unknown conditions. Although two-dimensional convolution on bi-temporal cascaded features considers spatial representation, it overlooks the temporal correlations of cross-temporal features, making adaptation to diverse observation environments and imaging conditions challenging.

Therefore, to effectively handle differences under various imaging conditions, and to deeply understand the essential changes in scene content, we propose a 3D Spatio-temporal Attention Difference module. This module robustly models heterogeneous difference features across both temporal and spatial dimensions, utilizing the attention mechanism to focus on highly discriminative spatio-temporal changes.

As illustrated in Fig. 3, bi-temporal features are no longer merged directly. Instead, channels are alternately drawn from X_1 and X_2. Specifically, X_1 provides channels at even indices (*e.g.* 0, 2, 4, 6), while X_2 provides channels at odd indices (*e.g.* 1, 3, 5, 7). Then, we treat the fused features of adjacent three channels as a temporal sequence and utilize 3D convolution kernels of varying spatial scales to extract difference information, represented as follows:

$$X_{diff} = \mathcal{K}_{3\times3\times3}\left(X_{fuse}\right) \parallel \mathcal{K}_{3\times5\times5}\left(X_{fuse}\right) \parallel \mathcal{K}_{3\times7\times7}\left(X_{fuse}\right), \tag{4}$$

where $\mathcal{K}_{3\times3\times3}$ denotes a 3D convolution with kernel size $3 \times 3 \times 3$, and $\parallel$ indicates the concatenation operation in dimension T. After merging the spatial dimensions of the difference features into a new dimension, they are reshaped to $[T, H * W]$. A one-dimensional convolution is then applied at each time step to generate attention weights, which are constrained within the range $[0, 1]$, using the Softmax activation function. These weights serve as effective distributions for weighting the temporal features of the input tensor. The attention weights are applied to the original difference features through tensor multiplication. This process is computed as follows:

$$X^t_{diff} = \mathcal{R}e\left(X_{diff}\right) \cdot \delta\left(\mathcal{K}_1\left(\mathcal{R}e\left(X_{diff}\right)\right)\right), \tag{5}$$

where $\mathcal{R}e$ denotes reshape operation, δ refers to the Softmax activation function, $\cdot$ denotes tensor multiplication, and $\mathcal{K}_1$ denotes a 1D convolution with kernel size 1.

In the spatial attention phase, a two-dimensional convolution compresses the feature X^t_{diff} into a spatial tensor of dimensions $H \times W$, which represents a map of importance scores for each spatial location. The Sigmoid activation function then maps the output to probability values within the range $[0, 1]$. Finally, the attention weights are applied to the input features, allowing the model to focus on important spatial locations while ignoring less critical ones. This is calculated as follows:

$$\tilde{X}_{diff} = X^t_{diff} \odot \sigma\left(\mathcal{K}_{7\times7}\left(X^t_{diff}\right)\right), \tag{6}$$

where σ refers to the Sigmoid activation function, $\odot$ denotes point-to-point multiplication, and $\mathcal{K}_{7\times7}$ denotes a 2D convolution with kernel size 7×7.

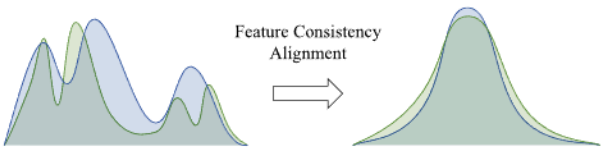

Fig. 4. Illustration of the proposed Feature Consistency Alignment loss $\mathcal{L}_{fca}$.

4.3. Loss functions

To train the HeteCD model, two distinct loss functions were employed: the segmentation loss $\mathcal{L}_{seg}$, and the FCA loss $\mathcal{L}_{fca}$.

Given the significant imbalance between positive and negative samples in change detection tasks, we introduced the DICE loss function to guide the model to focus more on identifying and exploring foreground regions. During model training, we supervised the change segmentation map using a combination of cross-entropy loss and DICE loss, as follows:

$$\mathcal{L}_{seg} = -\sum_{c=1}^{M}\left(y_c \log\left(p_c\right) + \frac{2y_c p_c}{{y_c}^2 + {p_c}^2}\right), \tag{7}$$

where M indicates the number of categories, i is the category index, y_c represents the probability of the true class being i, and p_c is the probability of the predicted class being i. As shown in Fig. 4, to address the distribution gap between heterogeneous data, we constructed a FCA loss to encourage the feature spaces of the Siamese network to be as close as possible. This loss addresses both the distribution and semantic consistency of features. The Kullback–Leibler divergence is used to quantify the closeness between the feature distributions of two types of data. Minimizing this metric facilitates a more consistent alignment of the network's output feature distributions. The FCA loss approximates heterogeneous feature distributions across both channel and spatial dimensions. A temperature parameter $\mathcal{T}$ is introduced to smooth the probability distributions, guiding the model to focus on internal feature representations and relative class relationships within the heterogeneous features. The FCA loss is summarized as follows:

$$\begin{aligned}\mathcal{L}_{fca} = &\frac{\mathcal{T}^2}{HW}\sum_{i=1}^{HW} p\left(X_{t1}^{i}\right)\cdot\log\left(\frac{p\left(X_{t1}^{i}\right)}{p\left(X_{t2}^{i}\right)}\right) + \\ &\frac{\mathcal{T}^2}{C}\sum_{c=1}^{C} p\left(X_{t1}^{c}\right)\cdot\log\left(\frac{p\left(X_{t1}^{c}\right)}{p\left(X_{t2}^{c}\right)}\right),\end{aligned} \tag{8}$$

where H and W are the spatial dimensions of the feature X, and C is channels dimension. Finally, the segmentation loss and feature consistency loss are weighted as follows:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\frac{1}{e}\mathcal{L}_{fca}, \tag{9}$$

where e is the number of current epoch, and α is the balance parameter is set to 2.

5. Experiments

In this section, we design a series of experiments to benchmark the XiongAn dataset. Additionally, we conducted several experiments to compare the performance of the proposed HeteCD network with SOTA fully-supervised homogeneous change detection algorithms. The experimental results are presented and discussed in this section.

5.1. Dataset description and experimental setting

(1) XiongAn dataset: From the XiongAn dataset, 18 of 22 image pairs were allocated to the training set, with the remaining 3 pairs designated for testing. Prior to training, panoramic images were segmented into patches using a window size of 512 pixels and a stride of 256 pixels, resulting in 1941 training samples and 419 test samples. Notably, 1164 training samples and 275 test samples exhibited changes. The red, green, and blue bands were selected to effectively capture variations in color, texture, and structural features, thereby reducing data redundancy and enhancing computational efficiency. For SAR data, the HH, VV, and HV polarizations were chosen from the initial four (HH, HV, VH, VV). The HH polarization captures surface characteristics, VV is sensitive to vertical structures, and HV provides additional information on surface roughness and texture. This combination facilitates a comprehensive analysis of building structures from multiple perspectives, thereby increasing the accuracy and robustness of change detection. To enhance model generalization, several random data augmentation techniques were applied to the input images during training, including random cropping, flipping, color space transformations, and rotations. Finally, all input images were normalized by scaling pixel values by 255.

(2) HTCD dataset (Shao et al., 2021): This dataset is designed for change detection using heterogeneous data from satellites and unmanned aerial vehicles (UAVs). This dataset spans two temporal snapshots from 2008 to 2020, covering approximately 36 square kilometers around Kishinev and its surroundings. The images primarily document urban changes, such as modifications to buildings and roads, with man-made features precisely annotated. Natural changes, however, are not included. The satellite images have a resolution of 0.5971 m, while the UAV images offer a higher resolution of 7.465 centimeters. During training, the images are cropped into 256 × 256 patches with 3080 pairs of training samples and 1321 pairs of test samples, and other parameters are consistent with those used in experiments on the XiongAn dataset.

(3) Experimental Setting: Our experiments were conducted on a server running Ubuntu 18.04 with an NVIDIA RTX 3090 GPU equipped with 24 GB of graphical memory. For fairness, all backbone models were pretrained on ImageNet using the Pytorch framework, and all models were trained for 200 epochs using open-source code. The models are trained with the AdamW optimizer, with the initial learning rate set to 5e−4.

5.2. Evaluation metrics

For the quantitative evaluation of change detection, commonly used metrics include Intersection over Union (IoU), Precision, Recall, and the F1-score (F1).

IoU measures the overlap between the predicted and ground truth regions, defined as the ratio of the intersection to the union of these regions. Mathematically, it is expressed as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{10}$$

where TP, FP, and FN represent the true positive, false positive, and false negative pixels, respectively.

Precision quantifies the accuracy of positive predictions, representing the proportion of true positives (TP) among all predicted positives (TP + FP). It is computed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{11}$$

Recall, also known as sensitivity, indicates the proportion of true positives (TP) identified from all actual positives (TP + FN). It is given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

F1 is the harmonic mean of Precision and Recall, providing a balanced measure of both metrics:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

Table 2
Quantitative Results on the XiongAn. The Best Results are Marked in Bold.

Method	Precision	Recall	F1	IoU	Params (M)	FLOPs (G)	Inference Speed (ms)
STANet	49.10	**80.47**	60.99	43.87	13.34	89.51	176.65
HANet	74.53	73.22	73.87	58.57	**2.61**	70.78	39.24
INF	74.06	80.18	77.00	62.60	35.73	329.02	50.35
SNUNet	74.20	71.84	73.00	57.48	3.01	55.14	39.06
BIT	77.13	67.06	71.75	55.94	2.99	34.70	48.20
DTCDSCN	83.12	65.82	73.47	58.06	31.26	52.90	43.66
ChangeFormer	78.44	75.48	76.93	62.51	41.03	811.15	232.24
CGNet	**85.75**	72.80	78.75	64.95	33.68	329.58	102.83
TTP	80.69	65.84	72.51	56.88	150.65	452.23	339.42
Changer	80.89	76.29	78.53	64.65	11.35	**23.10**	**20.44**
BAN	82.81	76.74	79.66	66.19	3.95	35.09	87.71
ChangeLN	79.14	79.08	79.11	65.44	12.35	26.63	40.80
HeteCD	81.18	80.02	**80.60**	**67.50**	50.19	58.23	156.44

Table 3
Quantitative Results on the HTCD. The Best Results are Marked in Bold.

Method	Precision	Recall	F1	IoU
STANet	51.31	72.38	60.05	42.91
HANet	69.91	73.22	60.89	48.24
INF	89.96	91.13	90.54	82.72
SNUNet	63.09	71.74	67.14	50.53
BIT	47.74	87.34	61.74	44.65
DTCDSCN	**94.78**	80.13	86.84	76.74
ChangeFormer	94.61	90.09	92.30	85.70
CGNet	92.58	89.27	90.90	83.31
TTP	93.46	90.84	92.13	85.41
Changer	90.45	87.80	89.11	80.36
BAN	78.78	85.77	82.13	69.68
HeteCD	94.21	**94.49**	**94.34**	**89.30**

5.3. Comparison with advanced methods

Given the current lack of established datasets and fully supervised methods for heterogeneous change detection, existing approaches predominantly rely on homogeneous data assumptions. Therefore, we ported 12 state-of-the-art homogeneous CD methods through targeted architectural adjustments to serve as benchmarks for our evaluation. The implemented benchmark comprises:

(1) STANet (Chen and Shi, 2020) is a siamese network-based spatio-temporal attention neural network that models spatio-temporal dependencies through a designed change detection self-attention mechanism, introducing self-attention in multi-scale sub-regions to generate more distinctive features that accommodate objects of varying scales.
(2) HANet (Han et al., 2023a) is a siamese network that integrates multi-scale features and refines detailed representations through its core component, the HAN module, which is a lightweight and efficient self-attention mechanism.
(3) IFN (Zhang et al., 2020) is a deep supervised image fusion network for high-resolution multi-temporal remote sensing change detection that extracts deep features from dual-time images and utilizes a differential discriminator network (DDN) with an attention module to enhance change map quality.
(4) SNUNet (Fang et al., 2022) is a dense connection siamese network for change detection that combines a siamese network with NestedUNet, alleviating the loss of deep spatial information through compact information transfer between the encoder and decoder, as well as between decoders.
(5) BIT (Hao Chen and Shi, 2021) is a bi-temporal image transformer network that models context in the spatial–temporal domain by representing images as tokens and refining them back to pixel space within a deep feature differentiation-based change detection framework.
(6) DTCDSCN (Liu et al., 2021) is a dual-task constrained deep siamese convolution network consisting of a change detection network and two semantic segmentation networks, featuring a dual attention module (DAM) and an improved focus loss to address sample imbalance issues.
(7) ChangFormer (Bandara and Patel, 2022) is a transformer-based siamese network for change detection from homogeneous remote sensing images, integrating a hierarchical transformer encoder with an MLP decoder to efficiently capture multi-scale long-range details.
(8) CGNet (Han et al., 2023b) is a change-guided network that enhances change feature representation and edge detection by generating change maps from deep semantic features to guide multi-scale fusion, incorporating a Change Guide Module to capture long-range pixel dependencies.
(9) TTP (Chen et al., 2023b) is a change detection network that leverages bi-temporal feature fusion and knowledge from the SAM base model to enhance change detection in remote sensing images, incorporating time-travel activation gates for better identification of changes.
(10) Changer (Fang et al., 2023) is a change detection architecture that incorporates selectable interaction layers in its feature extractor and introduces a Flowing Dual Alignment Fusion module for interactive alignment and feature fusion of bi-temporal features.
(11) BAN (Li et al., 2024) is a universal change detection framework based on foundational models, designed to extract knowledge for change detection; it comprises three components: a frozen foundational model, a Bi-Temporal Adapter Branch, and a bridging module between them.
(12) ChangeLN (Jing et al., 2024) integrates 3-D Neighborhood Cross-Differencing and Detail Refinement Decoder to precisely capture land cover changes through cross-temporal feature interaction and deep differential feature extraction.

Quantitative Comparison: Table 2 presents the quantitative performance of various change detection methods on the XiongAn dataset. Among the evaluated approaches, HeteCD exhibits superior effectiveness in key metrics, underscoring its robustness in change detection tasks. Specifically, HeteCD achieves the highest F1 (80.60) and IoU (67.50), outperforming all other methods. This indicates a balanced precision and recall, as well as enhanced overlap accuracy in detected changes. Regarding computational efficiency, HeteCD demonstrates moderate parameter counts (50.19 million) and FLOPs (58.23 billion), positioning it between lightweight models like HANet and more resource-intensive counterparts such as ChangeFormer. However, HeteCD's parameter and FLOP requirements are relatively higher compared to some of the most efficient models, which may pose challenges in resource-constrained environments. Additionally, its inference speed, though acceptable, does not match the real-time capabilities of faster models like Changer. In summary, HeteCD offers a compelling balance of high accuracy, as evidenced by its leading F1 and IoU scores, while

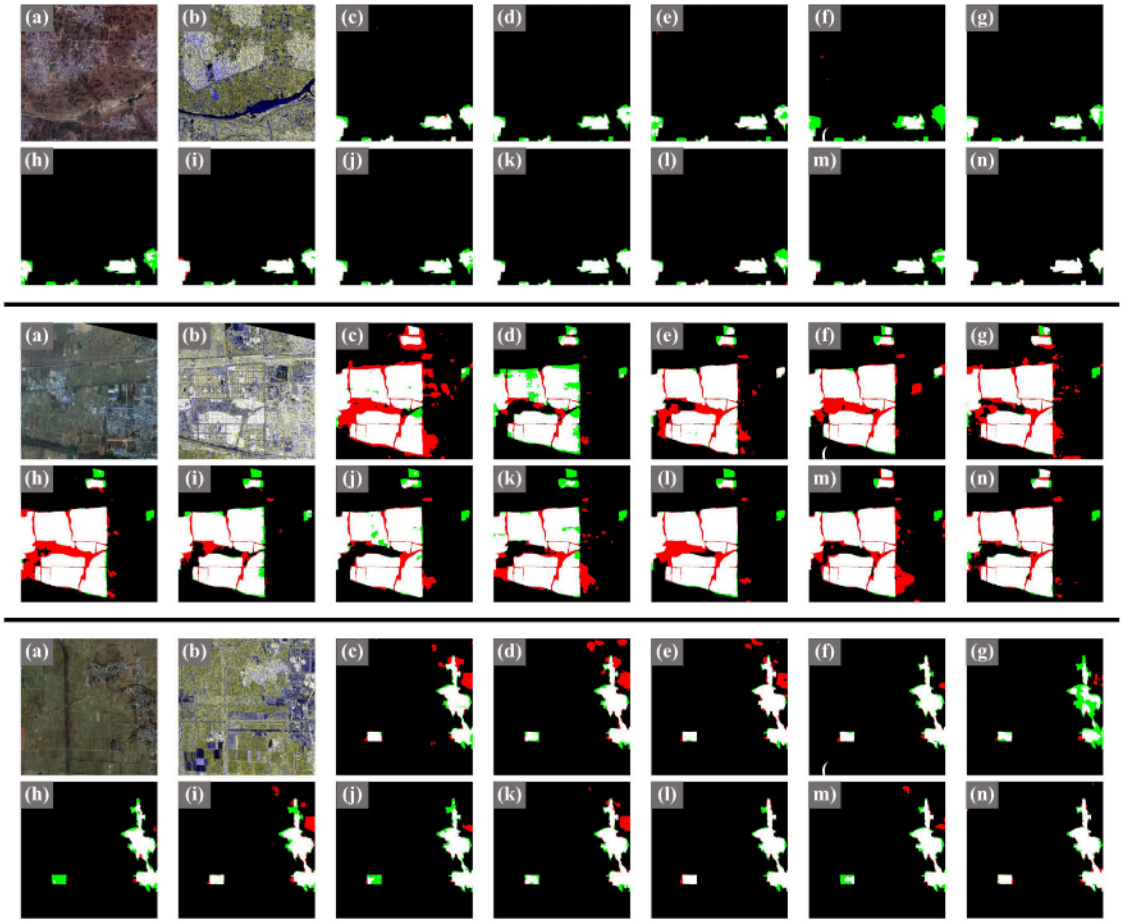

Fig. 5. The qualitative comparison of different methods on the DSIFN dataset. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) STANet. (d) HANet. (e) IFN. (f) SNUNet. (g) BIT. (h) DTCDSCN. (i) ChangeFormer. (j) CGNet. (k) TTP. (l) Changer. (m) BAN. (n) ours. We highlight the TP areas in white, the FP areas in red, and the FN areas in green. The black color denotes the TN areas.

maintaining moderate computational requirements. These attributes render HeteCD a robust choice for heterogeneous change detection, particularly in applications where accuracy is paramount and computational resources are sufficiently available. Future work may focus on optimizing its efficiency to further enhance its applicability across diverse operational contexts.

To verify the performance of the proposed method in other heterogeneous scenarios, we conducted comparative experiments on the HTCD dataset. As demonstrated in Table 3, HeteCD excels across multiple key performance indicators, particularly excelling in recall, F1 score, and IoU metrics. HeteCD achieves a high recall of 94.49%, an F1 score of 94.34%, and leads in IoU with 89.30%, outperforming other methods.

Qualitative Comparison: Fig. 5 illustrates the change detection performance across three distinct scenarios. In the first scenario, involving the demolition of low-rise buildings, most existing methods successfully detect these changes. However, the proposed method not only comprehensively and accurately identifies all building alterations but also exhibits superior detail capture. The second scenario depicts the dynamic development of high-rise residential areas during new urban construction. Comparative analysis of change detection results between optical images and SAR images reveals that, relative to competing methods, HeteCD significantly reduces the false positive rate at building boundaries, underscoring its stability and reliability in complex environments. The final scenario showcases the transformation of rural low-rise residential areas into high-rise structures, as well as the construction of new buildings on previously undeveloped land. While most existing methods perform poorly in detecting small-scale new constructions, often resulting in missed detections or false alarms, the proposed method accurately and comprehensively identifies all building changes, with only minimal false positives in contentious regions and areas of high SAR intensity. Overall, the proposed method markedly outperforms existing techniques in detail recognition and change detection accuracy, demonstrating robust adaptability across various complex scenarios.

Table 4
Ablation experiment on the XiongAn dataset. The best results are marked in bold.

Non-Shared Backbone	3D-STA	FCA Loss	F1	IoU
			77.28	62.97
✔			77.58	63.29
✔	✔		78.88	65.11
✔		✔	78.53	64.48
✔	✔	✔	**80.60**	**67.50**

5.4. Ablation studies

To investigate the impact of the proposed module and loss function on HeteCD, comprehensive ablation studies were performed using the XiongAn dataset. As indicated in Table 4, there is a noticeable improvement in model performance with the incremental addition of components.

As shown in Table 4, particularly in rows 3 and 5, the proposed difference module significantly improves the performance of change

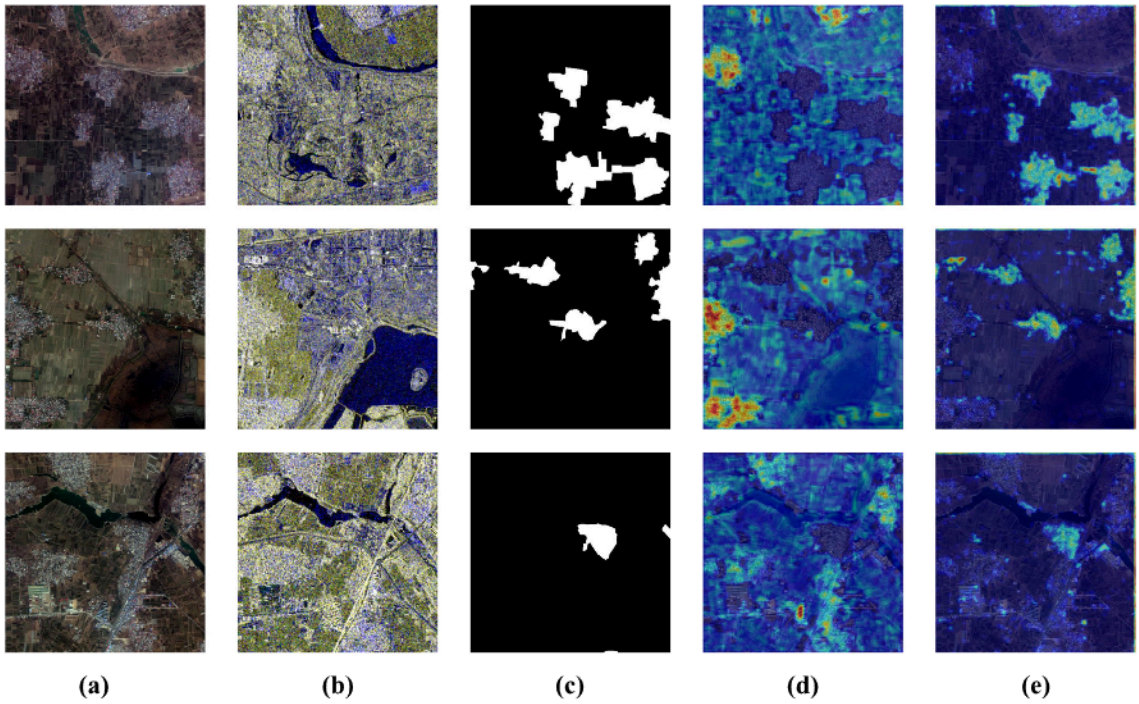

Fig. 6. Example of 3D Spatio-temporal Attention Difference module visualization by Grad-CAM. (a) Pre-temporal image. (b) Post-temporal image. (c) Binary change ground truth. (d) Grad-CAM of the model without the proposed difference module. (e) Grad-CAM of the model with the proposed difference module.

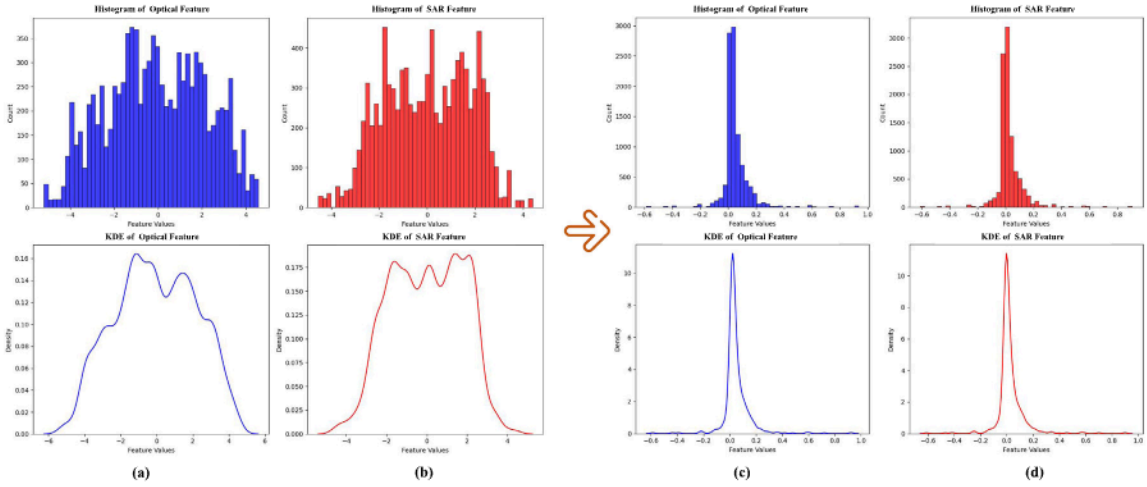

Fig. 7. Feature Distribution Histogram Visualization. (a) Feature distribution of the optical image without FCA. (b) Feature distribution of the SAR image without FCA. (c) Feature distribution of the optical image with FCA. (d) Feature distribution of the SAR image with FCA.

detection. Notably, integrating the 3D Spatio-temporal Attention Difference module into the baseline leads to a substantial increase in the mIoU metric, from 63.29% to 65.11%. To further demonstrate the efficacy of the proposed difference module in extracting robust differential features, this study uses Gradient-weighted Class Activation Mapping (Selvaraju et al., 2017) (Grad-CAM) for visual comparison with and without the module. The model without this strategy relies on traditional convolution to model differential features. As illustrated in Fig. 6, the 3D Spatio-temporal Attention Difference module extracts change features more comprehensively compared to traditional convolution techniques. Additionally, it significantly enhances the regions of object changes, directing gradient propagation to focus on the areas of change. In contrast, traditional convolution methods concentrate on high-brightness regions in SAR imagery, failing to understand the differential extraction mechanism in bi-temporal features, and instead forcefully fitting the change regions through subsequent decoders.

In addition to the 3D Spatio-temporal Attention Difference module, the proposed FCA loss function also yields significant improvements. When combined with the introduced modules, the FCA loss function provides complementary advantages. Specifically, training the baseline network with the FCA loss results in an increase of 0.95% in the F1 score and 1.19% in the mIoU metric for the heterogeneous change detection task. When integrated with the 3D Spatio-temporal Attention Difference module, the model achieves an IoU of 67.50%. Histograms and Kernel Density Estimation (KDE) provide a clear visualization of the spatial distribution of deep features, which is highly effective in analyzing feature consistency. As illustrated in Fig. 7, under the supervision of the FCA loss, the feature spaces of heterogeneous images, such as optical and SAR, progressively converge. This is achieved through minimizing the Kullback–Leibler divergence, which ensures that the distributions are not only aligned but also smoothed. Such smoothing reduces noise and emphasizes meaningful features, leading

to the observed changes in histogram extrema. Specifically, the redistribution of feature peaks and filtering of redundant features enhance the adaptability of HeteCD to multi-source data and significantly improve the accuracy of change detection. This alignment and refining of the histogram make it less noisy and more representative of core feature alignments, thereby facilitating a more effective change detection by the decoder network.

6. Conclusion

In this paper, we present a benchmark for cross-source change detection in urban scenes, named XiongAn. To address the distribution gap of heterogeneous data, we propose a fully supervised heterogeneous change detection framework, termed HeteCD. This framework incorporates a novel Feature Consistency Alignment loss to unify the high-dimensional feature spaces of heterogeneous images. Additionally, we design a 3D Spatio-temporal Attention Difference module to extract highly discriminative differential feature information from bi-temporal features. Experimental results demonstrate that HeteCD outperforms SOTA methods in both quantitative metrics and qualitative results on the XiongAn dataset. Given the prevalent issue of missed detections for small and insignificant targets in existing algorithms, future research will focus more on addressing subtle changes in heterogeneous data.

CRediT authorship contribution statement

Wei Jing: Writing – original draft, Methodology. **Haichen Bai:** Writing – review & editing. **Binbin Song:** Validation, Conceptualization. **Weiping Ni:** Data curation. **Junzheng Wu:** Validation, Data curation. **Qi Wang:** Funding acquisition.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62471394 and U21B2041, and the Innovation Foundation for Doctoral Dissertation of Northwestern Polytechnical University under Grant CX2024108.

References

Abuelgasim, A.A., Ross, W., Gopal, S., Woodcock, C., 1999. Change detection using adaptive fuzzy neural networks: Environmental damage assessment after the Gulf War. Remote Sens. Environ. 70 (2), 208–223.

Alberga, V., 2009. Similarity measures of remotely sensed multi-sensor images for change detection applications. Remote. Sens. 1 (3), 122–143. http://dx.doi.org/10.3390/rs1030122, URL https://www.mdpi.com/2072-4292/1/3/122.

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

Chen, C.-P., Hsieh, J.-W., Chen, P.-Y., Hsieh, Y.-K., Wang, B.-S., 2023a. SARAS-net: scale and relation aware siamese network for change detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 12. AAAI, pp. 14187–14195.

Chen, K., Liu, C., Li, W., Liu, Z., Chen, H., Zhang, H., Zou, Z., Shi, Z., 2023b. Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection. arXiv preprint arXiv:2312.16202.

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), http://dx.doi.org/10.3390/rs12101662, URL https://www.mdpi.com/2072-4292/12/10/1662.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., 2004. Assessment of a semantic statistical approach to detecting land cover change using inconsistent data sets. Photogramm. Eng. Remote Sens. 70, 931–938, URL https://api.semanticscholar.org/CorpusID:130608878.

Coppin, P., Jonckheere, I., Nackaerts, K., et al., 2004. Digital change detection methods in ecosystem monitoring: a review. Int. J. Remote Sens. 25 (9), 1565–1596.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2015. SAR-SIFT: A SIFT-like algorithm for SAR images. IEEE Trans. Geosci. Remote Sens. 53 (1), 453–466. http://dx.doi.org/10.1109/TGRS.2014.2323552.

Desclée, B., Bogaert, P., Defourny, P., 2006. Forest change detection by statistical object-based method. Remote Sens. Environ. 102 (1), 1–11. http://dx.doi.org/10.1016/j.rse.2006.01.013, URL https://www.sciencedirect.com/science/article/pii/S0034425706000344.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fang, S., Li, K., Li, Z., 2023. Changer: Feature interaction is what you need for change detection. IEEE Trans. Geosci. Remote Sens. 61, 1–11. http://dx.doi.org/10.1109/TGRS.2023.3277496.

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.

Han, M., Li, R., Zhang, C., 2022a. LWCDNet: A lightweight fully convolution network for change detection in optical remote sensing imagery. IEEE Geosci. Remote. Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2022.3159545.

Han, T., Tang, Y., Chen, Y., 2022b. Heterogeneous image change detection based on two-stage joint feature learning. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 3215–3218. http://dx.doi.org/10.1109/IGARSS46834.2022.9883323.

Han, C., Wu, C., Guo, H., Hu, M., Chen, H., 2023a. HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 3867–3878. http://dx.doi.org/10.1109/JSTARS.2023.3264802.

Han, C., Wu, C., Guo, H., Hu, M., Li, J., Chen, H., 2023b. Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 8395–8407. http://dx.doi.org/10.1109/JSTARS.2023.3310208.

Hao Chen, Z.Q., Shi, Z., 2021. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens. 1–14. http://dx.doi.org/10.1109/TGRS.2021.3095166.

Huang, R., Wang, R., Guo, Q., Zhang, Y., Fan, W., 2022. IDET: Iterative difference-enhanced transformers for high-quality change detection. http://dx.doi.org/10.48550/ARXIV.2207.09240, URL https://arxiv.org/abs/2207.09240.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586. http://dx.doi.org/10.1109/TGRS.2018.2858817.

Jing, W., Chi, K., Li, Q., Wang, Q., 2024. 3-D neighborhood cross-differencing: A new paradigm serves remote sensing change detection. IEEE Trans. Geosci. Remote Sens. 62, 1–11. http://dx.doi.org/10.1109/TGRS.2024.3422210.

Jing, W., Yuan, Y., Wang, Q., 2023. Dual-field-of-view context aggregation and boundary perception for airport runway extraction. IEEE Trans. Geosci. Remote Sens. 61, 1–12. http://dx.doi.org/10.1109/TGRS.2023.3271676.

Koller, R., Samimi, C., 2011. Deforestation in the Miombo woodlands: a pixel-based semi-automated change detection method. Int. J. Remote Sens. 32 (22), 7631–7649. http://dx.doi.org/10.1080/01431161.2010.527390.

Lebedev, M., Vizilter, Y., Vygolov, O., Knyaz, V., Rubis, A., 2018. Change detection in remote sensing images using conditional adversarial networks. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLII-2, 565–571. http://dx.doi.org/10.5194/isprs-archives-XLII-2-565-2018.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

Li, K., Cao, X., Meng, D., 2024. A new learning paradigm for foundation model-based remote sensing change detection. IEEE Trans. Geosci. Remote Sens. 1. http://dx.doi.org/10.1109/TGRS.2024.3365825.

Lihe, Z., He, J., Yuan, Q., Jin, X., Xiao, Y., Zhang, L., 2024. PhDnet: A novel physic-aware dehazing network for remote sensing images. Inf. Fusion 106, 102277. http://dx.doi.org/10.1016/j.inffus.2024.102277, URL https://www.sciencedirect.com/science/article/pii/S1566253524000551.

Liu, J., Gong, M., Qin, K., Zhang, P., 2018a. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans. Neural Networks Learn. Syst. 29 (3), 545–559. http://dx.doi.org/10.1109/TNNLS.2016.2636227.

Liu, Z., Li, G., Mercier, G., He, Y., Pan, Q., 2018b. Change detection in heterogenous remote sensing images via homogeneous pixel transformation. IEEE Trans. Image Process. 27 (4), 1822–1834. http://dx.doi.org/10.1109/TIP.2017.2784560.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X., 2021. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geosci. Remote. Sens. Lett. 18 (5), 811–815. http://dx.doi.org/10.1109/LGRS.2020.2988032.

Luppino, L.T., Bianchi, F.M., Moser, G., Anfinsen, S.N., 2019. Unsupervised image regression for heterogeneous change detection. IEEE Trans. Geosci. Remote Sens. 57 (12), 9960–9975. http://dx.doi.org/10.1109/TGRS.2019.2930348.

Lv, Z., Huang, H., Li, X., Zhao, M., Benediktsson, J.A., Sun, W., Falco, N., 2022. Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective. Proc. IEEE 110 (12), 1976–1991. http://dx.doi.org/10.1109/JPROC.2022.3219376.

Niu, X., Gong, M., Zhan, T., Yang, Y., 2019. A conditional adversarial network for change detection in heterogeneous images. IEEE Geosci. Remote. Sens. Lett. 16 (1), 45–49. http://dx.doi.org/10.1109/LGRS.2018.2868704.

Rawashdeh, S.B.A., 2011. Evaluation of the differencing pixel-by-pixel change detection method in mapping irrigated areas in dry zones. Int. J. Remote Sens. 32 (8), 2173–2184. http://dx.doi.org/10.1080/01431161003674634.

Satalino, G., Mattia, F., Balenzano, A., Lovergine, F.P., Rinaldi, M., De Santis, A.P., Ruggieri, S., Nafría García, D.A., Gómez, V.P., Ceschia, E., Planells, M., Toan, T.L., Ruiz, A., Moreno, J., 2018. Sentinel-1 & sentinel-2 data for soil tillage change detection. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 6627–6630. http://dx.doi.org/10.1109/IGARSS.2018.8519103.

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: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 618–626. http://dx.doi.org/10.1109/ICCV.2017.74.

Shao, R., Du, C., Chen, H., Li, J., 2021. SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network. Remote. Sens. 13 (18), http://dx.doi.org/10.3390/rs13183750, URL https://www.mdpi.com/2072-4292/13/18/3750.

Sun, Y., Lei, L., Li, X., Sun, H., Kuang, G., 2021. Nonlocal patch similarity based heterogeneous remote sensing change detection. Pattern Recognit. 109, 107598. http://dx.doi.org/10.1016/j.patcog.2020.107598, URL https://www.sciencedirect.com/science/article/pii/S0031320320304015.

Wan, L., Xiang, Y., You, H., 2019. An object-based hierarchical compound classification method for change detection in heterogeneous optical and SAR images. IEEE Trans. Geosci. Remote Sens. 57 (12), 9941–9959. http://dx.doi.org/10.1109/TGRS.2019.2930322.

Wan, L., Zhang, T., You, H.J., 2018. Multi-sensor remote sensing image change detection based on sorted histograms. Int. J. Remote Sens. 39 (11), 3753–3775. http://dx.doi.org/10.1080/01431161.2018.1448481.

Wang, J.-J., Dobigeon, N., Chabert, M., Wang, D.-C., Huang, T.-Z., Huang, J., 2024a. CD-GAN: A robust fusion-based generative adversarial network for unsupervised remote sensing change detection with heterogeneous sensors. Inf. Fusion 107, 102313. http://dx.doi.org/10.1016/j.inffus.2024.102313, URL https://www.sciencedirect.com/science/article/pii/S1566253524000915.

Wang, Z., Ma, Y., Zhang, Y., 2023. Review of pixel-level remote sensing image fusion based on deep learning. Inf. Fusion 90, 36–58. http://dx.doi.org/10.1016/j.inffus.2022.09.008, URL https://www.sciencedirect.com/science/article/pii/S1566253522001403.

Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R., 2024b. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, No. 6. AAAI, pp. 5805–5813.

Wang, M., Zhu, B., Zhang, J., Fan, J., Ye, Y., 2024c. A lightweight change detection network based on feature interleaved fusion and bistage decoding. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 2557–2569. http://dx.doi.org/10.1109/JSTARS.2023.3344635.

Weismiller, R.A., Kristof, S.J., Scholz, D.K., Anuta, P.E., Momin, S., 1977. Change detection in coastal zone environments. Photogramm. Eng. Remote Sens. 43.

Wen, D., Huang, X., Zhang, L., Benediktsson, J.A., 2016. A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation. IEEE Trans. Geosci. Remote Sens. 54 (1), 609–625. http://dx.doi.org/10.1109/TGRS.2015.2463075.

Ye, Y., Wang, M., Zhou, L., Lei, G., Fan, J., Qin, Y., 2023. Adjacent-level feature cross-fusion with 3-D CNN for remote sensing image change detection. IEEE Trans. Geosci. Remote Sens. 61, 1–14. http://dx.doi.org/10.1109/TGRS.2023.3305499.

Zhan, T., Gong, M., Jiang, X., Li, S., 2018. Log-based transformation feature learning for change detection in heterogeneous images. IEEE Geosci. Remote. Sens. Lett. 15 (9), 1352–1356. http://dx.doi.org/10.1109/LGRS.2018.2843385.

Zhang, H., Ma, G., Zhang, Y., Wang, B., Li, H., Fan, L., 2023. MCHA-Net: A multi-end composite higher-order attention network guided with hierarchical supervised signal for high-resolution remote sensing image change detection. ISPRS J. Photogramm. Remote Sens. 202, 40–68. http://dx.doi.org/10.1016/j.isprsjprs.2023.05.033, URL https://www.sciencedirect.com/science/article/pii/S0924271623001570.

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, URL https://www.sciencedirect.com/science/article/pii/S0924271620301532.

Zhou, W., Troy, A., Grove, M., 2008. Object-based land cover classification and change analysis in the Baltimore Metropolitan Area using multitemporal high resolution remote sensing data. Sensors 8 (3), 1613–1636. http://dx.doi.org/10.3390/s8031613, URL https://www.mdpi.com/1424-8220/8/3/1613.