

# ChangeRD: A Registration-Integrated Change Detection Framework for Unaligned Remote Sensing Images

Wei Jing<sup>a,b</sup>, Kaichen Chi<sup>b</sup>, Qiang Li<sup>b</sup> and Qi Wang<sup>b,\*</sup>

<sup>a</sup>National Elite Institute of Engineering, Northwestern Polytechnical University, Xi'an 710072, China

<sup>b</sup>School of Artificial Intelligence, Optics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

## ARTICLE INFO

### Keywords:

Change detection, registration, remote sensing, deep learning, perspective transformation.

## ABSTRACT

Change Detection (CD) is important for natural disaster assessment, urban construction management, ecological monitoring, etc. Nevertheless, the CD models based on the pixel-level classification are highly dependent on the registration accuracy of bi-temporal images. Besides, differences in factors such as imaging sensors and season often result in pseudo-changes in CD maps. To tackle these challenges, we establish a registration-integrated change detection framework called ChangeRD, which can explore spatial transformation relationships between pairs of unaligned images. Specifically, ChangeRD is designed as a multitask network that supervises the learning of the perspective transformation matrix and difference regions between images. The proposed Adaptive Perspective Transformation (APT) module is utilized to enhance spatial consistency of features from different levels of the Siamese network. Furthermore, an Attention-guided Central Difference Convolution (AgCDC) module is proposed to mine the deep differences in bi-temporal features, significantly reducing the pseudo-change noise caused by illumination variations. Extensive experiments on unaligned bi-temporal images have demonstrated that ChangeRD outperforms other SOTA CD methods in terms of qualitative and quantitative evaluation. The code for this work will be available on GitHub.

## 1. Introduction

Change detection is the process of detecting changes in the target area by interpreting multi-temporal images (Radke et al., 2005; Lu et al., 2004; Liu et al., 2023; Ru et al., 2021; Zheng et al., 2022). With the vigorous development of earth observation technologies, it has greatly facilitated the implementation of tasks such as disaster assessment (Abuelgasim et al., 1999), agricultural investigation (Satalino et al., 2018), and urban planning (Wen et al., 2016). This has progressively assumed a crucial role in monitoring surface environments and human activities. However, the spatial mismatch and intra-class variability of bi-temporal images make change detection on unregistered images a challenging task.

Traditional change detection methods focus on extracting change information based on the spectral differences of pixels, such as difference, ratio and regression analysis methods (Weismiller et al., 1977; Coppin et al., 2004). To further exploit the spectral information in remote sensing images, methods such as change vector analysis (Du et al., 2020), principal component analysis (Deng et al., 2008), and slow feature analysis (Wu et al., 2014) have been introduced into the task of change detection. These methods are easy to implement and have achieved good performance in simple scenes for low-resolution remote sensing images. However, they often fail in more complex and high-resolution scenes due to their neglect of semantic information around the pixels. The advent of object-based methods (Gong et al., 2017) and machine learning classifiers (Wu et al., 2017) has greatly


improved the accuracy of change detection. These methods separate objects in images and compare them after assigning categories using classification algorithms to detect the areas where changes have occurred. However, the limitations of handcrafted features and nonlinear feature mapping of classifiers restrict their application in practical scenarios.

With the wide application of high-resolution remote sensing imaging in the field of earth observation, change detection has flourished and also faced more challenges (Bai et al., 2023; Lei et al., 2022a; Ning et al., 2024a). Firstly, the complex and diverse scenes in VHR images result in a large amount of noise and task-irrelevant information, which interferes with the accuracy of models (Zhang et al., 2020b). Secondly, VHR images exacerbate the intra-class variability caused by the same object having different spectral responses due to imaging season, weather, and other factors (Zhu et al., 2021). Together, these issues lead to an increased risk of pseudo-changes being misdetected, further challenging the accuracy and reliability of change detection (Saha et al., 2019).

In recent years, deep learning has demonstrated powerful representation learning capabilities that enable the extraction of deep semantic information inherent in images (Lecun et al., 2015; Yuan et al., 2022, 2019; He et al., 2016). With the rise of large-scale remote sensing data, deep learning has been leveraged to extract high-dimensional spatio-temporal features from multi-temporal images. These discriminative features have led to the development of various object-level and pixel-level change detection methods (Caye Daudt et al., 2018; Wu et al., 2023; Bandara & Patel, 2022; Lin et al., 2023; Wu et al., 2021). However, these methods, particularly pixel-wise classification models, often overly rely on aligned pairs of images, which are highly sensitive to registration accuracy. In practice, obtaining well-registered pairs of multi-temporal

\*

\*Corresponding author

 wei\_adam@mail.nwpu.edu.cn (W. Jing);

chikaichen@mail.nwpu.edu.cn (K. Chi); liqmgcs@gmail.com (Q. Li);

crabwq@gmail.com (Q. Wang)

images is often challenging, and current approaches that first focus on registration before detection are both inefficient and expensive. Additionally, due to differences in imaging conditions, the same object in multi-temporal images often exhibits different characteristics, making it crucial to distinguish between pseudo-changes in unaligned images and extract highly discriminative difference features in multi-temporal images.

Image registration aims to align remote sensing images acquired at different time periods to ensure precise positioning within a unified coordinate system, where pixels in the registered images correspond accurately to their geographic locations (Thevenaz & Unser, 2000). Traditional remote sensing image registration methods involve a complex series of steps, including atmospheric correction, projection transformation, feature matching, and spatial transformation (Chang et al., 2019). In this paper, the registration process is explicitly integrated into a comprehensive change detection framework ChangeRD, introducing only a minimal number of additional parameters. First, we simulate an unregistered state by randomly offsetting the vertices of the registered bi-temporal images and calculating the perspective transformation matrix to warp the entire image. Subsequently, ChangeRD learns to predict the vertex offsets during the optimization process. The adaptive perspective transformation (APT) module then spatially transforms the input images and feature groups at various levels to align the bi-temporal images. Notably, we ensure the robustness of this process through dual supervision, including the vertex offset loss and transformation loss of the input images. Furthermore, considering the pseudo-changes caused by the same-object-different-spectrum phenomenon, we propose an Attention-guided Central Difference Convolution (AgCDC) module. By generating local difference features within the image, this module can mitigate the impact of spectral variation to some extent and reduce feature discrepancies between bi-temporal features caused by the same-object-different-spectrum phenomenon. To evaluate the performance of ChangeRD on unregistered image pairs, we designed a series of experiments on multiple public datasets. Compared to other algorithms, ChangeRD greatly reduces the dependency on registration accuracy for change detection, demonstrating superior performance. The main contributions of this work are summarized as follows:

- 1) An adaptive perspective transformation (APT) module is constructed, capable of spatially transforming intermediate layer features within the siamese network, thereby mitigating the impact of registration accuracy on change detection tasks.
- 2) Inspired by central difference convolution, we designed an Attention-guided Central Difference Convolution (AgCDC) module that extracts intrinsic difference information from bi-temporal images, alleviating the impact of pseudo-changes caused by illumination variations.
- 3) A change detection framework with an integrated registration process is proposed, which is capable of not only performing spatial transformations on features at various levels within the model but also mitigating the impact of the same-object-different-spectrum phenomenon through the AgCDC module.

- 4) Extensive experiments demonstrate that ChangeRD effectively addresses the challenges posed by poor registration, surpassing other SOTA change detection algorithms.

## 2. Related work

### 2.1. Remote Sensing Image Registration

Due to differences in sensors, platforms, earth surface changes, and observation angles, the position and shape of surface features in bi-temporal remote sensing images often vary. To improve the accuracy and reliability of remote sensing data, supporting high-precision remote sensing application research, image registration is typically required (Zhu et al., 2018; Li et al., 2023).

Traditional registration algorithms can be categorized into two main types: region-based (Chen et al., 2003; Liang et al., 2014; Suri & Reinartz, 2010) and feature-based (Chang et al., 2019; Lowe, 2004). The former relies on pre-established similarity measures and transformation models. It calculates image similarities based on shallow information such as image intensity or phase and adjusts the transformation model parameters through an optimization strategy to achieve an ideal similarity threshold, ultimately realizing high-precision image alignment. Suri & Reinartz (2010) proposed an automatic registration algorithm for dense urban scenes based on a two-step procedure using mutual information histograms, which calculates local deformations to achieve fine image registration. Although such methods are easy to understand and can preserve the detailed information of the original image, their accuracy heavily depends on similarity measures, and they suffer from low registration efficiency and poor real-time performance. Feature-based methods, on the other hand, first extract and match geometric features such as points, lines, and planes in images and use the spatial correspondence of these geometric features to solve the transformation parameters of the image model. A widely used feature matching algorithm is the Scale Invariant Feature Transform (SIFT) operator (Lowe, 2004). This algorithm extracts points invariant to changes in lighting, noise, and other factors by searching for them across different spatial feature scales. Models based on geometric mapping relationships are suitable for the registration tasks of images with large geometric deformations, but they suffer from issues such as error accumulation and poor registration accuracy (Feng et al., 2021).

The above-mentioned traditional registration algorithms struggle to fully utilize the deep semantic information of images, resulting in limited adaptability to various scenes. In recent years, numerous studies have attempted to apply deep learning to the field of image registration (Lee et al., 2021; Zampieri et al., 2018; Girard et al., 2019). Based on their application forms, deep learning-based algorithms can be primarily divided into: 1) Constructing similarity measures using deep learning and optimizing transformation models through reinforcement learning, or directly solving image displacement fields using end-to-end networks. For instance, Zampieri et al. (2018) used Fully Convolutional Network (FCN) to extract scale-invariant features from images and

estimated the global displacement field for multi-modal remote sensing image registration through a cascading network. Girard et al. (2019) predicted displacement fields and pixel-level building segmentation via cascading FCN, achieving coarse-to-fine registration. 2) The second approach involves embedding deep learning into traditional feature-based registration frameworks, replacing conventional feature extraction, description, and matching methods. (Verdie et al., 2015; DeTone et al., 2018; Zeng et al., 2021). DeTone et al. (2018) proposed a self-supervised method to extract image intrinsic features that are insensitive to lighting and imaging angles. Zeng et al. (2021) treated registration and stitching as a unified task. Compared to SIFT-based registration methods, the Siamese-network-based methods showed significant improvements in both accuracy and efficiency.

## 2.2. Deep Learning-Based Change Detection

The emergence of numerous deep learning-based change detection models has been facilitated by the availability of a large volume of annotated remote sensing data, thanks to the advancements in remote sensing big data (Chi et al., 2023; Jing et al., 2023). Deep learning, compared to traditional change detection algorithms, has a more robust feature representation capability. It captures both low-level details and high-level semantics in images, enabling adaptation to a variety of complex scenarios. Broadly speaking, single-branch and dual-branch structures are common deep learning-based change detection frameworks.

Structures with a single branch utilize strategies like differencing or concatenation prior to extracting features to merge bitemporal images, addressing change detection as a segmentation issue. The integration of an enhanced ConvLSTM within the U-Net architecture for seamless end-to-end change detection was achieved by Sun et al. (2022), which also employed Atrous convolution to explore multi-scale spatial details. Lin et al. (2023) approached change detection as analogous to understanding videos, focusing on the spatiotemporal interplay within the encoder and introduced the P2V-CD model to separate the spatiotemporal aspects of bitemporal images. Ning et al. (2024b) proposed a novel Multi-Stage Progressive Change Detection Network, which addresses the issues of class imbalance and cross-domain differences in optical remote sensing images through stable invariant region detection, knowledge distillation, and a coarse-to-fine change detection structure. Dual-branch structures opt for a late fusion approach using siamese networks as a typical model, which isolates distinctive features from paired images for change detection (Zhang et al., 2020a; Lei et al., 2022b). To improve the definition and consistency of objects within the resulting change maps, Zhang et al. (2020a) devised a deeply supervised network that discriminates differences, merging deep features from initial inputs with bi-temporal difference features through an attention mechanism to reconstruct the change maps. Addressing the detection of non-pertinent changes, Lei et al. (2022b) leveraged a siamese network to discern contrasts between foreground and background, also incorporating distant interactions to amplify the clarity and

coherence of the changes. Recently, the transformer architecture has shown exceptional promise in computer vision, surpassing CNN-based approaches in areas like classification and detection (Dosovitskiy et al., 2020). Its primary self-attention mechanism allows the capturing of interrelations across different image regions, thereby depicting global dependencies (Liu et al., 2021b). Bandara & Patel (2022) integrated a hierarchical Transformer structure into siamese networks, which efficiently portrayed the multi-scale distant details essential for precise change detection and achieved state-of-the-art results on various CD datasets. Zheng et al. (2024) proposed a deep probabilistic change model (DPCM) that integrates probabilistic graphical models with deep neural networks to address various change detection tasks in a unified manner. The introduction of the sparse change transformer significantly reduces the computational complexity for high-resolution images.

Despite substantial progress and notable achievements in image registration and change detection, the integration of these tasks into a cohesive framework remains incomplete. Prompted by this gap, this paper investigates the combination of image registration and change detection, proposing the ChangeRD model to tackle this challenge.

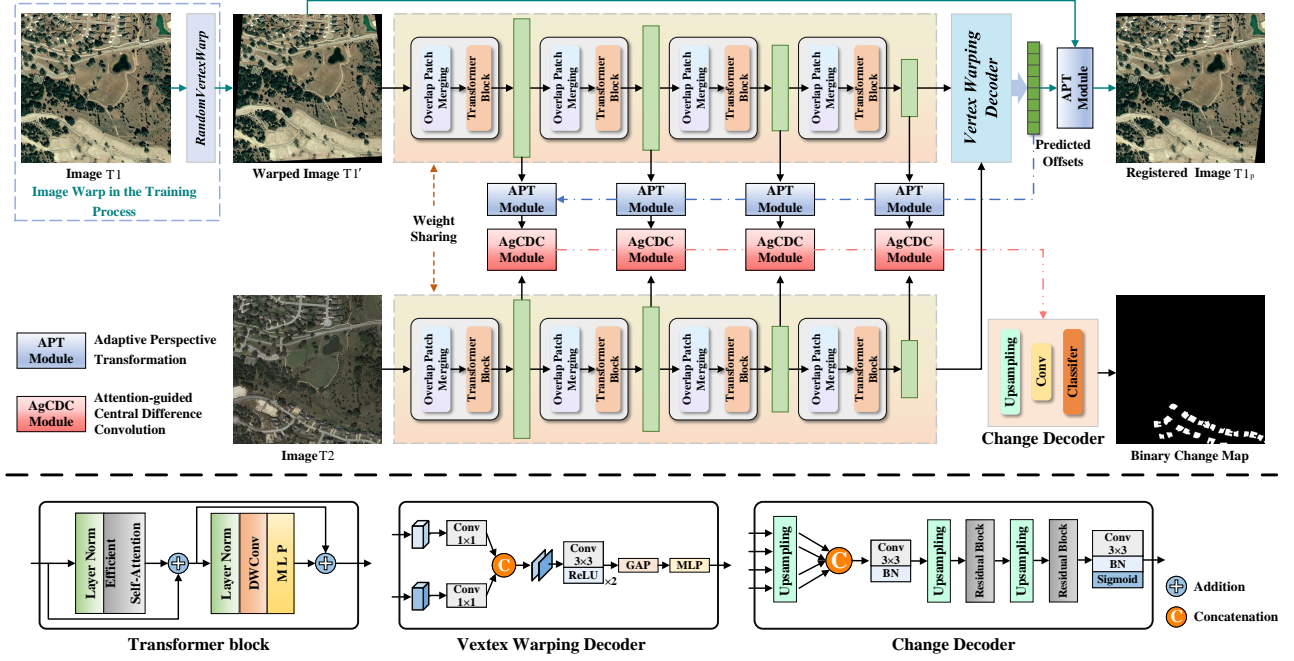
## 3. Methodology

### 3.1. Overall Framework

Extensive research in transfer learning has convincingly demonstrated that the features learned by deep models on one task can be effectively repurposed for a variety of other tasks. This stems from the deep abstract features' ability to encapsulate essential characteristics of the data that are relevant across multiple contexts (Niu et al., 2020; Lu et al., 2022). Change detection, which focuses on identifying differences between images over time, and registration, which aims to align multiple images from different times, sensors, or perspectives to a common reference frame, both fundamentally rely on the identification of key features like edges, textures, and shapes. Consequently, the features output by the backbone network, especially the deep abstract features that convey a comprehensive understanding of the image content, can be applied not only to enhance change detection but also to significantly improve registration tasks. Guided by this theory, an integrated change detection framework incorporating registration is constructed, as illustrated in Fig. 1.

To learn the spatial transformation relationship of unregistered images, during the training phase, we simulate the unregistered state by warping the four vertices of aligned images and performing perspective transformation. Regarding the extraction of features, the Siamese Vision Transformer network is used for capturing both shallow details and deep semantic information from the bi-temporal images. These spatiotemporal features encapsulate the intrinsic information of the bi-temporal images, contributing to varying degrees to image registration and change detection. For spatial registration, the deepest features are utilized to predict the offsets of the bi-temporal images vertices through the Vertex Warping





**Figure 1:** Overall Framework of ChangeRD. The unaligned bi-temporal images are subjected to feature-level and global image registration by the APT module, and the difference information is extracted using the AgCDC module.

Decoder. Specifically, we first use linear convolution and concatenation operations to preliminarily fuse the bi-temporal features, which can be expressed as:

$$X_f = (\mathcal{K}_{1 \times 1} * X_{t1}) \parallel (\mathcal{K}_{1 \times 1} * X_{t2}), \quad (1)$$

where  $\mathcal{K}$  is convolution kernel,  $*$  denotes convolution calculation,  $\parallel$  denotes concatenation operation, and  $X_t$  and  $X_f$  are the features before and after fusion, respectively. Following this, a subsequent stage involves applying two convolutional layers to project the fused feature  $X$  into a higher-dimensional space, while incorporating the rectified linear unit (ReLU) activation function to introduce nonlinearity. The above can be defined as follows:

$$X'_f = \delta(\mathcal{K}_{3 \times 3} * \delta(\mathcal{K}_{3 \times 3} * X_f)), \quad (2)$$

where  $\delta$  refers to the ReLU function. Finally, the vertex offsets of bi-temporal images are predicted using global average pooling (GAP) and multi-layer perception machine (MLP):

$$OFF_p = \mathcal{F}_{MLP}(\mathcal{F}_{GAP}(X'_f)), \quad (3)$$

where  $OFF_p$  represents the predicted vertex offsets. Four matching pairs of points are generated from the predicted offsets to initialize an Adaptive Perspective Transform (APT) module, and then the warped images T1' are aligned with T2. It is worth noting that during the registration stage, we adopt Mean Squared Error (MSE) loss and L1 loss functions to supervise the predicted vertex offsets and the registration accuracy of the bi-temporal images, respectively. The offset loss is defined as follows:

$$\mathcal{L}_{off} = \frac{1}{8} \sum_{i=1}^8 (OFF_y^i - OFF_p^i)^2, \quad (4)$$

where  $OFF_y$  and  $OFF_p$  are the true and predicted vertex offsets, respectively. The registration loss is defined as follows:

$$\mathcal{L}_{reg} = \frac{1}{m} \sum_{i=1}^m |T1^i - T1_p^i|, \quad (5)$$

where T1 and T1<sub>p</sub> are the unwarped image and the predicted registered image, respectively.

For change detection, features of different scales bring varying degrees of improvement to the model performance. Deep semantic features are beneficial for determining whether the current target has changed, while shallow details can help refine the edges of change maps. To unify the geospatial aspects of bi-temporal features, the APT module is used to perform feature-level alignment of multi-scale features as follows:

$$\tilde{X}_{t1}^i = \mathcal{T}^i(X_{t1}^i), \quad (6)$$

where  $i$  is the index of different scale features and  $\mathcal{T}$  denotes APT.

To mitigate the impact of spectral differences caused by factors such as varying illumination and vegetation growth status at different times, we design the Attention-guided Central Difference Convolution (AgCDC) module to extract the difference information between the registered bi-temporal features.

$$X_d^i = D^i(\tilde{X}_{t1}^i, X_{t2}^i), \quad (7)$$

where  $D$  denotes the difference calculation. Finally, the difference information is input into the change decoder to predict

binary change maps. Specifically, features of different scales are first upsampled to the same resolution using transposed convolutions, followed by concatenation and convolutional layers for preliminary fusion of multi-scale features. The process is defined as follows:

$$X_d = B(\mathcal{K}_{3 \times 3} * (\mathcal{U}(X_d^4) \parallel \mathcal{U}(X_d^3) \parallel \mathcal{U}(X_d^2) \parallel X_d^1)), \quad (8)$$

where  $B$  denotes batch normalization and  $\mathcal{U}$  indicates the upsampling. To fully explore the bi-temporal difference information, the combination of transpose and residual blocks is adopted to further learn the fused multi-scale difference features. During this process, the difference features are upsampled to their initial resolution as follows:

$$\tilde{X}_d = \mathcal{R}(\mathcal{U}(\mathcal{R}(\mathcal{U}(X_d)))), \quad (9)$$

where  $\mathcal{R}$  denotes residual block. In the end, a 3x3 convolution and a softmax activation function are used as a classifier to generate the binary change maps of the bi-temporal images. The cross-entropy loss is adopted to supervise the change detection task as follows:

$$\mathcal{L}_{cd} = - \sum_{c=1}^M (y_c \log(p_c)), \quad (10)$$

where  $M$  indicates the number of categories,  $c$  is the category index,  $y_c$  represents the probability of the true class being  $c$ , and  $p_c$  represents the probability of the predicted class being  $c$ . To sum up, the overall loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_{cd} + \alpha \mathcal{L}_{off} + \beta \mathcal{L}_{reg}, \quad (11)$$

where  $\alpha$  and  $\beta$  are the balancing factors, which are empirically set to 0.001 and 0.5, respectively.

### 3.2. Adaptive Perspective Transformation

High registration of bi-temporal images is a prerequisite for high-precision change detection of surface features, especially for pixel-level detection models. However, the cumbersome process of registering first and then detecting significantly increases the cost. Perspective transformation can achieve spatial registration of bi-temporal images by calculating the transformation matrix through solving a set of linear equations. Although its registration effect is limited by linear constraints, the perspective transformation algorithm based on linear interpolation is easy to solve and can be performed at high speed on Graphics Processing Units (GPUs), meeting the requirements of change detection models for registration accuracy. Based on the above considerations, we have developed an Adaptive Perspective Transformation (APT) module embedded in ChangeRD, which completes the spatial registration of input images and multi-scale features during the forward inference process.

The perspective transformation algorithm typically requires four pairs of matched feature points to calculate the transformation matrix  $H$ . However, for image pairs with dramatic scene changes, it is often difficult to extract robust

#### Algorithm 1 Perspective Transformation Matrix Calculation

**Require:**  $src, dst$

**Ensure:** Perspective transformation matrix  $H$

```

1:  $bs, \_, \_ \leftarrow src.size()$ 
2: Initialize  $A \in \mathbb{R}^{bs \times 8 \times 8}$ 
3: for  $i = 1$  to  $bs$  do
4:   for  $j = 0$  to 3 do
5:      $A_{i,2j,0:3} \leftarrow [src_{i,j,0}, src_{i,j,1}, 1]$ 
6:      $A_{i,2j+1,3:6} \leftarrow [src_{i,j,0}, src_{i,j,1}, 1]$ 
7:      $A_{i,2j,6:8} \leftarrow -dst_{i,j,0} \cdot src_{i,j,:}$ 
8:      $A_{i,2j+1,6:8} \leftarrow -dst_{i,j,1} \cdot src_{i,j,:}$ 
9:   end for
10: end for
11: Reshape  $dst$  into a  $bs \times 8 \times 1$  matrix  $B$ 
12: Solve the linear system  $Ah = B$  for  $h$ 
13:  $h \leftarrow Concatenate(h, \{1\})$ 
14: Reshape  $h$  into a  $bs \times 3 \times 3$  matrix  $H$ 
15: return  $H$ 
```

matching features. We cleverly circumvent this problem by predicting the offsets of the four vertices of the pre-temporal images T1 and post-temporal images T2. The offsets allow us to directly obtain the source points  $src$  and target points  $dst$ . We then constructed an algorithm for calculating the transformation matrix  $H$  from the  $src$  and  $dst$ , which can run efficiently on GPUs. As shown in **Alg. 1**, given  $src$  and  $dst$ , this algorithm constructs the coefficient matrix  $A$  and target matrix  $B$ , then solves the linear system  $Ah = b$  for  $h$ , reshaping it into a 3x3 matrix  $H$ . More specifically, for each sample in the batch, the algorithm initializes  $A$ , fills the first six columns of each row with the horizontal and vertical coordinates of four source points, and the last two columns with the product of the corresponding horizontal coordinate in  $dst$  and the entire row of the corresponding source point. The algorithm then reshapes  $dst$  and  $B$ , and solves the linear system  $Ah = b$  for  $h$ . Finally,  $h$  is reshaped into the general form of the transformation matrix, a 3x3 matrix  $H$ , and returned. Note that for features of different resolutions, the APT module achieves transformation matrices with varying parameters by adaptively scaling the offsets and adjusting the vertex coordinates.

After obtaining the transformation matrix  $H$ , we implemented a batch perspective transformation algorithm within the APT module as **Alg. 2**. Specifically, we first obtain the shape of the input tensor  $T1$  and the batch size  $bs$ . Next, an input grid  $G_i$  is created to perform operations with each tensor input. Then, the perspective transformation matrix  $H$  is multiplied by  $G_i$  to obtain the output grid  $G_o$ . We normalize  $G_o$  to adapt to the dimensions of the target tensor. Finally, using the bilinear interpolation grid sampling method, we transform  $T1$  into the registered tensor  $T1'$  based on  $G_o$ . Through the aforementioned operations, the input images and features at different levels can be efficiently and adaptively registered.

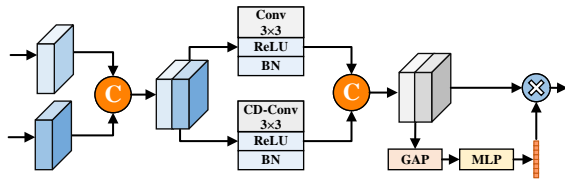
**Algorithm 2** Batch Perspective Transformation**Require:**  $T1, H$ **Ensure:** Transformed tensor  $T1'$ 

```

1:  $bs, height, width \leftarrow T1.shape$ 
2: Construct the input grid  $Gi$  with shape  $(bs, height * width, 3)$ 
3: for  $i = 1$  to  $bs$  do
4:    $Go_{i,:,:} \leftarrow H_{i,:,:} \cdot Gi_{i,:,:}$ 
5:    $Go_{i,:2,:} \leftarrow Go_{i,:2,:} / Go_{i,2,:}$ 
6: end for
7: Reshape  $Go_{i,:2,:}$  into a  $(bs, height, width, 2)$  matrix  $Go'$ 

8:  $Go'_{...,0} \leftarrow Go'_{...,0} / ((width - 1) / 2) - 1$ 
9:  $Go'_{...,1} \leftarrow Go'_{...,1} / ((height - 1) / 2) - 1$ 
10:  $T1' \leftarrow Grid\_Sample(T1, Go')$ 
11: return  $T1'$ 

```

**Figure 2:** Structure of the proposed AgCDC module.**3.3. Attention-guided Central Difference Convolution**

Bi-temporal images can exhibit different spectral or textural characteristics of the same land cover due to factors such as illumination and seasonality over a long time interval. One of the major challenges in change detection tasks is mitigating the impact of such spectral differences and eliminating pseudo-changes caused by differences in imaging times. Previous methods have largely relied on differencing or convolution to extract difference information from bi-temporal features, but they have struggled to capture the nuanced anti-counterfeiting essence of detailed information. For this, we propose an approach that leverages the spatial differencing properties of Central Difference Convolution (CDConv) (Yu et al., 2020) to extract essential discriminative information from bi-temporal features. By computing difference information with neighboring pixels, CDConv is able to capture discriminative features that are robust to factors such as illumination and seasonality. The general form of CDConv is as follows:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)), \quad (12)$$

where  $x, y$  are the input and output feature maps respectively,  $\mathcal{R}$  denotes all the positions that can be covered by the convolution operation,  $p_0$  denotes current location on both input and output feature maps while  $p_n$  enumerates the locations in  $\mathcal{R}$ .

Although central differencing features of the same land cover are not affected by changes in illumination, they can easily lose the semantic information of the land cover itself. To address this problem, we combine CDConv with general convolution in parallel and use an attention mechanism to adaptively guide the model to focus on features that are more conducive to differentiating land cover changes. The designed Attention-guided Central Difference Convolution (AgCDC) module is shown in Fig. 2. The bi-temporal features are first fused into a feature group through concatenation. Next, we use parallel  $3 \times 3$  Conv and CDConv to respectively extract semantic and illumination-invariant features of changing land cover, and concatenate them into  $F$ :

$$X = X_{t1} \parallel X_{t2},$$

$$F = B(\delta(\mathcal{K}_{3 \times 3} * X)) \parallel B(\delta(\mathcal{K}_{3 \times 3}^{cdc} * X)), \quad (13)$$

where  $\mathcal{K}^{cdc}$  is central difference convolution kernel. The attention mechanism is used to enhance important features in  $F$  while suppressing unimportant ones. Specifically,  $F$  is first aggregated into a one-dimensional vector  $V$  through global average pooling (GAP) as follows:

$$V = \mathcal{F}_{GAP}(F). \quad (14)$$

Then, MLP is trained to learn the importance of different feature channels, resulting in a weight vector  $\tilde{V}$ .

$$\tilde{V} = \mathcal{F}_{MLP}(V). \quad (15)$$

Finally, the weight vector  $\tilde{V}$  is used to perform channel-wise multiplication with the fused feature  $F$  to generate the output difference feature  $\tilde{F}$  as follows:

$$\tilde{F} = F \otimes \tilde{V}, \quad (16)$$

where  $\otimes$  denotes channel-wise multiplication.

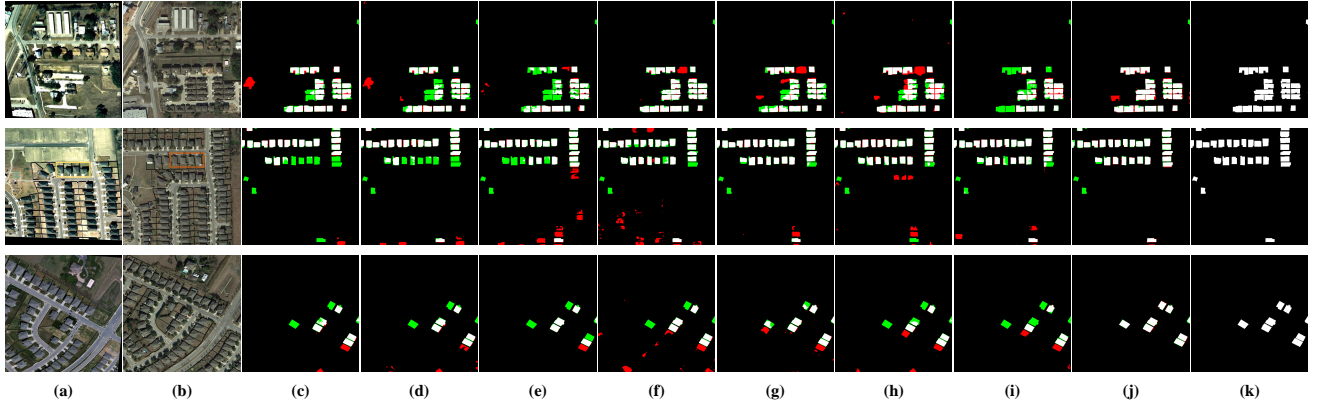
**4. EXPERIMENTS AND DISCUSSION**

To evaluate the performance of the proposed method, a set of experiments was carried out in this section using three openly accessible datasets: LEVIR-CD (Chen & Shi, 2020), WHU-CD (Ji et al., 2019), and SVCD (Lebedev et al., 2018).

**4.1. Dataset Description**

1) LEVIR-CD: This dataset comprises 637 images with a very high resolution of 0.5 meters per pixel. Each image has dimensions of  $1024 \times 1024$  pixels. These bi-temporal images exhibit significant land cover changes, particularly noticeable in the growth of buildings, over a time span ranging from 5 to 14 years. The dataset consists of 31,333 change instances in total. We use standardized training, validation, and testing sets. During model training, validation, and testing, we extracted non-overlapping patches from the original images using the sliding window approach, with each patch having dimensions of  $512 \times 512$ .

2) WHU-CD: This dataset comprises two aerial images, measuring  $32507 \times 15354$  in dimensions, with a spatial resolution of 0.3 meters per pixel. Following a similar approach as



**Figure 3:** The qualitative comparison of different methods on the LEVIR-CD dataset, the yellow boxes show the obvious wrong segmentation. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) FC-Siam-conc. (d) FC-Siam-diff. (e) DSIFN. (f) SNUNet. (g) DTCDCSCN. (h) BIT. (i) ChangeFormer. (j) Ours. (k) Groud truth. 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. Please zoom in for a better view.

used in the LEVIR-CD dataset, we extracted sample patches from the original images using the sliding window strategy with non-overlapping patches. The training, validation, and testing sets are distributed in an 8:1:1 ratio, resulting in 1536, 192, and 192 samples, respectively. It is worth noting that WHU-CD has a relatively smaller sample size compared to other datasets, which provides a better opportunity to evaluate the ability of models to combat overfitting.

3) SVCD: This dataset comprises 16,000 pairs of authentic seasonal change remote sensing images obtained from Google Earth. Each pair of images is composed of pixels with dimensions of  $256 \times 256$  and varying spatial resolutions ranging from 3 to 100 cm per pixel. The dataset is divided into training, validation, and testing sets, with 10,000, 3,000, and 3,000 samples, respectively. In contrast to the WHU-CD dataset, the SVCD dataset offers a larger sample size, enabling a comprehensive evaluation of their fitting capability. Moreover, since the majority of the bi-temporal image pairs originate from different seasons, this dataset serves as a valuable benchmark to assess the robustness of models against the same-object different-spectrum phenomena induced by seasonal factors.

## 4.2. Experiment Setup

The experiments were conducted on a server equipped with four NVIDIA GTX 3090 GPUs (24GB VRAM) running the Ubuntu 18.04 operating system. To ensure a fair comparison, all models were trained for 400 epochs using publicly available code and the AdamW optimizer. A learning rate schedule with linear decay was employed, starting from an initial value of  $1e-4$  and dynamically adjusted throughout the training process. All comparison models were retrained on the unaligned datasets with identical hyperparameters. Furthermore, a batch size of 8 was used for the LEVIR and WHU datasets, while a batch size of 32 was applied for the SVCD dataset due to its smaller image sizes. During the inference process, the images are input into the model only after normalization, with the batch size maintained at 1.

## 4.3. Evaluation Metrics

The broadly used criteria, Intersection over Union (IoU), Precision (P), Recall (R), and F1-score (F1), are applied for quantitative evaluation of change detection.

**IoU** is a metric defined as the ratio of the intersection of predicted and ground truth regions to their union. It is calculated as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}}, \quad (17)$$

where TP, TN, FP, and FN refer to the true positive, true negative, false positive, and false negative pixels, respectively.

**Precision** refers to the probability of correctly predicted positive samples (TP) out of all samples predicted as positive (TP+FP):

$$P = \frac{\text{TP}}{\text{FP} + \text{TP}}. \quad (18)$$

**Recall** is defined as the proportion of correctly predicted positive samples (TP) out of all positive samples (TP+FN) as follows:

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (19)$$

**F1** is the harmonic mean of precision and recall, which provides a comprehensive measure of overall performance:

$$F1 = 2 \times \frac{P \times R}{P + R}. \quad (20)$$

## 4.4. Comparison with Advanced Methods

To evaluate the superiority of ChangeRD, we compare it with 9 SOTA algorithms. The brief descriptions of the comparative algorithms are as follows:

- 1) FC-Siam-Conc (Caye Daudt et al., 2018): A fully convolutional Siamese network based on U-Net architecture that extracts difference information from multi-scale bi-temporal features using convolutional layers.



Quantitative Results on the LEVIR-CD. The Best Results are Marked in Bold, the 2nd-Best is Marked with Underline.

Method	Evaluation Metrics						
	P (%)	R(%)	F1(%)	IoU(%)	Params(M)	FLOPs(G)	GPU Memory(MB)
FC-Siam-conc	88.58	82.86	85.62	74.86	1.55	<u>21.32</u>	<u>262.44</u>
FC-Siam-diff	88.01	82.51	85.17	74.18	<b>1.35</b>	<b>18.91</b>	<b>247.69</b>
CDNet	74.62	79.79	77.12	62.76	<u>1.43</u>	93.90	391.25
DSIFN	87.48	82.52	84.93	73.80	35.73	329.02	902.40
SNUNet	79.23	85.85	82.41	70.08	10.20	177.51	842.03
DTCDCSCN	<u>90.74</u>	<u>86.37</u>	<u>88.50</u>	<u>79.38</u>	31.26	52.90	287.48
BIT	88.04	85.80	86.91	76.84	3.50	42.53	159.45
ChangeFormer	90.30	<u>86.69</u>	88.46	79.31	41.03	811.15	1331.78
P2V-CD	76.86	79.34	78.08	64.04	5.42	131.83	591.95
ChangeRD	<b>92.92</b>	<b>90.08</b>	<b>91.48</b>	<b>84.30</b>	29.47	73.10	375.23

- 2) FC-Siam-Diff (Caye Daudt et al., 2018): A fully convolutional Siamese network based on U-Net architecture that computes the absolute difference of multi-scale bi-temporal features as change-related features.
- 3) CDNet (Alcantarilla et al., 2018): This algorithm analogizes change detection to semantic segmentation and designs an efficient CNN architecture based on the U-Net to directly detect changes between image pairs.
- 4) DSIFN (Zhang et al., 2020a): Fusion of multi-level deep features and image difference features using attention mechanisms, improving the integrity of change map boundaries and the compactness within regions.
- 5) SNUNet (Fang et al., 2022): A Siamese network based on NestedUNet that aggregates and refines features from multiple semantic levels, suppressing semantic gaps and localization errors to some extent.
- 6) DTCDCSCN (Liu et al., 2021a): Combining semantic segmentation and change detection tasks to extract discriminative features of land cover, while introducing dual attention modules to better utilize channel and spatial information.
- 7) BIT (Chen et al., 2022): Effective modeling of contextual information in the spatial-temporal domain using transformers to accurately predict changes in bi-temporal features.
- 8) ChangFormer (Bandara & Patel, 2022): Hierarchical transformer encoders are used to extract bi-temporal features, and a feature difference module is designed to compute feature discrepancies at different scales.
- 9) P2V-CD (Lin et al., 2023): Interpreting change detection as a video understanding problem, it constructs pseudo-translational videos with rich temporal information based on input image pairs to mine temporal changes.

**Comparison Experiments on the LEVIR-CD:** Fig. 3 shows the qualitative results of various algorithms on the LEVIR-CD dataset, demonstrating the superior detection performance of ChangeRD. In row 1 of Fig. 3, after undergoing perspective transformation, the objects in the pre-temporal image are

**Table 1**

Quantitative Results on the WHU-CD. The Best Results are Marked in Bold, the 2nd-Best is Marked with Underline.

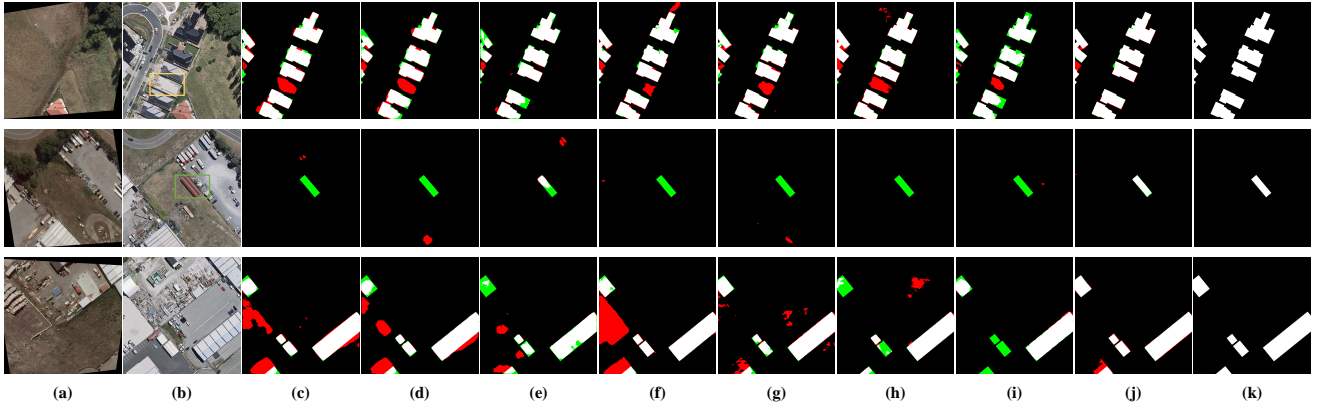
Method	Evaluation Metrics			
	P (%)	R(%)	F1(%)	IoU(%)
FC-Siam-conc	85.59	74.21	79.50	65.97
FC-Siam-diff	81.97	72.02	76.67	62.17
CDNet	<u>93.37</u>	56.23	70.19	54.07
DSIFN	91.77	64.36	75.66	60.84
SNUNet	25.12	78.83	38.09	23.53
DTCDCSCN	90.49	<u>82.09</u>	<u>86.08</u>	<u>75.57</u>
BIT	90.89	<u>81.33</u>	<u>85.84</u>	<u>75.20</u>
ChangeFormer	91.14	80.21	85.33	74.41
P2V-CD	87.34	71.42	78.58	64.72
ChangeRD	<b>95.92</b>	<b>88.73</b>	<b>92.18</b>	<b>85.50</b>

distorted to a certain extent, resulting in significant missed detection of changed building instances, indicated by green pixels in the change map. In row 2, the bi-temporal images mainly display spatial translation differences. Most comparative algorithms exhibit false negatives in the orange box area due to the overlap between the buildings in the yellow box of the pre-temporal phase and the orange box of the post-temporal phase, causing the algorithms to mistakenly classify this area as unchanged. Similarly, at the bottom of the scene, there are some false positives caused by the misinterpretation of vertically arranged buildings after vertical translation, impacting the performance of comparative algorithms. Row 3 focuses on detecting a small number of changed buildings within a dense building cluster. In comparison to other methods, ChangeRD remains unaffected by changes in illumination and demonstrates minimal occurrences of false negatives and false positives in this particular scene.

Table ?? displays the quantitative results, demonstrating the exceptional performance of ChangeRD across all evaluation metrics. Our method achieves an F1 score of 91.48% and an IoU of 84.30%, outperforming the second-best method, DTCDCSCN, by 2.98 and 4.92 percentage points, respectively.

**Comparison Experiments on the WHU-CD:** Fig. 4 presents the visual comparison of different methods. Owing to the spectral similarity between the impervious surface within the yellow rectangle and the roof, competitors erroneously classified this area as newly constructed buildings.





**Figure 4:** The qualitative comparison of different methods on the WHU-CD dataset, the yellow boxes show the obvious wrong segmentation. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) FC-Siam-conc. (d) FC-Siam-diff. (e) DSIFN. (f) DTCDSN. (g) BIT. (h) ChangeFormer. (i) P2V-CD. (j) Ours. (k) Groud truth. 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. Please zoom in for a better view.

**Table 2**

Quantitative Results on the SVCD. The Best Results are Marked in Bold, the 2nd-Best is Marked with Underline.

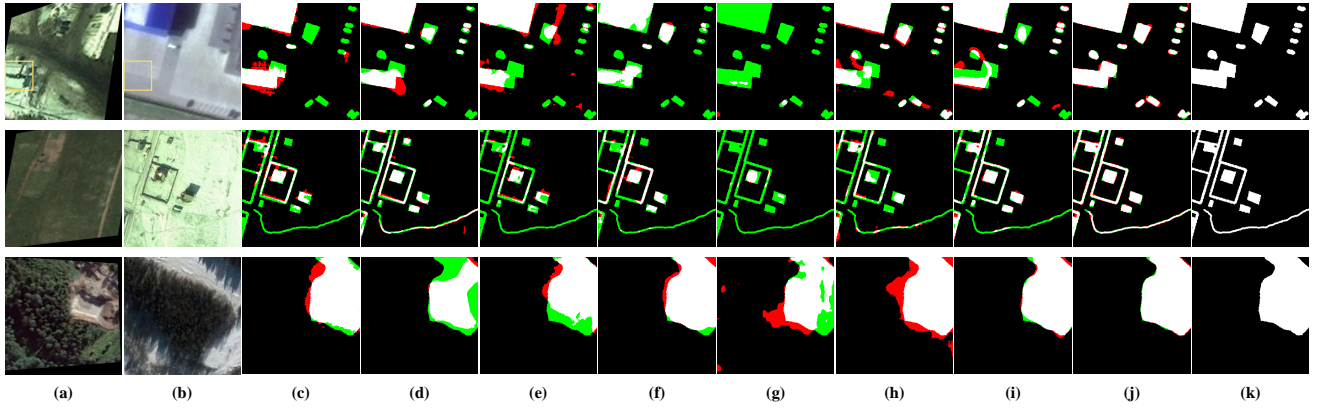
Method	Evaluation Metrics			
	P (%)	R(%)	F1(%)	IoU(%)
FC-Siam-conc	66.15	40.83	50.49	33.77
FC-Siam-diff	85.34	55.42	67.20	50.60
CDNet	86.39	70.55	77.67	63.49
DSIFN	94.17	82.39	87.88	78.39
SNUNet	85.04	78.49	81.63	68.97
DTCDSN	94.96	82.41	88.24	78.96
BIT	83.36	65.67	73.46	58.06
ChangeFormer	89.98	<u>87.76</u>	<u>88.86</u>	<u>79.95</u>
P2V-CD	91.10	80.52	85.48	74.65
ChangeRD	<b>95.90</b>	<b>95.57</b>	<b>95.73</b>	<b>91.82</b>

Furthermore, in comparison to other methods, ChangeRD exhibits superior precision in detecting building edges. In the second row of Fig. 4, the majority of methods failed to identify the building within the green rectangle due to its similarity to nearby vans. Our method accurately identifies and avoids misclassifying the building, even in areas with a high concentration of vehicles. The background of the input images in the third row is relatively cluttered, and after warping, the original land covers do not align properly. Consequently, the detection results of comparative algorithms exhibit numerous false positives. In these three typical scenarios, ChangeRD consistently outperforms its competitors.

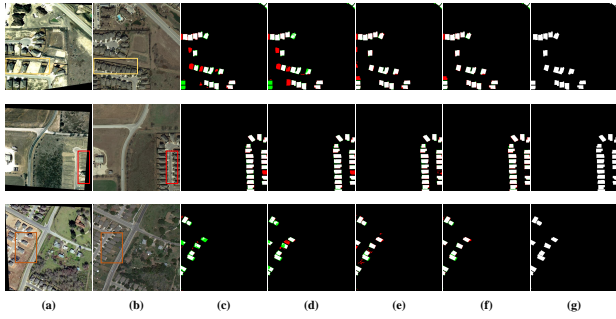
The quantitative results are presented in Table 1. Due to the relatively homogeneous scenes and limited data volume of the WHU-CD dataset, all methods demonstrated varying degrees of overfitting, with SNUNet being the most severely impacted. In contrast, ChangeRD exhibited superior generalization performance compared to the comparative methods, establishing a substantial lead in both the F1 and IoU composite metrics. Specifically, ChangeRD surpassed the second-best method by 6.1% in terms of F1 and 9.93% in terms of IoU.

**Comparison Experiments on the SVCD:** Differing from the other two datasets, the SVCD dataset consists of patches with a size of  $256 \times 256$ . When applying the same offsets to distort the pre-temporal images, the algorithms encounter more pronounced distortions on this dataset. Additionally, the SVCD dataset encompasses not only building changes but also variations in roads, vehicles, and vegetation. The images in this dataset possess lower resolution and feature more challenging and complex scenes, posing greater interpretation difficulties. These aforementioned differences pose challenges for the performance of models on this dataset. Fig. 5 presents the qualitative results of the comparative methods on the SVCD dataset. In row 1, this scene involves changes in both buildings and vehicles. Several earlier methods (i.e., CDNet, DSIFN, SNUNet, DTCDSN, BIT) exhibit insensitivity to changes in small-sized vehicles, leading to numerous missed detections. Furthermore, the comparative methods exhibit poor performance in detecting the decrease in buildings within the yellow rectangle. The second row of Fig. 5 includes the presence of roads and fences, in addition to buildings. It can be observed that nearly all methods can detect changes in buildings, but they demonstrate different degrees of missed detections for elongated roads and fences. The third row depicts a scene with typical seasonal transitions, involving changes from bare soil to forests. Due to spatial misalignment, the outlines of the detected change regions by the comparative algorithms deviate to some extent from the ground truth labels. Additionally, the transition from green grass to white snow leads to a significant number of false positives in earlier algorithms. In all three typical scenarios, ChangeRD exhibits minimal occurrences of missed detections and false positives, effectively tackling the challenges and surpassing the comparative methods.

Table 2 presents the quantitative results on the SVCD dataset. Competitors demonstrate low Recall, indicating a notable amount of missed detections in their prediction results. This can be attributed to the inherent challenges in the SVCD dataset, which include subtle road changes, small



**Figure 5:** The qualitative comparison of different methods on the SVCE dataset., the yellow boxes show the obvious wrong segmentation. Please zoom-in for the best view. (a) Pre-temporal image. (b) Post-temporal image. (c) FC-Siam-conc. (d) FC-Siam-diff. (e) DSIFN. (f) DTCDCN. (g) BIT. (h) ChangeFormer. (i) P2V-CD. (j) Ours. (k) Groud truth. 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. Please zoom in for a better view.



**Figure 6:** The qualitative results of ablation study. (a) Warped pre-temporal image. (b) Post-temporal image. (c) Baseline. (d) Baseline + AgCDC. (e) Baseline + APT. (f) Proposed ChangeRD. (g) Ground truth.

vehicle variations, and pronounced spatial misalignments. ChangeRD achieves a 7.81 percentage point higher recall compared to its competitors. Additionally, benefiting from the AgCDC module, which effectively captures intrinsic high-discriminative differences in features, ChangeRD is able to identify a larger number of change areas in scenarios with substantial seasonal variations. In terms of the combined F1 and IoU metrics, ChangeRD maintains an advantage of 6.87% and 11.87%, respectively, over the second-best method.

#### 4.5. Ablation Studies

To investigate the effects of the APT (Adaptive Perspective Transformation) module and AgCDC (Attention-guided Central Difference Convolution) module on change detection for unaligned images, we conduct comprehensive ablation studies on the LEVIR dataset by removing specific components for comparison. The qualitative results are presented in Fig. 6, while the quantitative results are provided in Table 3.

1) *Baseline with transformer backbone:* The transformer network is adopted as the baseline model, as depicted in Fig. 6. During the decoding stage, the offset prediction and image alignment processes are removed, and no adaptive per-

**Table 3**

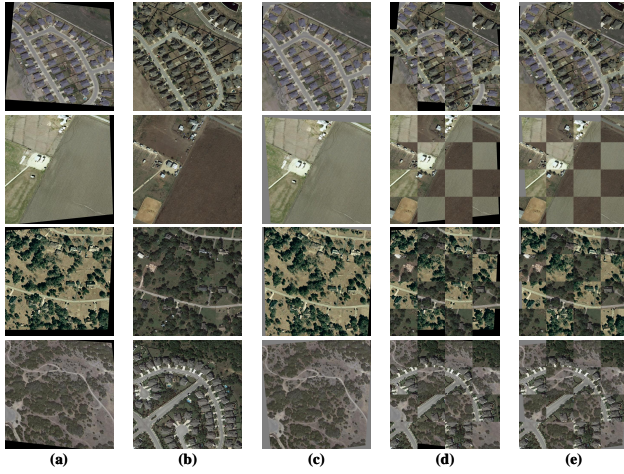
Ablation experiment on the LEVIR dataset. The best results are marked in bold.

AgCDC	APT	P (%)	R(%)	F1(%)	IoU(%)	Params (M)
		90.90	86.79	88.80	79.85	<b>25.81</b>
✓		89.93	89.39	89.66	81.26	29.38
	✓	91.66	89.76	90.70	82.99	25.89
✓	✓	<b>92.92</b>	<b>90.08</b>	<b>91.48</b>	<b>84.30</b>	29.47

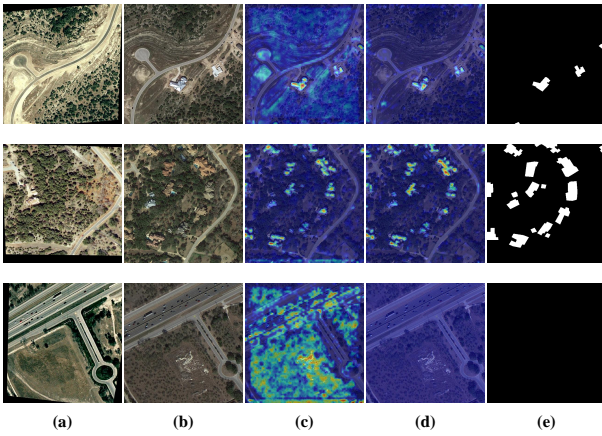
spective transformation is applied to the multi-scale features. Moreover, the AgCDC module is substituted with a regular convolutional layer. To ensure fairness and maintain consistency across all comparative experiments, we keep other hyperparameters unchanged. The baseline has 25.81M parameters and reaches an F1 score of 88.80% and an IoU of 79.85%.

2) *Effect of APT module:* To mitigate the influence of registration accuracy on change detection tasks, we design the APT module to align multi-scale features before extracting difference features. As shown in Fig. 6, row 1(c) and (d), when the APT module is removed, multiple buildings are misclassified within the yellow rectangular region due to the spatial misalignment of the bi-temporal images. This issue is also noticeable in the red rectangular region of row 2. Quantitatively, as indicated in Table 3, row 3, integrating the APT module into the baseline results in a substantial improvement of 1.9% in the overall F1 and 3.14% in IoU, with a marginal increase of only 0.08M parameters.

3) *Effect of AgCDC module:* The AgCDC module is proposed to extract discriminative difference information from the bi-temporal features that is robust to spectral variations. As shown in Fig. 6, the orange rectangular region in row 3, compared to the network without the AgCDC module, exhibits fewer false change noises in Fig. 6 (d) and (f). As shown in Table 3, the AgCDC module improves the F1 score by 0.86% and IoU by 1.41% compared to the baseline net-



**Figure 7:** Registration visualization of ChangeRD on warped images. (a) Warped pre-temporal image. (b) Post-temporal image. (c) Registered pre-temporal image. (d) Mosaic stitching of unaligned bi-temporal images. (e) Mosaic stitching of aligned bi-temporal images.



**Figure 8:** Example of AgCDC module visualization by Gradient-weighted class activation maps (Grad-CAM). (a) Warped pre-temporal image. (b) Post-temporal image. (c) Grad-CAM of the model with AgCDC module. (d) Grad-CAM of the model without AgCDC module. (e) Ground truth.

work.

Overall, by leveraging the advantages of both the APT and AgCDC modules, ChangeRD mitigates the effects of spatial misalignment and illumination variations on change detection accuracy. It improves the IoU of the baseline network from 79.85% to 84.30%.

#### 4.6. Model Analysis

1) *Analysis of Adaptive Perspective Transformation:* Fig. 7 depicts the alignment of the bi-temporal images by ChangeRD in various scenarios. Row 1 is a typical residential area, characterized by abundant robust matching features, such as building corners and roads, present in the bi-temporal images. As shown in the fourth column, the registered pre-temporal

image exhibits good alignment with the post-temporal image. In row 2, only a few prominent objects are available for pairing. Nonetheless, ChangeRD successfully accomplishes a satisfactory alignment of the bi-temporal images in this scenario. The third row represents a scene with numerous irregularly shaped trees. Despite substantial variations in tree characteristics between different temporal instances, ChangeRD manages to achieve rudimentary alignment in this scene, primarily attributed to the distinctive characteristics of the road. However, in row 4, the background becomes more complex, and the bi-temporal images exhibit substantial spectral and textural changes. Consequently, the alignment performance of ChangeRD in this specific scenario is suboptimal. Nevertheless, it is important to note that in change detection tasks involving such complex scenes, the registration accuracy generally has minimal impact due to the large number of observable changes.

2) *Analysis of AgCDC module:* To further demonstrate the role of the AgCDC module in extracting robust difference features, we visualize the class activation maps (Selvaraju et al., 2017) with or without the AgCDC module. The model without the AgCDC module uses regular convolution for difference computation. As shown in Fig. 8, row 1(c), due to significant spectral variations in the bi-temporal images, the regular convolution generates a large amount of gradient noise in the forest region while extracting the added buildings. Similarly, in row 3, illumination variations cause the traditional convolution to fail severely, resulting in a high number of false changes, whereas the AgCDC module robustly handles such illumination variations. Furthermore, as shown in Fig. 8, row 2, the AgCDC module significantly enhances the object change areas, highlighting its advantage in extracting discriminative difference features.

## 5. CONCLUSION

In this paper, registration and change detection are integrated into a unified framework that simultaneously learns the spatial transformation relationship and change information from bi-temporal image pairs. To mitigate the impact of spatial mismatch on change detection tasks, we introduce the Adaptive Perspective Transformation (APT) module, which performs feature-level registration during the change detection process. Additionally, we design the Attention-guided Central Difference Convolution (AgCDC) module to extract illumination-insensitive discriminative difference information from the bi-temporal features. Experimental results show that ChangeRD surpasses 9 SOTA methods in terms of quantitative metrics and qualitative results on three public datasets. However, it is important to note that the proposed ChangeRD, as a supervised model, relies on pre-aligned bi-temporal images as training data, which may limit its robustness and applicability in certain scenarios. An important future improvement would be to develop an unsupervised registration branch based on deformation fields, which can address the limitations in the training strategy of ChangeRD. Moreover, the progress in change detection can also enhance the



development of image registration tasks in highly dynamic scenes, fostering deeper complementarity between these two domains. Additionally, future work could explore the integration of advanced machine learning techniques, such as self-supervised learning or domain adaptation, to further improve the robustness and applicability of our method in various real-world scenarios.

## 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.

## Acknowledgement

This study is supported in part by the National Natural Science Foundation of China under Grant U21B2041; and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University, China, 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, 208–223.
- Alcantarilla, P. F., Stent, S., Ros, G., Arroyo, R., & Gherardi, R. (2018). Street-view change detection with deconvolutional networks. *Auto. Robots*, 42, 1301–1322. URL: <https://doi.org/10.1007/s10514-018-9734-5>. doi:10.1007/s10514-018-9734-5.
- Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., & Liu, B. (2023). Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois. *ISPRS J. Photogramm. Remote Sens.*, 201, 193–208. URL: <https://www.sciencedirect.com/science/article/pii/S0924271623001235>. doi:https://doi.org/10.1016/j.isprsjprs.2023.05.006.
- Bandara, W. G. C., & Patel, V. M. (2022). A transformer-based siamese network for change detection. In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)* (pp. 207–210). doi:10.1109/IGARSS46834.2022.9883686.
- Caye Daudt, R., Le Saux, B., & Boulch, A. (2018). Fully convolutional siamese networks for change detection. In *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)* (pp. 4063–4067). doi:10.1109/ICIP.2018.8451652.
- Chang, H.-H., Wu, G.-L., & Chiang, M.-H. (2019). Remote sensing image registration based on modified sift and feature slope grouping. *IEEE Geosci. Remote Sens. Lett.*, 16, 1363–1367. doi:10.1109/LGRS.2019.2899123.
- Chen, H., Qi, Z., & Shi, Z. (2022). Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.*, 60, 1–14. doi: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. URL: <https://www.mdpi.com/2072-4292/12/10/1662>. doi:10.3390/rs12101662.
- Chen, H.-M., Varshney, P., & Arora, M. (2003). Performance of mutual information similarity measure for registration of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 41, 2445–2454. doi:10.1109/TGRS.2003.817664.
- Chi, K., Yuan, Y., & Wang, Q. (2023). Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond. *IEEE Trans. Geosci. Remote Sens.*, 61, 1–14. doi:10.1109/TGRS.2023.3285228.
- Coppin, P., Jonckheere, I., Nackaerts, K., & et al. (2004). Digital change detection methods in ecosystem monitoring: a review. *Int. J. Remote Sens.*, 25, 1565–1596.
- Deng, J. S., Wang, K., Deng, Y. H., & Qi, G. J. (2008). Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens.*, 29, 4823–4838. doi:10.1080/01431160801950162.
- DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. *arXiv:1712.07629*.
- 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 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Du, P., Wang, X., Chen, D., Liu, S., Lin, C., & Meng, Y. (2020). An improved change detection approach using tri-temporal logic-verified change vector analysis. *ISPRS J. Photogramm. Remote Sens.*, 161, 278–293. doi:10.1016/j.isprsjprs.2020.01.026.
- 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. doi:10.1109/LGRS.2021.3056416.
- Feng, R., Shen, H., Bai, J., & Li, X. (2021). Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions. *IEEE Geosci. Remote Sens. Mag.*, 9, 120–142. doi:10.1109/MGRS.2021.3081763.
- Girard, N., Charpiat, G., & Tarabalka, Y. (2019). Aligning and updating cadaster maps with aerial images by multi-task, multi-resolution deep learning. In *Asian Conference on Computer Vision (ACCV)* (pp. 675–690). Cham: Springer International Publishing.
- Gong, M., Zhan, T., Zhang, P., & Miao, Q. (2017). Superpixel-based difference representation learning for change detection in multispectral remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 55, 2658–2673. doi:10.1109/TGRS.2017.2650198.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
- 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, 574–586. doi:10.1109/TGRS.2018.2858817.
- 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. doi:10.1109/TGRS.2023.3271676.
- 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. Spatial Inf. Sci.*, XLII-2, 565–571. doi:10.5194/isprs-archives-XLII-2-565-2018.
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436.
- Lee, W., Sim, D., & Oh, S.-J. (2021). A cnn-based high-accuracy registration for remote sensing images. *Remote Sens.*, 13. URL: <https://www.mdpi.com/2072-4292/13/8/1482>. doi:10.3390/rs13081482.
- Lei, T., Wang, J., Ning, H., Wang, X., Xue, D., Wang, Q., & Nandi, A. K. (2022a). Difference enhancement and spatial-spectral nonlocal network for change detection in vhr remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 60, 1–13. doi:10.1109/TGRS.2021.3134691.
- Lei, T., Wang, J., Ning, H., Wang, X., Xue, D., Wang, Q., & Nandi, A. K. (2022b). Difference enhancement and spatial-spectral nonlocal network for change detection in vhr remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 60, 1–13. doi:10.1109/TGRS.2021.3134691.
- Li, X., Ai, W., Feng, R., & Luo, S. (2023). Survey of remote sensing image registration based on deep learning. *Natl. Remote Sens. Bull.*, 27, 267–284. doi:10.11834/jrs.20235012.
- Liang, J., Liu, X., Huang, K., Li, X., Wang, D., & Wang, X. (2014). Automatic registration of multisensor images using an integrated spatial and mutual information (smi) metric. *IEEE Trans. Geosci. Remote Sens.*, 52, 603–615. doi:10.1109/TGRS.2013.2242895.
- 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. doi:10.1109/TIP.2022.3226418.
- Liu, W., Lin, Y., Liu, W., Yu, Y., & Li, J. (2023). An attention-based multiscale transformer network for remote sensing



- image change detection. *ISPRS J. Photogramm. Remote Sens.*, 202, 599–609. URL: <https://www.sciencedirect.com/science/article/pii/S092427162300182X>. doi:<https://doi.org/10.1016/j.isprsjprs.2023.07.001>.
- Liu, Y., Pang, C., Zhan, Z., Zhang, X., & Yang, X. (2021a). Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.*, 18, 811–815. doi:[10.1109/LGRS.2020.2988032](https://doi.org/10.1109/LGRS.2020.2988032).
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10012–10022).
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60, 91–110. URL: <https://doi.org/10.1023/B:VISI.0000029664.99615.94>. doi:[10.1023/B:VISI.0000029664.99615.94](https://doi.org/10.1023/B:VISI.0000029664.99615.94).
- Lu, D., Mausel, P., BrondÅzio, E., & Moran, E. (2004). Change detection techniques. *Int. J. Remote Sens.*, 25, 2365–2401. doi:[10.1080/0143116031000139863](https://doi.org/10.1080/0143116031000139863).
- Lu, Y., Zhu, Q., Zhang, B., Lai, Z., & Li, X. (2022). Weighted correlation embedding learning for domain adaptation. *IEEE Trans. Image Process.*, 31, 5303–5316. doi:[10.1109/TIP.2022.3193758](https://doi.org/10.1109/TIP.2022.3193758).
- Ning, X., Zhang, H., Zhang, R., & Huang, X. (2024a). Multi-stage progressive change detection on high resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.*, 207, 231–244. URL: <https://www.sciencedirect.com/science/article/pii/S0924271623003404>. doi:<https://doi.org/10.1016/j.isprsjprs.2023.11.023>.
- Ning, X., Zhang, H., Zhang, R., & Huang, X. (2024b). Multi-stage progressive change detection on high resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.*, 207, 231–244. URL: <https://www.sciencedirect.com/science/article/pii/S0924271623003404>. doi:<https://doi.org/10.1016/j.isprsjprs.2023.11.023>.
- Niu, S., Liu, Y., Wang, J., & Song, H. (2020). A decade survey of transfer learning (2010–2020). *IEEE Trans. Artif. Intell.*, 1, 151–166. doi:[10.1109/TAI.2021.3054609](https://doi.org/10.1109/TAI.2021.3054609).
- Radke, R., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: a systematic survey. *IEEE Trans. Image Process.*, 14, 294–307. doi:[10.1109/TIP.2004.838698](https://doi.org/10.1109/TIP.2004.838698).
- Ru, L., Du, B., & Wu, C. (2021). Multi-temporal scene classification and scene change detection with correlation based fusion. *IEEE Trans. Image Process.*, 30, 1382–1394. doi:[10.1109/TIP.2020.3039328](https://doi.org/10.1109/TIP.2020.3039328).
- Saha, S., Bovolo, F., & Bruzzone, L. (2019). Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Trans. Geosci. Remote Sens.*, 57, 3677–3693. doi:[10.1109/TGRS.2018.2886643](https://doi.org/10.1109/TGRS.2018.2886643).
- Satalino, G., Mattia, F., Balenzano, A., Lovergine, F. P., Rinaldi, M., De Santis, A. P., Ruggieri, S., Nafria 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 *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)* (pp. 6627–6630). doi:[10.1109/IGARSS.2018.8519103](https://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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)* (pp. 618–626). doi:[10.1109/ICCV.2017.74](https://doi.org/10.1109/ICCV.2017.74).
- Sun, S., Mu, L., Wang, L., & Liu, P. (2022). L-unet: An lstm network for remote sensing image change detection. *IEEE Geosci. Remote Sens. Lett.*, 19, 1–5. doi:[10.1109/LGRS.2020.3041530](https://doi.org/10.1109/LGRS.2020.3041530).
- Suri, S., & Reinartz, P. (2010). Mutual-information-based registration of terrasars-x and ikonos imagery in urban areas. *IEEE Trans. Geosci. Remote Sens.*, 48, 939–949. doi:[10.1109/TGRS.2009.2034842](https://doi.org/10.1109/TGRS.2009.2034842).
- Thevenaz, P., & Unser, M. (2000). Optimization of mutual information for multiresolution image registration. *IEEE Trans. Image Process.*, 9, 2083–2099. doi:[10.1109/83.887976](https://doi.org/10.1109/83.887976).
- Verdie, Y., Yi, K. M., Fua, P., & Lepetit, V. (2015). TILDE: A temporally invariant learned DDetector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. URL: <https://doi.org/10.1109/Fcvpr.2015.7299165>. doi:[10.1109/cvpr.2015.7299165](https://doi.org/10.1109/cvpr.2015.7299165).
- Weismiller, R. A., Kristof, S. J., Scholz, D. K., Anuta, P. E., & Momin, S. (1977). Change detection in coastal zone environments. *Photogramm. Eng. Remote Sensing*, 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, 609–625. doi:[10.1109/TGRS.2015.2463075](https://doi.org/10.1109/TGRS.2015.2463075).
- Wu, C., Du, B., Cui, X., & Zhang, L. (2017). A post-classification change detection method based on iterative slow feature analysis and bayesian soft fusion. *Remote Sens. Environ.*, 199, 241–255. doi:<https://doi.org/10.1016/j.rse.2017.07.009>.
- Wu, C., Du, B., & Zhang, L. (2014). Slow feature analysis for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.*, 52, 2858–2874. doi:[10.1109/TGRS.2013.2266673](https://doi.org/10.1109/TGRS.2013.2266673).
- Wu, J., Fu, R., Liu, Q., Ni, W., Cheng, K., Li, B., & Sun, Y. (2023). A dual neighborhood hypergraph neural network for change detection in vhr remote sensing images. *Remote Sens.*, 15. URL: <https://www.mdpi.com/2072-4292/15/3/694>. doi:[10.3390/rs15030694](https://doi.org/10.3390/rs15030694).
- Wu, J., Li, B., Qin, Y., Ni, W., Zhang, H., Fu, R., & Sun, Y. (2021). A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.*, 105, 102615. URL: <https://www.sciencedirect.com/science/article/pii/S0303243421003226>. doi:<https://doi.org/10.1016/j.jag.2021.102615>.
- Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., & Zhao, G. (2020). Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5294–5304). doi:[10.1109/CVPR42600.2020.00534](https://doi.org/10.1109/CVPR42600.2020.00534).
- Yuan, Y., Li, Z., & Ma, D. (2022). Feature-aligned single-stage rotation object detection with continuous boundary. *IEEE Trans. Geosci. Remote Sens.*, 60, 1–11.
- Yuan, Y., Xiong, Z., & Wang, Q. (2019). Vssa-net: Vertical spatial sequence attention network for traffic sign detection. *IEEE Trans. Image Process.*, 28, 3423–3434.
- Zampieri, A., Charpiat, G., Girard, N., & Tarabalka, Y. (2018). Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In *Proc. Eur. Conf. Comput. Vis. (ECCV)* (pp. 679–696). Cham: Springer International Publishing.
- Zeng, L., Du, Y., Lin, H., Wang, J., Yin, J., & Yang, J. (2021). A novel region-based image registration method for multisource remote sensing images via cnn. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 14, 1821–1831. doi:[10.1109/JSTARS.2020.3047656](https://doi.org/10.1109/JSTARS.2020.3047656).
- Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., & Liu, G. (2020a). 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. URL: <https://www.sciencedirect.com/science/article/pii/S0924271620301532>. doi:<https://doi.org/10.1016/j.isprsjprs.2020.06.003>.
- Zhang, Y., Chen, G., Vukomanovic, J., Singh, K. K., Liu, Y., Holden, S., & Meentemeyer, R. K. (2020b). Recurrent shadow attention model (rsam) for shadow removal in high-resolution urban land-cover mapping. *Remote Sens. Environ.*, 247, 111945. doi:<https://doi.org/10.1016/j.rse.2020.111945>.
- Zheng, Z., Zhong, Y., Tian, S., Ma, A., & Zhang, L. (2022). Change-mask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 228–239. URL: <https://www.sciencedirect.com/science/article/pii/S0924271621002835>. doi:<https://doi.org/10.1016/j.isprsjprs.2021.10.015>.
- Zheng, Z., Zhong, Y., Zhao, J., Ma, A., & Zhang, L. (2024). Unifying remote sensing change detection via deep probabilistic change models: From principles, models to applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215, 239–255. URL: <https://www.sciencedirect.com/science/article/pii/S0924271624002624>. doi:<https://doi.org/10.1016/j.isprsjprs.2024.07.001>.
- Zhu, Q., Zhang, Y., Wang, L., Zhong, Y., Guan, Q., Lu, X., Zhang, L., & Li, D. (2021). A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS J. Photogramm.*

*Remote Sens.*, 175, 353–365. URL: <https://www.sciencedirect.com/science/article/pii/S0924271621000873>. doi:<https://doi.org/10.1016/j.isprsjprs.2021.03.016>.

Zhu, X., Zhang, Y., Cao, H., Tan, K., & Ling, X. (2018). A novel fine registration technique for very high resolution remote sensing images. In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)* (pp. 4085–4088). doi:10.1109/IGARSS.2018.8519137.