

A Comprehensive Review of Remote Sensing Image Change Detection Methods

Gai Liu

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China
212409020031@home.hpu.edu.cn

Abstract: Remote sensing image change detection aims to identify spatiotemporal changes in land cover, land use, and ground-object structures by comparing remote sensing images acquired over the same area at different times. It is an important research topic at the intersection of remote sensing image processing, geographic information science, and computer vision. With the rapid accumulation of high-resolution, multi-source, and multi-temporal remote sensing data, change detection has been widely applied to urban expansion monitoring, land-use surveys, disaster damage assessment, and ecological environmental protection. This review focuses on remote sensing image change detection. It summarizes the general workflow, outlines the principles and limitations of traditional methods, and systematically reviews deep learning-based change detection methods. In addition, mainstream public datasets and commonly used evaluation metrics are summarized. The core challenges faced by current research are analyzed from the perspectives of data quality, data resources, and methodological design, including image registration errors, pseudo-change interference, and the high cost of annotation. Finally, future research directions are discussed to provide references for method development, model optimization, and practical applications in remote sensing image change detection.

Keywords: Remote sensing, change detection, deep learning, Siamese network, Transformer.

1. Introduction

Remote sensing image change detection is an important research direction at the intersection of remote sensing image processing, geographic information science, and computer vision. Its core objective is to identify temporal changes in land cover, land use, or ground-object structures by comparatively analyzing remote sensing images acquired over the same area at different times [1]. With the rapid development of satellite, aerial, and unmanned aerial vehicle remote sensing technologies, multi-source, multi-temporal, and high-spatial-resolution remote sensing data have continued to increase, providing a rich data foundation for dynamic land-surface monitoring. Among various paradigms, bi-temporal remote sensing image change detection has become one of the most representative research paradigms because its data acquisition form is relatively clear, its task definition is explicit, and its application scope is broad. This task generally takes pre-change and post-change remote sensing images as input and outputs pixel-level or object-level changed regions, thereby supporting subsequent geospatial analysis and decision-making. Remote sensing image change detection has significant application value in natural resource surveys, urban governance, ecological and environmental protection, and disaster emergency management [2]. As global urbanization accelerates, urban spatial structures continue to expand and construction land increases. Automatically identifying urban expansion areas from bi-temporal or multi-temporal remote sensing images can provide timely and objective data support for territorial spatial planning, urban growth boundary control, infrastructure construction assessment, and illegal land-use supervision. In land-use analysis, change detection can reveal conversion relationships among cropland, forest, water bodies, grassland, and built-up land, thereby helping to analyze the intensity of human activities, patterns of land-resource development, and

their impacts on ecosystems. In disaster assessment, remote sensing change detection can rapidly identify surface damage caused by natural disasters and provide key information for disaster evaluation and post-disaster reconstruction. In ecological environmental monitoring, change detection can be used to monitor deforestation, wetland degradation, desertification, and related processes, thereby supporting the evaluation of ecological protection policies and regional environmental change trends [3].

Early remote sensing change detection methods mainly relied on hand-crafted features and traditional image processing techniques. Common approaches include image differencing, image ratioing, change vector analysis, principal component analysis, post-classification comparison, and threshold-based change extraction [1, 4]. These methods have advantages such as relatively simple computation, strong interpretability, and low dependence on large-scale data. However, traditional methods usually depend heavily on the quality of image preprocessing and expert experience. They are sensitive to radiometric differences and seasonal variations and often struggle to distinguish real ground-object changes from pseudo-changes caused by illumination, imaging angle, phenological differences, and other factors [3]. In recent years, the rise of deep learning has created new opportunities for remote sensing image change detection. Convolutional neural networks can automatically extract multi-level spatial and semantic features through end-to-end learning, thereby substantially improving the accuracy and robustness of changed-region identification [5]. Deep models represented by fully convolutional networks, U-Net, Siamese networks, and encoder-decoder architectures have gradually become mainstream methods for bi-temporal remote sensing image change detection. Furthermore, attention mechanisms, Transformers, multi-scale feature fusion, cross-temporal feature interaction, and semantic guidance have been widely introduced to enhance the ability of models to represent

complex change patterns, long-range dependencies, and fine-grained boundary information. Meanwhile, new learning paradigms such as self-supervised learning, semi-supervised learning, weakly supervised learning, and domain adaptation have also attracted increasing attention to alleviate the shortage of pixel-level annotations and insufficient cross-region generalization [6, 7].

This paper reviews research on remote sensing image change detection, with a particular focus on bi-temporal remote sensing image change detection. It aims to systematically summarize the basic theory of remote sensing change detection, the development process from traditional methods to deep learning methods, the technical system, and future trends, thereby providing a reference for method optimization and practical applications in remote sensing change detection.

2. General Workflow of Remote Sensing Change Detection

Remote sensing image change detection usually transforms multi-temporal remote sensing observations into interpretable change information through a series of technical procedures. Although traditional methods, machine learning methods, and deep learning models differ in their specific algorithms, most change detection systems follow a relatively unified workflow, including data acquisition, image preprocessing, feature extraction, change information generation, change region determination, post-processing, and accuracy assessment [3, 4].

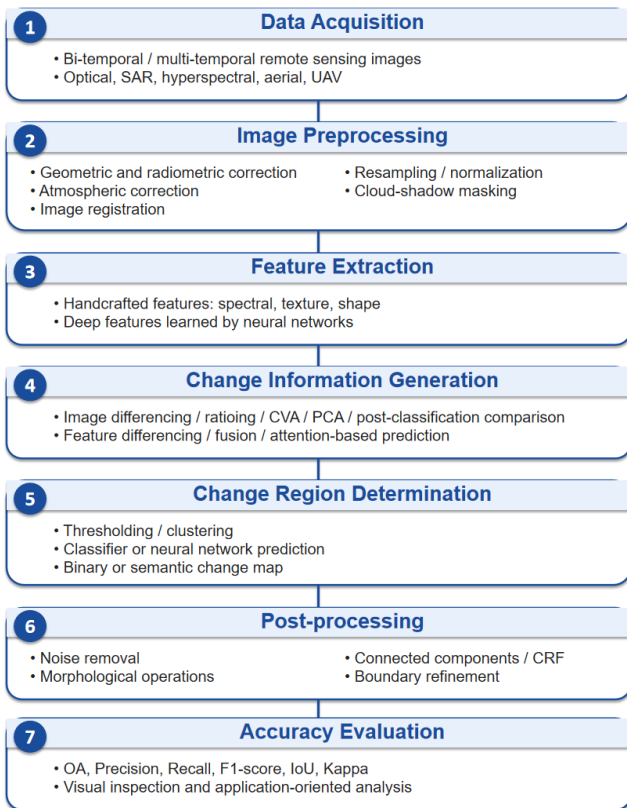


Figure 1. General Workflow of Remote Sensing Image Change Detection

2.1. Data Acquisition

Data acquisition is the initial step of remote sensing change detection. It mainly involves obtaining remote sensing images of the same geographical area at two or more different times.

These images can be acquired by optical satellites, synthetic aperture radar sensors, hyperspectral sensors, aerial platforms, or unmanned aerial vehicles. The choice of data source depends on the application objective, the spatial scale of the study area, the required temporal observation frequency, and the environmental conditions during image acquisition. Optical images contain rich spectral and spatial information and are therefore widely used for land-cover change, vegetation change, urban expansion, and building change analysis. However, optical images are easily affected by clouds, shadows, atmospheric conditions, and illumination differences. SAR images can be acquired under cloudy or nighttime conditions and are therefore suitable for disaster monitoring and areas frequently covered by clouds, although speckle noise and special imaging geometry increase interpretation difficulty. Hyperspectral images provide detailed spectral information and are suitable for identifying subtle material changes, whereas unmanned aerial vehicle images provide extremely high spatial resolution for local-scale monitoring. Therefore, spatial resolution, spectral resolution, temporal interval, imaging angle, and seasonal consistency should be comprehensively considered during data acquisition.

2.2. Image Preprocessing

Image preprocessing aims to reduce differences between multi-temporal images that are unrelated to real changes. Common preprocessing operations include geometric and radiometric correction, atmospheric correction, image registration, resampling, normalization, and cloud-shadow masking. Remote sensing images acquired at different times may differ not only because of real land-surface changes but also because of sensor status, illumination conditions, atmospheric effects, or differences in observation geometry. Among these preprocessing steps, image registration is particularly important. Even a small spatial misalignment between bi-temporal images may generate pseudo-changes near object boundaries, roads, buildings, rivers, and other high-contrast ground objects [8]. Radiometric inconsistency can also affect change detection results by making unchanged regions appear significantly different. Traditional pixel-level methods directly compare pixel values and are therefore especially sensitive to such errors. Deep learning methods may be more robust in some cases; however, when preprocessing quality is poor, models may still learn misleading patterns. Thus, image preprocessing provides the necessary geometric and radiometric foundation for reliable change detection.

2.3. Feature Extraction

Feature extraction transforms raw image data into representations that are more suitable for change analysis. In traditional change detection methods, features are usually manually designed and include spectral values, vegetation indices, texture features, shape features, edge information, and object-level attributes [4]. These features are typically designed according to prior knowledge of target ground objects and the specific application scenario. For example, spectral indices are suitable for vegetation change analysis, whereas texture and shape features help identify urban expansion or building changes. In deep learning-based change detection, feature extraction is usually performed automatically by neural networks [5]. Compared with handcrafted features, deep features are more suitable for complex

scenes, but their performance strongly depends on the quality of training samples and may be affected by regional, sensor, and seasonal differences.

2.4. Change Information Generation

After feature extraction, the next step is to generate change information by comparing multi-temporal images or their feature representations. Traditional methods usually generate change information through image differencing, image ratioing, change vector analysis [9], principal component analysis [10], or post-classification comparison [4]. These methods are easy to implement and highly interpretable, but they are sensitive to noise and registration errors. Deep learning methods directly predict change probability maps, semantic change maps, or category transition maps through bi-temporal image concatenation or Siamese feature extraction combined with feature differencing, fusion, or attention mechanisms [11, 12]. Model outputs may be binary change probability maps, semantic change maps, or multi-class land-cover transition maps. Compared with traditional comparison strategies, deep learning models can learn more complex relationships between bi-temporal images, but their outputs strongly depend on the quality of training data and the consistency of annotations.

2.5. Changed-Region Determination

Changed-region determination is the process of converting change information into explicit change detection results. In traditional unsupervised methods, this step usually relies on threshold segmentation or clustering. Threshold segmentation classifies pixels as changed or unchanged by setting a threshold on a change magnitude map. However, when the distributions of change magnitudes in changed and unchanged regions overlap substantially, threshold selection becomes very difficult. Clustering methods can reduce dependence on manually selected thresholds, but their results may still be affected by noise and feature selection [10]. In supervised learning and deep learning methods, changed-region determination is usually implemented through a classifier or the prediction layer of a neural network. The final result may be a binary change map indicating whether each pixel has changed, or a semantic change map indicating both the location and type of change. Binary change detection mainly answers the question of where changes occurred, whereas semantic change detection further answers which class was transformed into which other class. Whether to use binary or semantic output depends on specific application requirements. For example, disaster assessment may only require the extraction of damaged regions, whereas land-use monitoring usually requires more detailed class transition relationships.

2.6. Post-Processing

Post-processing is used to improve the spatial consistency, visual quality, and practical value of change detection results. Common post-processing methods include noise removal, morphological operations, connected-component analysis, conditional random fields, boundary refinement, and small-object filtering. These methods can remove isolated false alarms, fill small holes within changed regions, smooth irregular boundaries, and improve the spatial integrity of detected targets. The importance of post-processing varies across methods and application scenarios. Traditional pixel-level methods often generate fragmented change maps and

therefore usually require morphological optimization. Results generated by deep learning models are generally smoother, but they may still suffer from blurred boundaries, small false alarms, or incomplete target detection. In high-resolution remote sensing images, post-processing is particularly useful for refining building edges, road changes, and small-object changes. However, excessive post-processing may also remove real small-scale changes; therefore, parameter selection should be consistent with the target application and the minimum object scale of interest.

2.7. Accuracy Assessment

Accuracy assessment evaluates the reliability of change detection results by comparing the predicted change map with reference data or ground-truth labels [13]. Common evaluation metrics include overall accuracy, precision, recall, F1-score, intersection over union, Kappa coefficient, false alarm rate, and missed detection rate. Each metric reflects a different aspect of model performance. Overall accuracy measures the proportion of correctly classified pixels among all pixels, but it may be misleading when unchanged pixels dominate the image. Precision focuses on the reliability of detected changes, while recall measures the ability of the model to identify real changes. The F1-score balances precision and recall, whereas intersection over union evaluates the spatial overlap between predicted changed regions and reference changed regions. Evaluation metrics should be selected according to the application context, the degree of class imbalance, and the intended decision-making use of the detection results. The Kappa coefficient should be interpreted with caution, as it can produce counterintuitive results under extreme class imbalance [14]. A comprehensive accuracy assessment should combine quantitative metrics, visual inspection, and application-oriented error analysis.

3. Traditional Remote Sensing Change Detection Methods

Traditional remote sensing change detection methods mainly identify changed regions by comparing image gray values, spectral characteristics, manually designed features, or independently obtained classification results. These methods generally have clear theoretical foundations, relatively simple implementations, and high computational efficiency. They have been widely applied to land-use and land-cover monitoring, urban expansion analysis, vegetation degradation assessment, water-body extent extraction, and rapid disaster damage evaluation.

Image algebra methods are among the earliest and most widely used approaches to remote sensing change detection. Their basic principle is to directly compare pixel values or band combinations from images acquired at different times after geometric registration and radiometric normalization, and then identify changed regions according to the magnitude or direction of the resulting differences. Common techniques include image differencing, image ratioing, normalized difference index comparison, and change analysis based on vegetation, water, or built-up indices. Image differencing is one of the most fundamental and intuitive approaches. It calculates the absolute or signed difference between corresponding pixels in bi-temporal images and identifies changes using a predefined or adaptively determined threshold. Image ratioing emphasizes relative rather than absolute change by calculating the ratio between

corresponding gray or spectral values. This approach can reduce the influence of some multiplicative radiometric differences, although it may become unstable when denominator values are close to zero. Change vector analysis (CVA) characterizes overall spectral changes in a multi-band feature space by calculating the difference vector between corresponding pixels from two dates [9]. The magnitude of the vector represents the intensity of change, whereas its direction reflects the spectral trajectory of the change. When combined with directional partitioning, prior knowledge, or classification rules, CVA can further support the interpretation of possible change categories. By integrating information from multiple bands, CVA can describe land-surface changes more comprehensively than single-band differencing and is therefore particularly suitable for multispectral and hyperspectral remote sensing imagery.

Feature transformation methods constitute another classical technical route, among which principal component analysis (PCA) is one of the most representative. PCA transforms correlated input variables into a set of orthogonal components ordered according to their explained variance. When PCA is applied to stacked bi-temporal images, common background information may dominate the leading principal components, whereas some change-related information may appear in lower-variance components, particularly when changed areas occupy only a small proportion of the scene [10]. However, this distribution is not universal. Strong or spatially extensive changes may also contribute substantially to the leading components, while minor components may contain noise, misregistration effects, or radiometric inconsistencies. When PCA is applied to previously calculated difference images, major change patterns with relatively high variance are often represented in the leading components, although the exact distribution depends on change magnitude, spatial extent, noise level, and data normalization.

Post-classification comparison is another independent and widely used paradigm. In this approach, images from different dates are classified separately, and changes are identified by comparing the resulting thematic maps. An important advantage is that it can directly provide semantic transition information, such as vegetation-to-built-up or water-to-bare-land conversion. However, the final change map is affected by the classification errors of both dates, and inconsistencies between the two classification procedures may accumulate during comparison.

Traditional pixel-level methods have several advantages, including clear principles, simple implementation, high computational efficiency, and limited dependence on large training datasets. Consequently, they were extensively used in early studies of land-cover change, vegetation dynamics, and rapid disaster assessment. Nevertheless, their performance is highly dependent on preprocessing quality. Registration errors, illumination variation, seasonal differences, atmospheric effects, and sensor radiometric inconsistencies can introduce substantial pseudo-change signals into difference images. In addition, simple pixel-level operations usually describe only the magnitude or spectral direction of change and cannot directly characterize semantic change categories or spatial relationships. Threshold selection, expert knowledge, spatial regularization, or subsequent classification is therefore often required to generate the final change map.

Early classical machine learning methods also provided

important technical support and methodological exploration pathways for remote sensing change detection. Support vector machines (SVM), random forests, k-nearest neighbors, and decision trees laid an important foundation for the transition of remote sensing change detection from simple threshold-based discrimination to integrated multi-feature classification. SVM was one of the earliest supervised learning algorithms widely applied to remote sensing change identification. The method was systematically proposed by Cortes and Vapnik [15]. It can construct an optimal classification hyperplane in a high-dimensional feature space and handle nonlinear classification problems through kernel functions. Because remote sensing change samples are costly to acquire and feature dimensions such as spectra, texture, and indices are rich, SVM is suitable for change identification under small-sample and high-dimensional feature conditions. Decision tree algorithms have also demonstrated strong performance. Im and Jensen [16] proposed a change detection model based on neighborhood correlation image analysis and decision tree classification. This model uses correlation changes within local neighborhoods of two images as important features and combines them with decision tree classification for change identification. It addresses the limitations of traditional pixel-level methods, which rely excessively on single-pixel spectral differences, cannot effectively use spatial context, and are susceptible to noise. In addition, random forests [17] have been widely used in remote-sensing classification and have also supported related change detection tasks, while k-nearest neighbors (KNN) [18] and Bayesian-network-based methods [19] have been applied to remote-sensing change detection.

Although these methods improve change detection accuracy and feature fusion capability compared with traditional differencing and ratioing methods, their performance still depends on hand-crafted feature design, training sample quality, and model parameter settings. They also have generalization limitations when applied across regions, sensors, and times. Traditional change detection methods laid the foundation for subsequent deep learning methods in feature comparison, change modeling, and accuracy assessment. However, under massive high-resolution remote sensing data and complex change scenarios, their representation and generalization capabilities have gradually become insufficient.

4. Deep Learning-Based Change Detection Methods

In recent years, deep learning has substantially advanced remote sensing change detection by enabling models to learn hierarchical and discriminative representations directly from multi-temporal images. Compared with traditional methods, deep learning substantially reduces reliance on manually designed spectral, textural, and index-based features, although model performance still depends on data preprocessing, network architecture, loss design, training strategy, and annotation quality. Deep learning change detection methods can be categorized from several perspectives, including their basic computational units, bi-temporal feature interaction strategies, and prediction architectures. These dimensions are not mutually exclusive. For example, a model may simultaneously use a convolutional backbone, a Siamese feature extraction strategy, an attention-based interaction module, and an encoder-decoder prediction structure. The following discussion

therefore classifies representative methods according to their most distinctive technical characteristics.

4.1. CNN-Based Methods

Early deep learning methods for remote sensing change detection focused on learning discriminative difference representations from bi-temporal imagery. Some of these methods preceded the widespread adoption of fully convolutional architectures and therefore should be distinguished from later CNN-based dense prediction networks. Zhang et al. [20] proposed DRLnet, a deep difference representation learning framework that integrates difference measurement, representation learning, and unsupervised clustering into a unified optimization process. The model learns more discriminative representations of bi-temporal differences and reduces dependence on manually designed spectral and textural features as well as manually selected thresholds. However, DRLnet is more appropriately regarded as an early deep representation learning method than as a conventional convolutional change detection network. To improve the spatial consistency of pixel-level results, Gong et al. [21] proposed superpixel-based difference representation learning (SDRL). The method first partitions the image into spatially continuous and spectrally homogeneous superpixels and then performs difference representation learning at the regional level. Compared with independent pixel-wise discrimination, the introduction of superpixel constraints can suppress isolated noise and improve the regional continuity of change maps. Nevertheless, SDRL does not perform end-to-end two-dimensional convolutional feature learning in the same manner as modern CNN-based segmentation networks. Convolutional neural networks subsequently became a dominant paradigm because of their ability to learn local spatial patterns through shared convolutional kernels. A typical CNN progressively enlarges its effective receptive field through convolution, nonlinear activation, downsampling, dilation, or multi-scale feature aggregation. For change detection, bi-temporal images may be stacked as a joint input, processed by separate branches, or compared at multiple feature levels. Zhang and Shi [22] proposed the feature difference convolutional neural network (FDCNN) to address limited annotations and pseudo-change interference. The method adopts a weight-sharing dual-branch structure inspired by Siamese networks and uses transfer learning to extract multi-scale and multi-depth features from bi-temporal images. Difference maps generated at multiple feature levels are used for change discrimination, and a change-magnitude-guided loss is introduced to improve training under limited supervision. CNN-based methods substantially improve local pattern extraction and spatial feature learning compared with traditional pixel-level techniques. However, convolutional operations primarily aggregate information within local neighborhoods. Although deeper layers, dilation, and multi-scale fusion can enlarge the receptive field, the modeling of long-range dependencies may remain indirect. This limitation motivated the introduction of Siamese comparison mechanisms, attention modules, and Transformer-based architectures.

4.2. Siamese Network-Based Methods

Siamese networks are particularly suitable for change detection because they explicitly reflect the bi-temporal structure of the task. In a conventional Siamese architecture, two encoder branches process pre-change and post-change

images separately using shared parameters and project them into a common feature space. Weight sharing reduces inconsistencies caused by using different feature extractors and makes corresponding bi-temporal features more comparable. However, it cannot guarantee that all feature differences represent real land-surface changes, because illumination, phenology, atmosphere, viewing geometry, and registration errors may still generate pseudo-changes.

Not all dual-branch change detection models strictly share all parameters. Pseudo-Siamese, asymmetric, and partially shared networks may use different encoders when the two images originate from heterogeneous sensors or exhibit substantial domain differences. Therefore, weight sharing should be regarded as a defining feature of strict Siamese networks rather than a universal property of all dual-stream change detection models. Daudt et al. [11] proposed FC-Siam-diff and FC-Siam-conc, which are representative fully convolutional Siamese architectures. FC-Siam-diff calculates absolute differences between corresponding bi-temporal feature maps and transfers these differences to the decoder. FC-Siam-conc concatenates the features from the two temporal branches along the channel dimension, allowing subsequent layers to learn joint change representations. These architectures established two widely adopted feature interaction patterns: explicit differencing and feature concatenation. Zhan et al. [23] proposed a deep Siamese convolutional network for optical aerial image change detection and introduced a weighted contrastive loss. The loss encourages unchanged pixel pairs to have similar feature representations while maintaining a margin between changed pairs. By directly optimizing the discriminative structure of the feature space, the method improves the separability of changed and unchanged samples compared with approaches that rely solely on post hoc feature differencing. Zhang et al. [24] proposed DSIFN, which combines dual-stream VGG-style feature extraction, deep supervision, and an attention-based difference discrimination network. Multi-level features from the two temporal images are progressively fused with difference features to improve the representation of changes at different semantic scales. The method enhances the completeness of changed regions and the delineation of object boundaries in high-resolution imagery. Its main contribution lies in multi-level feature interaction and attentive difference fusion rather than explicit cross-sensor radiometric adaptation. Fang et al. [25] further enhanced information transfer through a densely connected Siamese nested architecture. Dense skip pathways reduce the semantic gap between shallow spatial features and deep semantic features and alleviate the loss of fine details during downsampling. Overall, Siamese change detection methods have evolved from simple feature differencing and concatenation toward metric learning, dense interaction, multi-scale fusion, and attention-guided comparison.

4.3. Attention and Transformer-Based Methods

Attention mechanisms improve feature selection by assigning different weights to spatial locations, channels, or temporal interactions. Transformer architectures further introduce self-attention-based representation learning, allowing distant regions to interact more directly than in conventional local convolution. Global self-attention can model pairwise dependencies across an entire feature map, whereas efficient variants may restrict attention to local

windows, sparse regions, semantic tokens, or reduced-resolution representations. Consequently, not every attention module enables direct global interaction in a single layer. Window-based and hierarchical mechanisms usually establish long-range dependencies through repeated local interactions and cross-window information exchange. In remote sensing change detection, attention and Transformer modules can help integrate spatially distributed evidence and suppress some irrelevant or inconsistent features. They may therefore improve robustness to structured pseudo-changes caused by shadows, seasonal variation, or background disturbance. However, attention alone cannot completely eliminate these effects, and its effectiveness still depends on training data, feature design, and temporal consistency. Chen and Shi [12] proposed STANet, which introduces a spatial-temporal attention module into a Siamese feature extraction and metric-learning framework. The model captures spatial dependencies within each temporal image and interactions between bi-temporal features, improving the selection of change-relevant information. STANet is better characterized as a Siamese attention-based network than as a conventional symmetric encoder-decoder architecture. Chen et al. [26] proposed the bitemporal image Transformer (BIT). Instead of applying computationally expensive self-attention directly to all image pixels, BIT converts convolutional feature maps into compact semantic tokens. A Transformer encoder models contextual relationships within each temporal image, and a Transformer decoder projects the enhanced token information back into the pixel-level feature space. This design reduces the computational burden of global modeling while strengthening semantic interaction between bi-temporal images. Bandara and Patel [27] proposed ChangeFormer, a Siamese hierarchical Transformer architecture for change detection. It uses Transformer-based encoders to learn multi-scale bi-temporal features and a lightweight decoder to progressively fuse these features and generate change maps. ChangeFormer demonstrates the effectiveness of Transformer-dominant feature learning for remote sensing change detection and improves global contextual representation compared with purely convolutional architectures. The Swin Transformer proposed by Liu et al. [28] is a general-purpose hierarchical visual backbone that uses window-based self-attention and shifted-window operations. Instead of performing global attention over all image patches, it restricts attention to local windows and enables cross-window communication by shifting the partitioning pattern between successive layers. This design reduces computational complexity for high-resolution imagery and has therefore been widely adopted as a backbone in remote sensing change detection. Although Transformer-based models provide strong contextual modeling capabilities, they may require substantial computational resources and large training datasets. Their ability to preserve small-object boundaries and fine local textures also depends on hierarchical feature design, positional representation, and decoder structure.

4.4. Encoder-Decoder Architectures

The encoder-decoder architecture is a fundamental design for dense change prediction. The encoder progressively extracts high-level semantic features, while the decoder restores spatial resolution and produces a pixel-level change map. Skip connections transfer high-resolution information from the encoder to the decoder and alleviate the loss of object

boundaries, small changed regions, and elongated structures caused by repeated downsampling. U-Net [29] is one of the most influential encoder-decoder architectures. Its approximately symmetric structure combines deep semantic representations with shallow spatial details through skip connections. Owing to its simplicity, flexibility, and localization capability, U-Net has been widely adapted for remote sensing change detection. Corley et al. [30] systematically evaluated U-Net-based baselines for change detection. Their experiments showed that carefully implemented U-Net and Siamese U-Net models remain highly competitive. On LEVIR-CD, these baselines achieved performance comparable to several more complex architectures, although they did not uniformly outperform the best reported Transformer- or attention-based models. Under the corrected and unified WHU-CD evaluation setting used in their study, the U-Net baselines outperformed the Transformer-based models included in the comparison. These results indicate that architectural complexity alone does not guarantee improved performance and that data partitioning, preprocessing, training configuration, and evaluation protocol can substantially affect reported accuracy. Peng et al. [31] proposed an end-to-end change detection method based on an improved UNet++ architecture. The method uses multiple side-output fusion to combine predictions from different semantic levels and implements deep supervision at multiple decoder stages. This strategy strengthens multi-scale feature learning and improves the delineation of fine-grained changed regions. Zhou et al. [32] proposed UNet++, a nested encoder-decoder architecture with dense skip pathways. Intermediate convolutional nodes gradually reduce the semantic gap between encoder and decoder features, improving feature fusion across different resolutions. Although UNet++ was originally developed for medical image segmentation, its nested skip structure has been extensively transferred to remote sensing change detection. Oktay et al. [33] proposed Attention U-Net, which incorporates attention gates into U-Net-style skip connections. These gates suppress irrelevant encoder activations and emphasize spatial regions relevant to the prediction task. When adapted to change detection, attention gates can improve the selective transfer of change-related details, although their effectiveness depends on the quality of the learned gating signals. DSIFN also follows an encoder-decoder-style dense prediction process with deep supervision and attention-based difference discrimination [24]. Its multi-level fusion design improves the interaction between semantic information and spatial details, particularly for high-resolution change objects with complex boundaries. Encoder-decoder architectures are not mutually exclusive with Siamese, Transformer, or Mamba paradigms. A change detection network may use a Siamese encoder, Transformer interaction modules, and a convolutional decoder simultaneously. Therefore, encoder-decoder should be understood primarily as a prediction topology rather than as an independent type of feature extractor.

4.5. CNN-Transformer Hybrid Architectures

CNN-Transformer hybrid architectures combine the local spatial inductive bias of convolution with the long-range dependency modeling capability of self-attention. CNN components are effective at extracting edges, textures, and local geometric structures, whereas Transformer components provide flexible contextual interaction over larger spatial

ranges. Their integration is therefore particularly attractive for very-high-resolution remote sensing imagery, where changed objects may vary substantially in scale and shape.

Chen et al. [34] proposed SUT, a Siamese UNet3+-Transformer architecture that combines Siamese feature extraction, full-scale dense skip connections, and Transformer modules. The network also introduces a progressive attention module to strengthen multi-scale change feature fusion and long-range spatial interaction. Its design mainly addresses insufficient cross-scale information integration and incomplete representation of complex changed regions. Deng et al. [35] proposed TChange, a hybrid Transformer-CNN network for very-high-resolution remote sensing image change detection. The method enhances interactions between features at different scales and shortens cross-scale information transmission paths. It aims to improve the simultaneous detection of large changed regions and small objects by integrating local detail extraction with contextual modeling. Ailimujiang et al. [36] proposed TUNetCD, a U-shaped change detection network based on the Swin Transformer. It extracts hierarchical features through a Swin Transformer encoder and introduces a difference enhancement module to strengthen bi-temporal change representations. The model focuses on improving multi-scale feature extraction and the use of explicit difference information. Yang et al. [37] proposed EHCTNet, an enhanced hybrid CNN-Transformer network with dual feature modeling pathways. The model integrates convolutional and Transformer modules to strengthen local-global complementarity and multi-scale change representation. This design improves the discrimination of changed objects in complex scenes where local textures and global spatial relationships are both important.

Hybrid architectures alleviate some limitations of pure CNNs and pure Transformers. However, many such networks introduce additional modules, interaction pathways, and decoder stages, which may increase structural complexity, memory consumption, and training cost. These disadvantages are not universal, because actual efficiency depends strongly on backbone size, token resolution, attention range, and fusion strategy. Lightweight hybrid designs remain an important research direction.

4.6. Mamba-Based Methods

Mamba is an emerging architecture based on selective state-space modeling. Gu and Dao [38] introduced input-dependent selective state-space layers, commonly associated with the S6 formulation, together with a hardware-aware selective scan algorithm. The selective mechanism enables the model to dynamically control which information is retained or discarded according to the input. Compared with standard self-attention, the computational complexity of the selective scan grows approximately linearly with sequence length, making it attractive for long-sequence and high-resolution vision tasks.

Mamba and S6 are related but not identical concepts. S6 refers primarily to the selective state-space layer obtained by introducing input-dependent parameters into earlier structured state-space formulations, whereas Mamba denotes the broader architecture and implementation framework built around selective state-space modeling. Chen et al. [39] were among the first to systematically investigate Mamba for remote sensing change detection and proposed ChangeMamba. The method uses a visual Mamba encoder to

extract bi-temporal features and introduces spatiotemporal relationship modeling tailored to change detection. It aims to capture both spatial dependencies within each image and temporal interactions between the two acquisition dates. ChangeMamba develops three task-specific frameworks: MambaBCD for binary change detection, MambaSCD for semantic change detection, and MambaBDA for building damage assessment. These frameworks demonstrate that state-space modeling can be adapted to different change detection tasks rather than being limited to binary change segmentation. To address the relatively weak local detail modeling that may arise when state-space scanning is directly applied to dense prediction, Zhang et al. [40] proposed CDMamba. The method combines Mamba-based long-range dependency modeling with convolutional local feature extraction. Its scaled residual ConvMamba module captures global contextual information through selective state-space operations while reinforcing local change details through convolution. CDMamba also introduces an adaptive global-local guided fusion module. This module uses global and local cues to regulate the interaction between bi-temporal features and generate more discriminative change representations. The design reflects a broader trend in visual Mamba research: combining state-space modeling with convolutional or local-enhancement components to balance global context and fine spatial structure. Liu et al. [41] proposed GSTM-SCD, a graph-enhanced spatiotemporal state-space model for semantic change detection in multi-temporal remote sensing imagery. The method combines graph-based spatial relationship modeling with Mamba-based spatiotemporal state-space modeling. Graph structures help describe irregular relationships among land-cover objects, while state-space modules efficiently capture long-range dependencies across spatial and temporal dimensions. GSTM-SCD further introduces a bidirectional three-dimensional change scanning strategy and a loss function designed for time-series semantic change detection. These components support the modeling of multi-temporal observations and complex temporal relationships and improve the modeling of complex spatial topology and long-range temporal relationships.

The original Mamba formulation does not possess the same explicit local spatial inductive bias as convolution. Consequently, many visual Mamba models incorporate convolution, local windows, multi-directional scanning, or hierarchical feature fusion. Future research should further explore how to construct Mamba-based change detection frameworks that balance long-range dependency modeling, local boundary preservation, computational efficiency, and cross-domain generalization. Other important directions include lightweight selective scanning, multi-sensor temporal alignment, uncertainty modeling, and robustness under limited annotations.

5. Datasets and Evaluation Metrics

5.1. Commonly Used Datasets

Datasets are the basic support for remote sensing image change detection research. Their quality directly affects the effectiveness of model training, the comparability of experimental results, and the credibility of algorithmic conclusions. Table 1 lists commonly used change detection datasets.

Table 1. Commonly used change detection datasets

Dataset	Source	Spatial resolution	Number of images	Change type	Main characteristics
LEVIR-CD [12]	Google Earth	0.5 m/pixel	637 pairs of bi-temporal images, each with a size of 1024×1024	Building changes, including new construction and demolition	Covers multiple regions in Texas, USA, with a time span of approximately 2002–2018. It contains 31,333 building change instances and is one of the most widely used benchmarks for building change detection.
WHU-CD [42]	Aerial remote sensing images derived from building change data related to the WHU Building Dataset	0.075 m/pixel	Originally one pair of large bi-temporal aerial images, approximately $32,507 \times 15,354$; commonly cropped into 256×256 image patches	Building changes, mainly including new construction and demolition	Covers the Christchurch region of New Zealand. The images are typically acquired in 2012 and 2016. Its extremely high spatial resolution is useful for evaluating fine building-boundary extraction, although the geographical source is relatively limited.
DSIFN-CD [24]	Google Earth high-resolution optical images	Approximately 2 m/pixel	A commonly used split contains 3,600 training pairs, 340 validation pairs, and 48 test pairs.	Binary change detection involving multiple land-cover object types, including buildings, roads, water bodies, farmland, and other surface changes.	Covers Chinese cities including Beijing, Chengdu, Shenzhen, Chongqing, Wuhan, and Xi'an. Its land-cover types and background complexity are higher than those of single-building datasets, and it is commonly used to evaluate model generalization to multiple change types.
CDD [43]	Google Earth remote sensing images and synthetic changed images	0.03–1 m/pixel	Approximately 16,000 cropped image pairs of 256×256 pixels, commonly split into 10,000 training, 3,000 validation, and 3,000 test pairs.	Binary change detection under seasonal variation, illumination differences, synthetic object changes, and small-object changes.	Includes real seasonal changes and synthetic change samples, making it suitable for analyzing model robustness under seasonal differences, illumination changes, and small-object change conditions.
SYSU-CD [44]	Aerial remote sensing images; multi-temporal data from Hong Kong	0.5 m/pixel	20,000 pairs of 256×256 bi-temporal images	Diverse urban change scenarios with binary change annotations, including newly built buildings, suburban expansion, groundwork before construction, vegetation changes, road expansion, and sea/coastal construction.	Contains 20,000 pairs of 256×256 aerial images with a spatial resolution of 0.5 m, acquired in Hong Kong between 2007 and 2014. It includes richer urban change scenarios and more complex backgrounds than building-only datasets such as LEVIR-CD.
S2Looking [45]	Off-nadir satellite optical images with different off-axis imaging angles	0.5–0.8 m/pixel	5,000 pairs of 1024×1024 bi-temporal images	Building construction and demolition	Designed for off-nadir satellite imagery and covering global rural areas, it contains more than 65,920 change instances. Compared with near-nadir datasets, it better reflects challenges caused by viewpoint differences, shadows, and registration deviations.
SECOND [46]	Multi-region aerial optical remote sensing images with varied acquisition conditions	Approximately 0.5–3 m/pixel	4,662 pairs of 512×512 bi-temporal images	Semantic class changes involving six major land-cover classes: non-vegetated ground surfaces, trees, low vegetation, water bodies, buildings, and playgrounds; approximately 30 change patterns.	It focuses not only on whether a change occurs but also on what class changes into what other class. It is suitable for studying semantic consistency, class transitions, and multi-task change detection.
xView2/xBD [47]	Pre- and post-disaster high-resolution satellite images from Maxar/DigitalGlobe and related sources	High-resolution optical satellite imagery with variable GSD; public tiles are commonly provided as 1024×1024 RGB images.	The complete xBD dataset contains 22,068 images, i.e., 11,034 pairs of pre-/post-disaster images, with 850,736 building polygons	Building localization and disaster damage severity assessment, including no damage, minor damage, major damage, and destroyed.	Designed for disaster response and humanitarian assistance, it covers 19 natural disasters and multiple disaster types. It is an important dataset for extending traditional change detection to building damage severity assessment.

5.2. Evaluation Metrics

Remote sensing image change detection can be regarded as a pixel-level binary classification, semantic segmentation, or multi-class classification problem, depending on the task definition. In binary change detection, a model usually needs to determine whether each pixel has changed. In building

change detection, the model focuses more on the spatial consistency of newly built, demolished, or damaged building regions. In semantic change detection, the model must not only identify change locations but also recognize the land-cover classes before and after the change. Therefore, model performance evaluation should not rely on a single metric; instead, it should be comprehensively assessed from multiple

perspectives.

In binary change detection, evaluation metrics are commonly defined based on the confusion matrix. TP (True Positive) denotes the number of pixels that are truly changed and correctly predicted as changed; TN (True Negative) denotes the number of pixels that are truly unchanged and correctly predicted as unchanged; FP (False Positive) denotes the number of pixels that are truly unchanged but incorrectly predicted as changed; and FN (False Negative) denotes the number of pixels that are truly changed but incorrectly predicted as unchanged.

Precision represents the proportion of truly changed pixels among pixels predicted as changed. It is defined as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

This metric mainly measures the ability of a model to control false alarms. A higher Precision indicates that fewer unchanged regions are incorrectly classified as changed regions, but it does not directly indicate whether real changed regions are sufficiently detected.

Recall represents the proportion of real changed regions that are correctly detected. It is defined as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Recall mainly measures the ability of a model to reduce missed detections. A higher Recall indicates that the model can discover more real changed regions, although in some cases it may be accompanied by an increase in false alarms.

The F1-score is the harmonic mean of Precision and Recall. It is defined as:

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

The F1-score considers both false alarms and missed detections and is suitable for comprehensively evaluating model performance when changed and unchanged pixels are imbalanced. In remote sensing change detection, unchanged regions usually far outnumber changed regions. Accuracy alone may therefore fail to reflect the ability of a model to identify minority changed regions. For this reason, the F1-score is one of the most commonly used core metrics in binary change detection studies.

IoU measures the degree of overlap between predicted changed regions and ground-truth changed regions. It is defined as:

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

Compared with the F1-score, IoU places greater emphasis on the consistency of spatial regions and is an important metric for evaluating the segmentation quality of changed regions. Overall Accuracy (OA) represents the proportion of all pixels that are correctly classified. It is defined as:

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

OA intuitively reflects the overall classification accuracy and is useful for providing a general overview of model performance. However, in remote sensing change detection, unchanged regions often occupy the vast majority of pixels. Thus, even if a model performs poorly on changed regions, it may obtain a high OA by correctly predicting many

unchanged pixels. Therefore, OA should generally not be used as the only evaluation metric.

The Kappa coefficient measures the agreement between prediction results and ground-truth annotations while considering the effect of random agreement. It is defined as:

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e}$$

Here, p_o denotes the observed agreement rate, which usually corresponds to the overall proportion of correctly classified samples; p_e denotes the random agreement rate, which can be calculated from the marginal distributions of predicted and true classes in the confusion matrix. Kappa is commonly used to evaluate classification agreement, but it can be sensitive to class imbalance and marginal distributions. Therefore, it should be interpreted together with Precision, Recall, F1-score, IoU, and other metrics.

mIoU (mean Intersection over Union) is the mean of the IoU values of all classes in multi-class semantic segmentation tasks. It is defined as:

$$\text{mIoU} = \frac{1}{C} \sum_{i=1}^C \text{IoU}_i$$

Here, C denotes the number of classes, and IoU_i denotes the intersection over union of the i -th class. mIoU is suitable for semantic change detection and multi-class ground-object change recognition tasks, as it measures the average segmentation quality across all evaluated semantic classes or change categories. Compared with the IoU of a single change class, mIoU focuses more on the overall balanced performance across classes and can prevent good performance on dominant classes from masking insufficient performance on minority classes.

6. Existing Challenges

In recent years, deep learning methods have significantly improved the extraction of changed regions, boundary delineation, and complex-scene representation. Nevertheless, their practical application is still constrained by multiple factors. Compared with traditional methods based on difference images or hand-crafted features, deep learning models have stronger nonlinear representation capabilities, but they also rely more heavily on high-quality data, sufficient annotated samples, and stable training distributions. Therefore, systematically summarizing current challenges helps clarify key directions for subsequent method improvement and application deployment.

6.1. Image Registration Errors and Pseudo-Change Interference

Image registration, which spatially aligns bi-temporal images to the same coordinate system, is a prerequisite for pixel-level change detection. When pre-change and post-change images are not precisely aligned, a pixel in one image is compared with a different ground location in the other image. At land-cover boundaries, such as field edges, building outlines, and coastlines, this spatial shift causes pixels from two different land-cover types to be compared with each other, producing false difference values that are difficult to distinguish from real change signals and may therefore be misclassified by the model as real changes. This problem is especially serious in high-spatial-resolution images, where object boundaries and texture details are more prominent and

slight spatial shifts can cause obvious local differences. Consequently, higher registration accuracy is required [8]. In addition to geometric registration errors, illumination, shadows, clouds and haze, atmospheric conditions, seasonal changes, and surface moisture differences can also significantly affect the consistency of multi-temporal images. For optical remote sensing images, changes in solar elevation angle may alter shadow direction and length, cloud and haze occlusion may reduce object visibility, and vegetation phenology and crop rotation may cause substantial spectral response differences. These variations do not necessarily correspond to real land-cover changes but may appear as strong differences in the model feature space. By contrast, synthetic aperture radar (SAR) images offer day-and-night and relatively weather-independent observation capabilities. However, speckle noise, incidence-angle differences, and complex backscattering mechanisms may also introduce pseudo-change interference.

6.2. Limited Datasets and High Annotation Costs

The lack of large-scale, high-quality annotated datasets is the primary bottleneck for applying supervised deep learning to change detection. Pixel-level change annotation requires annotators to compare images acquired at different times simultaneously and to determine changed regions, change boundaries, and change types based on ground-object knowledge. This process is time-consuming and requires professional expertise. In complex scenarios such as urban renewal, disaster damage, and ecological degradation, it is also more susceptible to subjective judgment. Different annotators may use inconsistent criteria for temporary changes, ambiguous boundaries, small-scale changes, or seasonal disturbances, resulting in label noise and annotation uncertainty. Compared with single-image annotation, this task is inherently more time-consuming because annotators must inspect two images simultaneously and determine whether observed spectral differences correspond to real land-surface changes. Annotation costs increase with spatial resolution because higher resolution means more pixels must be labeled per unit area. They also increase with task complexity, such as the difference between binary change detection and multi-class change detection annotation, and with the spatial extent of the study area. Existing benchmark datasets remain relatively small by deep learning standards.

In addition, existing change detection datasets still suffer from limited diversity. First, their geographical coverage remains insufficient, with regions experiencing active land-cover changes, such as Africa, South America, South Asia, and Southeast Asia, still underrepresented [6]. Second, most existing datasets focus mainly on building change detection and lack systematic annotations for vegetation changes, water-body dynamics, agricultural land transitions, and infrastructure changes beyond buildings. Meanwhile, widely adopted benchmark datasets with dense time-series change annotations or multimodal paired labels are still lacking, which limits the development of related methods for continuous change monitoring and cross-modal change analysis. Moreover, the annotation quality of existing datasets is not sufficiently transparent, as most dataset papers do not provide detailed reports on inter-annotator agreement, annotation quality-control procedures, or boundary accuracy statistics. Therefore, the performance differences reported among different methods may not be entirely attributable to

improvements in model design, but may also be affected by annotation inconsistency, boundary errors, or dataset bias [30].

6.3. Performance Instability in Complex Scenes

Compared with the relatively simple and well-structured suburban landscapes commonly found in mainstream benchmark datasets, current change detection methods exhibit a clear performance decline in geometrically and spectrally more complex scenes. This performance instability mainly arises from high-density urban areas, shadows and extreme illumination, heterogeneous landscapes, and gradual or subtle changes. In dense urban environments, numerous buildings generate complex and varying shadow patterns across acquisition dates, while mixed pixels along building edges may simultaneously contain building, shadow, and ground reflectance components whose proportions change under slight misregistration. Narrow streets, small structures, and complex roof geometries further produce change signals at spatial scales close to the sensor resolution limit. Shadows caused by clouds, terrain, and buildings are also frequently misclassified as changes because they introduce large and spatially coherent apparent reflectance differences; although deep learning models can suppress such effects when trained with sufficient shadow examples, shadow geometry is highly scene-specific and difficult to generalize across locations or dates. In heterogeneous landscapes, such as mixed farmland, savannas with scattered trees, and urban-wildland interfaces, fine-scale spatial variability produces mixed pixels whose spectral features vary nonlinearly with sub-pixel land-cover composition. Moreover, gradual and subtle changes, including vegetation degradation, soil moisture variation, water quality change, and early-stage construction, often generate weak, spatially diffuse spectral responses that are easily confused with radiometric noise. Consequently, methods optimized for abrupt and high-contrast changes, such as new building construction or clear-cut deforestation, may systematically miss these more complex change patterns.

6.4. Insufficient Interpretability of Results

Insufficient interpretability of results is an important challenge when remote sensing change detection moves from algorithmic research toward practical decision-making applications. Existing deep learning-based change detection models can usually output changed regions or change probability maps, but their discrimination processes are often based on high-dimensional feature mappings and complex network parameters, making it difficult to directly explain why a model classifies a region as changed or unchanged. For binary change detection, model results usually remain at the level of whether a change has occurred and lack further explanation of change type, change degree, change cause, and spatial context. In application scenarios such as urban management, disaster assessment, ecological supervision, and land-resource monitoring, providing only change masks is often insufficient for practical decision-making because managers usually need to understand the basis of changes, the confidence of results, and potential risks of false alarms or missed detections.

From the perspective of model mechanisms, insufficient interpretability in remote sensing change detection mainly stems from the black-box nature of deep neural networks. Although Convolutional Neural Networks (CNNs) can extract textures, edges, and spatial structures, the relationship

between hierarchical features and specific ground-object changes is not clear. Transformers and attention mechanisms can model long-range dependencies, but attention visualizations alone are not automatically sufficient as faithful explanations of model decisions. Meanwhile, multi-scale fusion, multi-branch structures, and cross-temporal interaction further increase the complexity of the discrimination process, making it difficult to trace the relationship between results and original image features. Existing explanation methods also have limitations: different methods may produce inconsistent results, and no unified standard exists for evaluating the faithfulness of explanations; saliency maps and attention visualizations usually only show regions of model focus and cannot explain specific causes; and most methods rely on post hoc analysis and cannot constrain the model during training to learn features consistent with geographic laws. In addition, change detection involves multi-temporal, multi-scale, and multi-source data and requires simultaneous explanation of temporal differences, spatial context, and sensor differences, making it more complex than ordinary natural image classification.

7. Future Development Trends

7.1. Multi-Source Remote Sensing Data Fusion and Cross-Modal Change Detection

At present, most change detection methods are still mainly based on optical images from a single data source and a single modality. Multi-source remote sensing data fusion is a key direction for improving change detection stability in complex environments. Optical images provide rich spectral and texture information and are suitable for identifying changes in object appearance and cover types. SAR images can provide observations under nighttime, cloudy, smoky, or dusty conditions and are less affected by weather than optical sensors, although precipitation and surface moisture can still influence backscatter. Hyperspectral images can characterize finer spectral differences and help identify changes in material composition, vegetation status, and water-body characteristics. Light Detection and Ranging (LiDAR) data provide three-dimensional structural information and are suitable for analyzing building height changes, forest structural changes, and terrain changes [48].

Different data sources are complementary, but they also involve problems such as inconsistent spatial resolution, differences in imaging mechanisms, difficulty in temporal synchronization, and different noise characteristics. Early fusion methods mainly adopted pixel-level concatenation, feature-level fusion, or decision-level ensemble strategies. These methods are relatively easy to implement but cannot fully represent physical differences and semantic relationships among modalities. In recent years, deep learning methods have increasingly introduced technical routes such as multi-branch networks, cross-modal attention mechanisms, shared-specific feature disentanglement, graph neural networks, and multimodal contrastive learning. Compared with simple data stacking, cross-modal representation learning is more conducive to mining complementary information among different sensors. However, multi-source fusion does not necessarily lead to performance improvement. If registration between modalities is inaccurate, acquisition times are inconsistent, or information quality differs substantially, the fusion process may introduce additional

noise.

7.2. Time-Series Change Detection

Traditional change detection mostly relies on bi-temporal images and focuses on determining whether changes occurred within a given time interval. However, many land-surface changes are continuous, staged, and periodic. Using only two time points makes it difficult to accurately reveal the occurrence time, duration, and trend of change. For example, urban expansion usually manifests as gradual increases in construction land, vegetation degradation may undergo long-term slow evolution, and disasters such as floods, fires, and landslides may include multiple stages, including pre-disaster, during-disaster, and post-disaster recovery. Therefore, multi-temporal and time-series change detection has become an important development direction in remote sensing change monitoring.

Multi-temporal data can help models distinguish real changes, short-term disturbances, and seasonal fluctuations. For agricultural and vegetation monitoring, time-series information helps distinguish normal phenological variation from abnormal degradation. For urban studies, multi-year images can be used to characterize expansion speed, spatial direction, and land-use conversion pathways. For disaster assessment, continuous observations can support analysis of disaster evolution processes and recovery status [6]. Compared with bi-temporal methods, time-series methods can provide richer information about change processes, but they also face challenges such as cloud occlusion, uneven temporal intervals, differences in historical data quality, and sensor replacement.

Breakthroughs in this direction require advances in both data and algorithms. At the data level, constructing a large-scale, open-access benchmark dataset for time-series change detection that combines harmonized Landsat-Sentinel imagery with multi-year dense change labels covering diverse change types and geographical regions would substantially promote the field. At the algorithmic level, Transformer or Mamba architectures equipped with temporal encoding, masking strategies, and multi-resolution fusion mechanisms could help handle irregular sampling and multi-resolution time series, while pretraining on harmonized Landsat-Sentinel archives could provide a representation foundation for large-scale continuous monitoring. Future research should move beyond merely identifying changed regions toward analyzing change processes, enabling comprehensive representation of change onset, duration, direction, intensity, and outcome.

7.3. Weakly Supervised, Semi-Supervised, and Self-Supervised Learning

Reducing dependence on large-scale manually annotated samples is a necessary condition for the widespread application of change detection methods. Weakly supervised learning generates pseudo-labels from low-cost supervisory information such as image-level labels, bounding boxes, point annotations, or scribble annotations. It uses class activation maps, prototype learning, consistency constraints, and spatial priors from foundation models such as the Segment Anything Model (SAM) to improve changed-region localization. Existing studies have shown that weakly supervised methods have shown the potential to approach fully supervised performance under some experimental settings [7]. However, incomplete CAM activation, insufficient spatial information

in weak labels, difficulty in ensuring cross-image semantic consistency, and inconsistent weak-supervision evaluation protocols still limit stable application.

Semi-supervised learning usually trains models using a small number of pixel-level annotated samples together with a large number of unlabeled bi-temporal images, and it is another important pathway for alleviating annotation scarcity. Existing semi-supervised change detection methods mainly include pseudo-label self-training, consistency regularization, teacher-student models, mean-teacher frameworks, entropy minimization, and uncertainty-guided sample selection. Nevertheless, semi-supervised learning in remote sensing change detection still faces challenges such as sparse changed regions, imbalanced positive and negative samples, difficult control of pseudo-label noise, and strong pseudo-change interference in unlabeled samples. Therefore, how to construct reliable pseudo-label update mechanisms, introduce change boundary and semantic consistency constraints, and combine uncertainty estimation to suppress the propagation of erroneous supervision remains a key problem in semi-supervised change detection.

In contrast, self-supervised learning completely avoids manual annotations during pretraining. It constructs supervision signals using the spatial structure, temporal differences, and seasonal consistency of remote sensing multi-temporal images themselves, and learns change-aware representations through contrastive learning, masked image modeling, and difference-pair pretraining. This approach is expected to make full use of large-scale unlabeled remote sensing archives such as Landsat and Sentinel and transfer to downstream change detection tasks with only a small number of annotated samples. Nevertheless, no mature consensus has been reached on the design of self-supervised pretraining tasks for bi-temporal or multi-temporal change detection, the relationship between pretraining scale and computational cost, or unified evaluation standards. Future breakthroughs may arise from the synergistic integration of weakly supervised, semi-supervised, and self-supervised learning with foundation segmentation models. On the one hand, weakly supervised and semi-supervised methods can use spatial priors and uncertainty estimation from foundation models to improve pseudo-label quality. On the other hand, large-scale difference-pair self-supervised pretraining is expected to form change-aware backbone networks, thereby reducing annotation costs while improving the spatial localization, semantic discrimination, and cross-region generalization capabilities of models for real land-surface changes.

7.4. Lightweight Deployment and Trustworthy Explanation

Practical applications of remote sensing change detection require not only high model accuracy but also good computational efficiency, semantic representation capability, and result trustworthiness. Large-scale remote sensing image processing usually involves massive data tiling, multi-scale inference, and result mosaicking, leading to substantial computational resource consumption. Lightweight models and efficient inference techniques are important for unmanned aerial vehicle inspection, disaster emergency response, mobile terminal monitoring, and edge computing scenarios. Existing studies mainly reduce computational complexity through model pruning, parameter quantization, knowledge distillation, lightweight convolution, and efficient attention mechanisms. However, lightweight processing may

weaken the ability of a model to perceive small-scale changes, complex boundaries, and long-range context. Therefore, a balance among speed, accuracy, and robustness is required.

Large models and vision-language models provide new possibilities for trustworthy change detection. By introducing remote sensing foundation models, vision-language models, and multimodal reasoning frameworks, future models may generate explanatory descriptions of change type, spatial distribution, and possible causes while outputting changed regions. However, such methods still face problems including factual consistency, insufficient domain knowledge, hallucination risks, and limited verifiability, and they cannot yet replace expert judgment. Trustworthy change detection should further combine uncertainty estimation, saliency analysis, attention visualization, expert-knowledge constraints, and human-machine collaboration mechanisms, enabling models to explain the basis of judgments, confidence levels, and potential sources of error [49].

8. Conclusions

This paper systematically reviewed remote sensing image change detection, including its basic workflow, traditional methods, deep learning methods, commonly used datasets, evaluation metrics, existing challenges, and future development trends. Traditional methods have advantages such as clear principles, simple computation, and relatively strong interpretability, making them suitable for some scenarios with small data volumes or clear prior knowledge. Deep learning methods can automatically learn complex spatial and semantic features and have demonstrated higher accuracy and stronger representation capability in high-resolution bi-temporal image change detection. Although substantial progress has been achieved, remote sensing change detection still faces a series of challenges. Future research should further strengthen multi-source and multimodal data fusion, fully use time-series remote sensing observations to characterize change processes, develop weakly supervised, semi-supervised, and self-supervised learning to reduce annotation dependence, and promote lightweight models, trustworthy explanations, and human-machine collaboration mechanisms in practical operations. With the continued development of remote sensing data acquisition capabilities, computing platforms, and intelligent algorithms, change detection technology will play an increasingly important role in urban governance, natural resource management, ecological environmental protection, and disaster emergency assessment.

References

- [1] Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6), 989–1003. <https://doi.org/10.1080/01431168908903939>
- [2] Hussain, M., Chen, D., Cheng, A., Wei, H., & Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80, 91–106. <https://doi.org/10.1016/j.isprsjprs.2013.03.006>
- [3] Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., & Lambin, E. (2004). Digital change detection methods in ecosystem monitoring: A review. *International Journal of Remote Sensing*, 25(9), 1565–1596. <https://doi.org/10.1080/0143116031000101675>

- [4] Lu, D., Mausel, P., Brondizio, E., & Moran, E. (2004). Change detection techniques. *International Journal of Remote Sensing*, 25(12), 2365–2407. <https://doi.org/10.1080/0143116031000139863>
- [5] Khelifi, L., & Mignotte, M. (2020). Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *IEEE Access*, 8, 126385–126400. <https://doi.org/10.1109/ACCESS.2020.3008036>
- [6] Cheng, G., Huang, Y., Li, X., Lyu, S., Xu, Z., Zhao, H., Zhao, Q., & Xiang, S. (2024). Change detection methods for remote sensing in the last decade: A comprehensive review. *Remote Sensing*, 16(13), 2355. <https://doi.org/10.3390/rs16132355>
- [7] Ding, L., Hong, D., Zhao, M., Chen, H., Li, C., Deng, J., Yokoya, N., Bruzzone, L., & Chanussot, J. (2025). A survey of sample-efficient deep learning for change detection in remote sensing: Tasks, strategies, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 13(3), 164–189. <https://doi.org/10.1109/MGRS.2025.3533605>
- [8] Dai, X., & Khorram, S. (1998). The effects of image misregistration on the accuracy of remotely sensed change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 36(5), 1566–1577. <https://doi.org/10.1109/36.718857>
- [9] Malila, W. A. (1980). Change vector analysis: An approach for detecting forest changes with Landsat. In *Proceedings of the 6th Annual Symposium on Machine Processing of Remotely Sensed Data* (pp. 326–335). Purdue University.
- [10] Celik, T. (2009). Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geoscience and Remote Sensing Letters*, 6(4), 772–776. <https://doi.org/10.1109/LGRS.2009.2025059>
- [11] Daudt, R. C., Le Saux, B., & Boulch, A. (2018). Fully convolutional Siamese networks for change detection. In *2018 IEEE International Conference on Image Processing* (pp. 4063–4067). IEEE. <https://doi.org/10.1109/ICIP.2018.8451652>
- [12] Chen, H., & Shi, Z. (2020). A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10), 1662. <https://doi.org/10.3390/rs12101662>
- [13] Congalton, R. G., & Green, K. (2019). *Assessing the accuracy of remotely sensed data: Principles and practices* (3rd ed.). CRC Press. <https://doi.org/10.1201/9780429052729>
- [14] Pontius, R. G., Jr., & Millones, M. (2011). Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15), 4407–4429. <https://doi.org/10.1080/01431161.2011.552923>
- [15] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- [16] Im, J., & Jensen, J. R. (2005). A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sensing of Environment*, 99(3), 326–340. <https://doi.org/10.1016/j.rse.2005.09.008>
- [17] Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- [18] Touazi, A., & Bouchaffra, D. (2015). A k-nearest neighbor approach to improve change detection from remote sensing: Application to optical aerial images. In *2015 15th International Conference on Intelligent Systems Design and Applications* (pp. 98–103). IEEE. <https://doi.org/10.1109/ISDA.2015.7489208>
- [19] Qin, D., Ma, J., & Yun, O. Y. (2006). Remote sensing data change detection based on the CI test of Bayesian networks. *Computers & Geosciences*, 32(2), 195–202. <https://doi.org/10.1016/j.cageo.2005.06.012>
- [20] Zhang, P., Gong, M., Zhang, H., & Liu, J. (2017). DRLnet: Deep difference representation learning network and an unsupervised optimization framework. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 3413–3419). IJCAI. <https://doi.org/10.24963/ijcai.2017/477>
- [21] Gong, M., Zhan, T., Zhang, P., & Li, W. (2017). Superpixel-based difference representation learning for change detection in multispectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2658–2673. <https://doi.org/10.1109/TGRS.2017.2650198>
- [22] Zhang, M., & Shi, W. (2020). A feature difference convolutional neural network-based change detection method. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10), 7232–7246. <https://doi.org/10.1109/TGRS.2020.2981051>
- [23] Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., & Qiu, X. (2017). Change detection based on deep Siamese convolutional network for optical aerial images. *IEEE Geoscience and Remote Sensing Letters*, 14(10), 1845–1849. <https://doi.org/10.1109/LGRS.2017.2738149>
- [24] Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., & Liu, G. (2020). A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166, 183–200. <https://doi.org/10.1016/j.isprsjprs.2020.06.003>
- [25] Fang, S., Li, K., Shao, J., & Li, Z. (2022). SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geoscience and Remote Sensing Letters*, 19, 8007805. <https://doi.org/10.1109/LGRS.2021.3056416>
- [26] Chen, H., Qi, Z., & Shi, Z. (2022). Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 5607514. <https://doi.org/10.1109/TGRS.2021.3095166>
- [27] Bandara, W. G. C., & Patel, V. M. (2022). A Transformer-based Siamese network for change detection. In *2022 IEEE International Geoscience and Remote Sensing Symposium* (pp. 207–210). IEEE. <https://doi.org/10.1109/IGARSS46834.2022.9883686>
- [28] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). IEEE. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [29] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
- [30] Corley, I., Lwande, J., Robinson, C., Dodhia, R., & Lavista Ferres, J. M. (2024). A change detection reality check. In *ICLR 2024 Workshop on Machine Learning for Remote Sensing*.
- [31] Peng, D., Zhang, Y., & Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sensing*, 11(11), 1382. <https://doi.org/10.3390/rs11111382>
- [32] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision*

- Support (Vol. 11045, pp. 3–11). Springer. https://doi.org/10.1007/978-3-030-00889-5_1
- [33] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 61–69). Springer. https://doi.org/10.1007/978-3-030-00937-3_7
- [34] Chen, M., Zhang, Q., Ge, X., Xu, B., Hu, H., Zhu, Q., & Zhang, X. (2023). A full-scale connected CNN-Transformer network for remote sensing image change detection. *Remote Sensing*, 15(22), 5383. <https://doi.org/10.3390/rs15225383>
- [35] Deng, Y., Meng, Y., Chen, J., Yue, A., Liu, D., & Chen, J. (2023). TChange: A hybrid Transformer-CNN change detection network. *Remote Sensing*, 15(5), 1219. <https://doi.org/10.3390/rs15051219>
- [36] Ailimujiang, G., Jiaermuhamaiti, Y., Jumahong, H., Wang, H., Zhu, S., & Nurmamaiti, P. (2022). A Transformer-based network for change detection in remote sensing using multiscale difference-enhancement. *Computational Intelligence and Neuroscience*, 2022, 2189176. <https://doi.org/10.1155/2022/2189176>
- [37] Yang, J., Wan, H., & Shang, Z. (2025). Enhanced hybrid CNN and transformer network for remote sensing image change detection. *Scientific Reports*, 15, 10161. <https://doi.org/10.1038/s41598-025-94544-7>
- [38] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces [Preprint]. *arXiv*. <https://arxiv.org/abs/2312.00752>
- [39] Chen, H., Song, J., Han, C., Xia, J., & Yokoya, N. (2024). ChangeMamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 4409720. <https://doi.org/10.1109/TGRS.2024.3417253>
- [40] Zhang, H., Chen, K., Liu, C., Chen, H., Zou, Z., & Shi, Z. (2025). CDMamba: Incorporating local clues into Mamba for remote sensing image binary change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 4405016. <https://doi.org/10.1109/TGRS.2025.3545012>
- [41] Liu, X., Dai, C., Ding, L., Zhang, Z., Li, Y., Zuo, X., Li, M., Wang, H., & Miao, Y. (2025). GSTM-SCD: Graph-enhanced spatio-temporal state space model for semantic change detection in multi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230, 73–91. <https://doi.org/10.1016/j.isprsjprs.2025.09.003>
- [42] Ji, S., Wei, S., & Lu, M. (2019). Building instance change detection from large-scale aerial images using convolutional neural networks. *Remote Sensing*, 11(11), 1343. <https://doi.org/10.3390/rs11111343>
- [43] Lebedev, M. A., Vizilter, Y. V., Vygolov, O. V., Knyaz, V. A., & Rubis, A. Y. (2018). Change detection in remote sensing images using conditional adversarial networks. *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2, 565–571. <https://doi.org/10.5194/isprs-archives-XLII-2-565-2018>
- [44] Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., & Zhang, L. (2022). A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 5604816. <https://doi.org/10.1109/TGRS.2021.3085870>
- [45] Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., & Jiang, B. (2021). S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24), 5094. <https://doi.org/10.3390/rs13245094>
- [46] Yang, K., Xia, G.-S., Liu, Z., Du, B., Yang, W., Pelillo, M., & Zhang, L. (2022). Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–18. <https://doi.org/10.1109/TGRS.2021.3113912>
- [47] Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., & Gaston, M. (2020). xBD: A dataset for assessing building damage from satellite imagery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–8). IEEE.
- [48] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102926. <https://doi.org/10.1016/j.jag.2022.102926>
- [49] Gevaert, C. M. (2022). Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102869. <https://doi.org/10.1016/j.jag.2022.102869>