remote sensing

MDPI

Review

Advances and Challenges in Deep Learning-Based Change Detection for Remote Sensing Images: A Review through Various Learning Paradigms

Lukang Wang [1], Min Zhang [2,3,*], Xu Gao [1] and Wenzhong Shi [2,3]

[1] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; wanglukang@cumt.edu.cn (L.W.); xugao@cumt.edu.cn (X.G.)
[2] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong; john.wz.shi@polyu.edu.hk
[3] Otto Poon Charitable Foundation Smart Cities Research Institute, The Hong Kong Polytechnic University, Hong Kong
* Correspondence: lsgi-min.zhang@polyu.edu.hk

Abstract: Change detection (CD) in remote sensing (RS) imagery is a pivotal method for detecting changes in the Earth's surface, finding wide applications in urban planning, disaster management, and national security. Recently, deep learning (DL) has experienced explosive growth and, with its superior capabilities in feature learning and pattern recognition, it has introduced innovative approaches to CD. This review explores the latest techniques, applications, and challenges in DL-based CD, examining them through the lens of various learning paradigms, including fully supervised, semi-supervised, weakly supervised, and unsupervised. Initially, the review introduces the basic network architectures for CD methods using DL. Then, it provides a comprehensive analysis of CD methods under different learning paradigms, summarizing commonly used frameworks. Additionally, an overview of publicly available datasets for CD is offered. Finally, the review addresses the opportunities and challenges in the field, including: (a) incomplete supervised CD, encompassing semi-supervised and weakly supervised methods, which is still in its infancy and requires further in-depth investigation; (b) the potential of self-supervised learning, offering significant opportunities for Few-shot and One-shot Learning of CD; (c) the development of Foundation Models, with their multi-task adaptability, providing new perspectives and tools for CD; and (d) the expansion of data sources, presenting both opportunities and challenges for multimodal CD. These areas suggest promising directions for future research in CD. In conclusion, this review aims to assist researchers in gaining a comprehensive understanding of the CD field.

Keywords: change detection; deep learning; remote sensing; semi-supervised; weakly supervised; unsupervised; self-supervised; Foundation Models; multimodal

Citation: Wang, L.; Zhang, M.; Gao, X.; Shi, W. Advances and Challenges in Deep Learning-Based Change Detection for Remote Sensing Images: A Review through Various Learning Paradigms. *Remote Sens.* **2024**, *16*, 804. https://doi.org/10.3390/rs16050804

Academic Editor: Eufemia Tarantino

Received: 24 January 2024
Revised: 23 February 2024
Accepted: 23 February 2024
Published: 25 February 2024

1. Introduction

Remote sensing (RS) image change detection (CD) is a technique that utilizes multi-temporal RS imagery and auxiliary data to extract and analyze information on surface changes [1]. CD serves as a crucial tool for earth surface observation and is extensively used for updating land use changes [2], assessing natural hazards [3], and analyzing urban sprawl [4].

With the rapid development of deep learning (DL) technology, it has shown great potential and broad application prospects in the field of CD [5–7]. DL, with its excellent capabilities in feature learning and pattern recognition, has brought innovative solutions and methods to CD. Compared to traditional methods based on manually designed features [8–11], DL automatically learns high-level feature representations from data, significantly improving the performance of CD. The transition to DL-based methods signifies a

paradigm shift in CD, moving from labor-intensive feature engineering to an era where models can autonomously learn and adapt. This development promises to accelerate the pace of innovation in CD, offering new avenues for research and practical applications that were previously unattainable with conventional methods.

With the continuous advancement in Earth observation technologies [12,13], the acquisition of RS imagery has made significant progress, as evidenced by improved spatial, temporal, and spectral resolutions. This advancement has led to an increase in the volume, complexity, and heterogeneity of RS data. Such developments present unprecedented opportunities to deeply understand the changes and evolution on the Earth's surface, but also bring substantial challenges in data processing [14,15]. These challenges include the considerable effort and time required to annotate large datasets, coarse-grained data labels, and effective leveraging of the vast amount of unlabeled Earth observation data. Against this backdrop of challenges, combined with practical application needs, CD tasks face a diversity of data sample scenarios. Hence, employing various innovative methods to tackle these varied data sample scenarios is crucial. This facilitates the maximization of the continually evolving potential of remote sensing technology, enabling practitioners to extract more meaningful solutions from the burgeoning data resources.

In these varied data scenarios, choosing the appropriate learning paradigm becomes particularly crucial. Traditional fully supervised learning paradigms perform well with sufficient labeled data but may encounter overfitting issues in data-scarce situations. Semi-supervised and self-supervised learning paradigms can enhance model performance by effectively exploiting the abundant information contained within unlabeled samples. Weakly supervised learning paradigms can achieve CD tasks using coarsely labeled data, such as image-level, bounding box, or scribble labels. Additionally, transfer learning and domain adaptation techniques play an active role in handling CD tasks across different data sources. The distinct advantages and application contexts of these learning paradigms provide diversified solutions for CD tasks, not only enriching the choice of methodologies but also opening new possibilities for CD research adaptable to various data scenarios.

Most existing reviews [5–7,16,17] in the field of CD have predominantly focused on fully supervised or unsupervised methods, often overlooking the nascent areas of semi-supervised and weakly supervised methods. Given that these emerging directions have seen limited exploration in the past reviews, this review adopts a comprehensive perspective, examining DL methods for RS image CD across various learning paradigms, as shown in as Figure 1. To better illustrate these paradigms, a schematic diagram of CD across different learning paradigms is presented in Figure 2. This review extends from the basic network architectures to the latest methods within various paradigms, providing a thorough summary and analysis of common frameworks. Furthermore, this review keenly focuses on the challenges and promising prospects brought forth by the rapid development of DL in CD, particularly emphasizing areas like self-supervised learning and Foundation Models. This comprehensive viewpoint underscores the timeliness and importance of this review, especially against the backdrop of evolving DL technologies, which are reshaping the field of CD. It not only offers researchers profound insights into the latest advancements within the domain but also delineates potential future research directions and challenges.

The rest of the paper is organized as follows. Section 2 introduces the basic network architectures of DL used for CD. Section 3 provides a comprehensive review of CD methods under different learning paradigms. Section 4 discusses the adaptation, analysis, pros and cons, and application scenarios of different learning paradigms for CD. Section 5 discusses the opportunities and challenges of CD based on DL. Finally, we draw conclusions in Section 6.

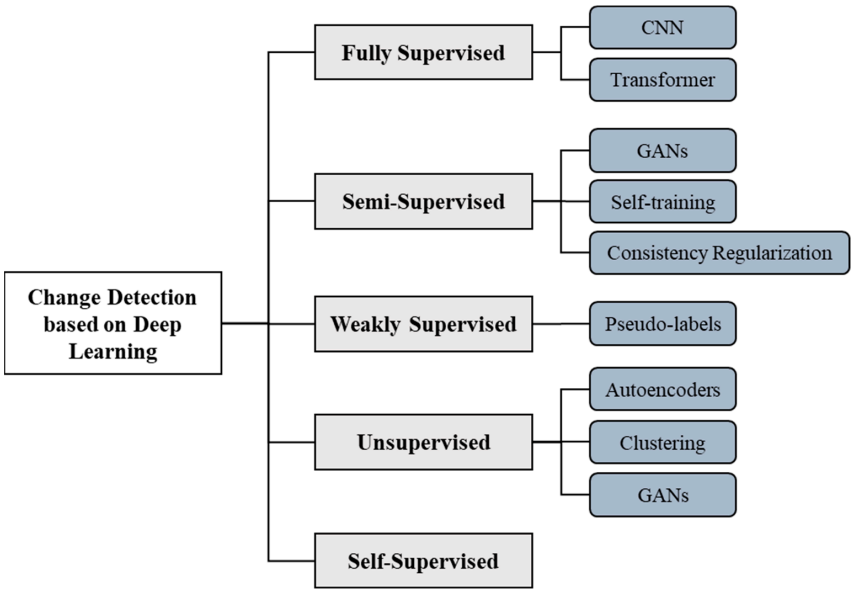

Figure 1. Taxonomy of DL-based CD.

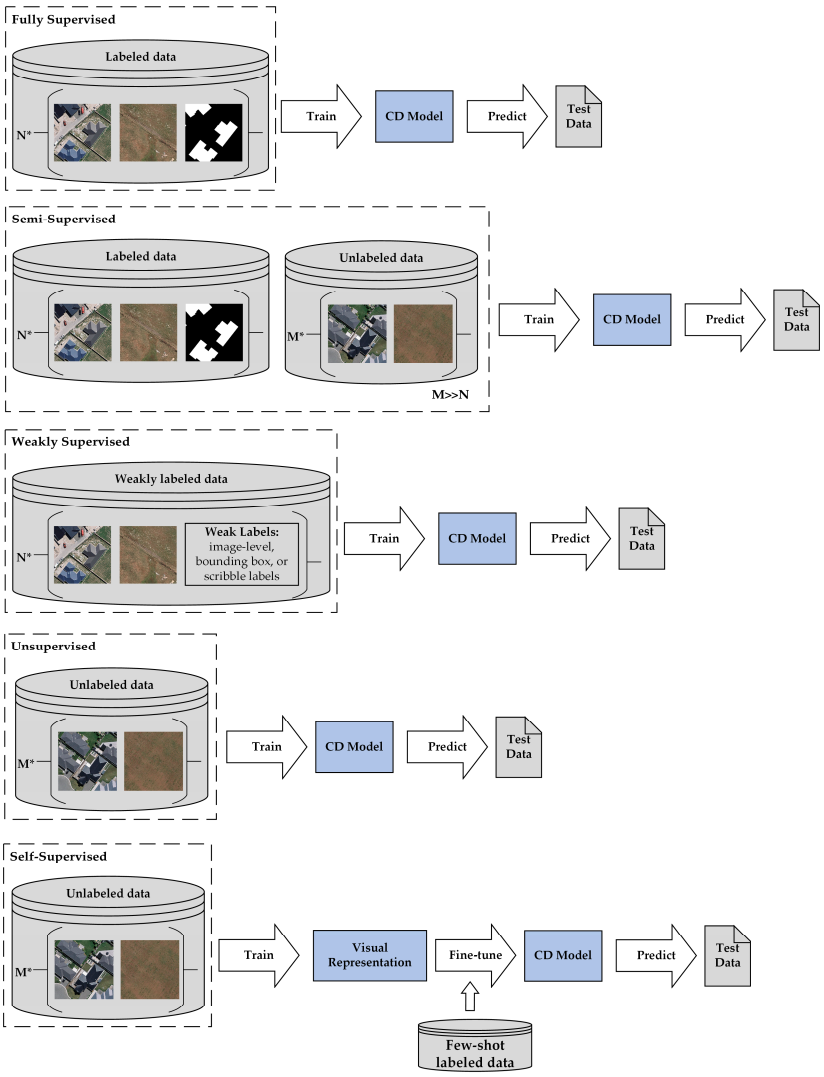

Figure 2. Schematic diagram of CD across different learning paradigms.

2. Basic Network Architectures of DL

In this section, we explore the basic DL network architectures used for CD, encompassing key structures, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), AutoEncoder (AE), and Transformer.

2.1. Convolutional Neural Network

CNN originated in the late 1980s, initially proposed by Yann LeCun with the LeNet-5 model for handwritten digit recognition [18]. With the expansion of dataset sizes and increased computational power, CNNs began to demonstrate their potent feature extraction and pattern recognition capabilities. The remarkable success of AlexNet [19] in the 2012 ImageNet image classification competition marked a significant milestone, signifying the successful application of CNNs in large-scale image recognition tasks.

The core concept of CNN is the extraction of features from input data through convolution operations. By stacking multiple convolutional and pooling layers, CNN progressively build a high-level abstract representation of the input data, as shown in Figure 3. Generally, CNNs are regarded as hierarchical feature extractors that map raw pixel intensities into feature vectors across various abstract layers. The fundamental components of CNN include:

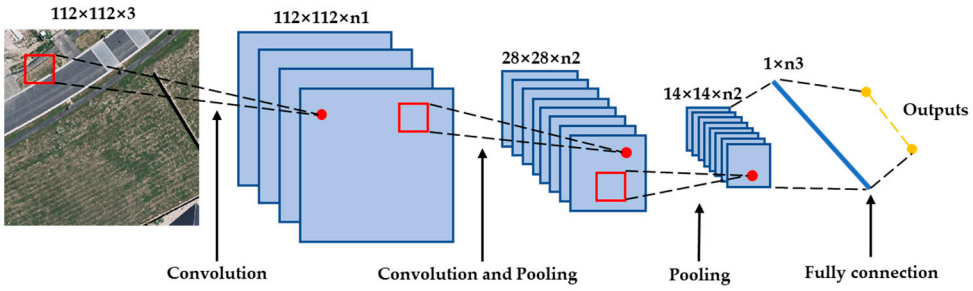

Figure 3. Schematic diagram of CNN.

Convolutional Layer. These are the essential building blocks of CNN, where convolution operations are performed to extract features from the input data. Convolution can be seen as a filtering operation that scans across the input data with sliding filters to obtain a series of local feature responses. The convolution operation is mathematically represented as:

$$(f * g)(i, j) = \sum_{m} \sum_{n} f(m, n) g(i - m, j - n) \quad (1)$$

where f represents the input data, g is the convolution kernel, and (i, j) represents the position in the output feature map. Convolution layers often involve considerations such as kernel size, which determines the dimension of the sliding filters and influences the network's ability to extract features of varying scales; stride, which indicates the distance the filter moves across the input data; and padding, which involves adding zeros around the input data to maintain its spatial dimensions after convolution.

Activation Function. Typically following the convolutional layers, activation functions introduce non-linearity, mapping inputs to a new space and allowing the network to learn complex features. The choice of an appropriate activation function, such as Sigmoid, Tanh, ReLU, or Softmax, is crucial for the network's training and performance.

Pooling Layer. Also known as subsampling, pooling layers reduce data dimensions and the number of parameters while preserving essential features, thus improving computational efficiency and mitigating overfitting. Pooling functions, like max pooling or average pooling, aggregate information within local regions to produce a reduced feature map.

Fully Connected Layer. In this layer, each neuron is connected to all neurons in the previous layer, creating a fully connected network that combines the network's high-level abstractions to form complex mappings of the input data.

CNNs play a pivotal role in the field of image processing. Their robust feature extraction and hierarchical abstraction capabilities efficiently capture both local and global information in images, enabling recognition of edges, textures, shapes, and other features. This makes them the architecture of choice for tasks such as image classification [20–23], object detection [24–26], image segmentation [27–29], and change detection [30–33].

2.2. Recurrent Neural Network

The RNN excels at processing sequential data, capturing temporal dependencies inherent within it. This makes RNNs highly effective for tasks involving sequential or time-related information. The RNN has a long history, dating back to the 1980s [34]. However, training RNNs has historically been challenging due to issues such as vanishing and exploding gradients. It was not until the early 2000s, with the development of technologies like Long Short-Term Memory (LSTM) [35] and Gated Recurrent Units (GRU) [36], that RNNs began to see broader application.

The RNN is distinguished by its feedback connections, allowing the network to transmit information over time and consider entire input sequences for prediction purposes. Specifically, at each time step, an RNN receives an input and the hidden state from the previous time step. It then performs a linear transformation via weight matrices, followed by a non-linear transformation through an activation function to generate the current time step's hidden state. This hidden state is conveyed to the network's input layer at the subsequent time step, creating a loop, as shown in Figure 4. The computations of an RNN can be described as:

$$h_t = f(W_{ih}x_t + b_{ih} + W_{hh}h_{t-1} + b_{hh}) \tag{2}$$

$$y_t = f(W_{ho} \cdot h_t + b_{ho}) \tag{3}$$

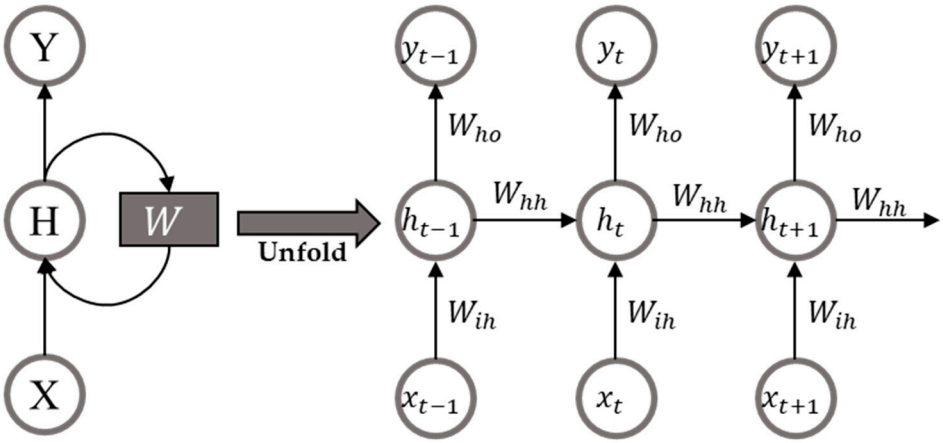

Figure 4. Schematic diagram of RNN.

Equation (2) represents the computation from the input layer to the hidden layer, where h_t is the hidden state at the current time step, x_t is the input at the current time step, W_{ih} and W_{hh} are the weight matrices for the input and the hidden state, respectively, b_{ih} and b_{hh} are the corresponding bias terms, and f is the activation function. Equation (3) represents the computation from the hidden layer to the output layer, where y_t is the output at the current time step, W_{ho} is the weight matrix connecting the hidden state to the output, and b_{ho} is the bias term.

The RNN is commonly employed in tasks that involve modeling sequential data, such as language modeling [37], machine translation [38], and time series forecasting [39]. In the realm of image processing, the RNN is typically used in conjunction with the CNN [40–42]. While the CNN effectively captures local features in image processing, the RNN leverages its sequential processing capacity to integrate global information or capture temporal

dependencies. Especially in tasks involving long-term sequential CD, the RNN plays a vital role.

2.3. AutoEncoder

The AE, originating from research in neural networks and information theory, can be traced back to the 1990s [43]. With the advancement of DL, the AE has garnered increased attention. In 2006, Hinton proposed a method of unsupervised pre-training followed by supervised fine-tuning [44], enabling the deep AE to effectively solve many practical problems.

An AE is an unsupervised learning neural network model, consisting of two main components: an encoder and a decoder, as shown in Figure 5. The encoder maps input data to a low-dimensional hidden representation, aiming to capture the most significant features of the input, as calculated in Equation (4). The decoder maps the hidden representation back to the original input space, attempting to reconstruct the original data. Its goal is to ensure that the hidden representation retains as much of the original information as possible, as described in Equation (5).

$$h = f(W_{eh} \cdot x + b_{eh}) \tag{4}$$

$$\hat{x} = g(W_{dh} \cdot h + b_{dh}) \tag{5}$$

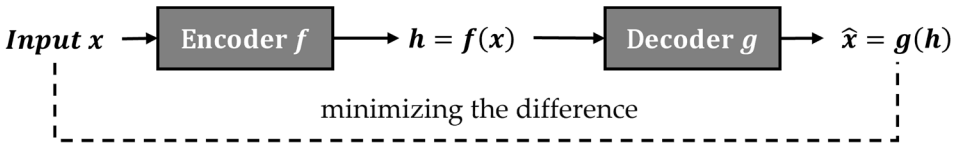

Figure 5. Schematic diagram of AE.

In these equations, x represents the input data, W_{eh}, W_{dh} are the weight matrices for the encoder and decoder, respectively, b_{eh}, b_{dh} are bias terms, f, g denotes the encoding and decoding functions, h is the hidden representation obtained, and $\hat{x}$ is the output of the decoder. A loss function is used to measure the difference between the reconstructed input and the original input, with common choices being Mean Squared Error (MSE) or Cross-Entropy (CE) loss, as shown in the following equation:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}||x_i - \hat{x}_i||^2 \tag{6}$$

The AE continually learns to extract useful features from input data and aims to accurately reconstruct the original input during the decoding phase. This capability makes the AE a powerful tool for feature learning. In the field of image processing, the AE is primarily used for feature learning and extraction [45], dimensionality reduction [46], and as an initialization tool for generative models [47].

2.4. Transformer

The Transformer, a neural network architecture introduced by Google in 2017 [48], was initially developed for natural language processing tasks, achieving remarkable success, particularly in machine translation. As research progressed, the Transformer's versatility became apparent, showcasing its powerful sequence modeling capabilities in image processing [49–52] and speech recognition [53–55] etc. This versatility has made the Transformer one of the most prominent models in the field of DL, with an increasingly broad range of applications.

At its core, the Transformer relies on the self-attention mechanism to process input sequences. This mechanism enables the model to dynamically focus on different parts of the input sequence while processing each position, thereby capturing global contex-

tual information. The Transformer comprises components such as multi-head attention and feedforward neural networks, as shown in Figure 6. It achieves efficient modeling of input sequences by stacking multiple layers of these components. The fundamental components include:

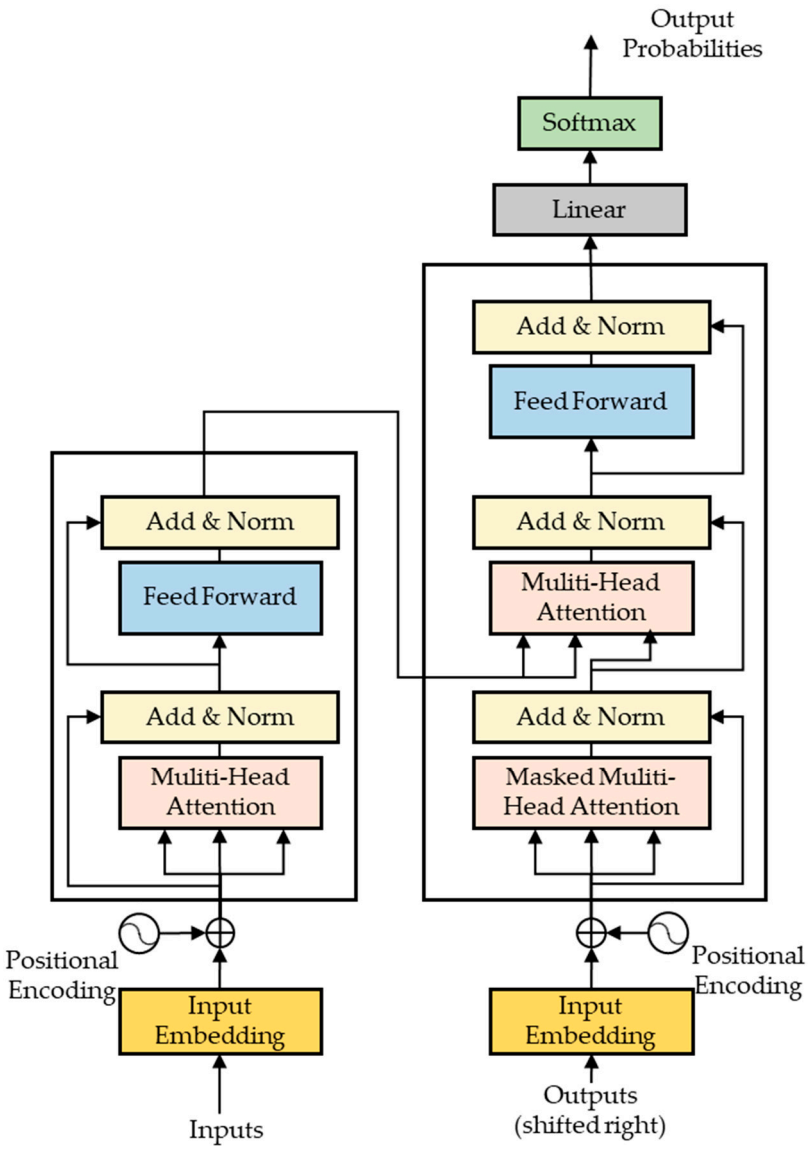

Figure 6. Schematic diagram of Transformer.

Multi-Head Attention. The Transformer introduces the multi-head attention mechanism, allowing the model to focus on different aspects in various representational subspaces. In multi-head attention, the input is mapped to different linear subspaces (heads), and attention weights are computed separately for each. The outputs from these subspaces are then concatenated and linearly transformed to produce the final output.

Positional Encoding. As the Transformer lacks a recurrent structure, it requires a method to handle the relative positional information within a sequence. Positional encoding is added to the input embeddings to provide this information. Typically, positional encoding is a matrix of the same dimension as the input, with values calculated based on position and dimension.

Feedforward Neural Network. Following the multi-head attention layer, each position passes through a feedforward neural network. This network typically consists of two linear

layers and a nonlinear activation function, independently processing the elements in the sequence at each position.

Residual Connections and Layer Normalization. To prevent gradient vanishing or explosion, each sublayer's input (such as multi-head attention and feedforward neural network) is passed through a residual connection. This means that the input is added to the output of the sublayer, preserving the original information. Additionally, each sublayer's output undergoes normalization to ensure the network's stability and convergence during training.

When the concept of the Transformer was introduced into the visual domain, the Vision Transformer (ViT) [52] redefined image classification tasks by treating images as a sequence of regular patches, providing a new paradigm for image processing. Furthermore, in object detection [56,57], segmentation tasks [58–60] and change detection [61–63], dividing image regions into sequences allows the Transformer to understand images both globally and locally, offering new perspectives and methodologies for these key tasks.

3. DL-Based CD Methods across Various Learning Paradigms

In this section, we will review and analyze DL-based methods for RS image CD from the perspectives of different learning paradigms, including fully supervised learning, semi-supervised learning, weakly supervised learning, and unsupervised learning. The discussion will encompass a variety of data sample scenarios, exploring how these learning paradigms utilize different types of data samples to address specific challenges.

3.1. Fully Supervised Learning

Fully supervised CD methods leverage multi-temporal RS imagery, richly labeled with dense change labels, to construct and train neural network models. These trained models are then applied to pairs of images with unknown labels for detecting changes. Currently, research in CD using fully supervised learning is both widespread and deeply developed. With ongoing technological advancements, two primary frameworks have emerged as particularly successful in this field: those based on CNN and those based on Transformers. These frameworks have achieved significant accomplishments in the field of CD.

3.1.1. Fully Supervised CD Methods Based on CNN

Fully supervised CD methods based on CNN typically utilize an encoder–decoder architecture. This structure allows for efficient feature extraction from input data and precise reconstruction. In the encoding phase, layers of convolutions and pooling gradually map the raw data into a high-dimensional feature representation, effectively capturing spatial and semantic information. The decoding phase, through deconvolution or up=sampling operations, reconstructs these high-dimensional features into segmentation results matching the input data. Numerous image segmentation networks based on this encoder–decoder structure, such as U-Net [29], U-Net++ [64], FCN [65], SegNet [66], Deeplab [28], and PSPNet [27], have been successfully applied to CD tasks.

Early methods [67–72] based on CNN typically began with image fusion of bi-temporal RS imagery, as illustrated in Figure 7, using techniques like direct stacking [73], differencing [74], or principal component analysis (PCA) [75]. These fused images were then input into DL models with single input channels to achieve CD. With further research, Siamese network structures have become mainstream [30–33,76–83]. In these architectures, bi-temporal images are processed through feature encoders with identical structures, enabling the fusion of features at the level of the feature maps. This method allows models to better understand and capture changes between the bi-temporal images, thus enhancing the accuracy and robustness of CD. Additionally, designing mechanisms for feature transfer or deep feature fusion methods between the encoder and decoder helps preserve more contextual information, which is crucial for improving CD accuracy. Integrating various attention mechanisms for feature transfer or fusion has become a prevalent approach, including channel attention [84–86], spatial attention [87,88], and channel-spatial attention [89,90].

These mechanisms assign different weights to each input element, emphasizing important features and enabling the model to focus on task-relevant aspects, reducing sensitivity to irrelevant features. The Siamese structure has shown excellent performance in CD tasks, becoming a significant research direction in the field.

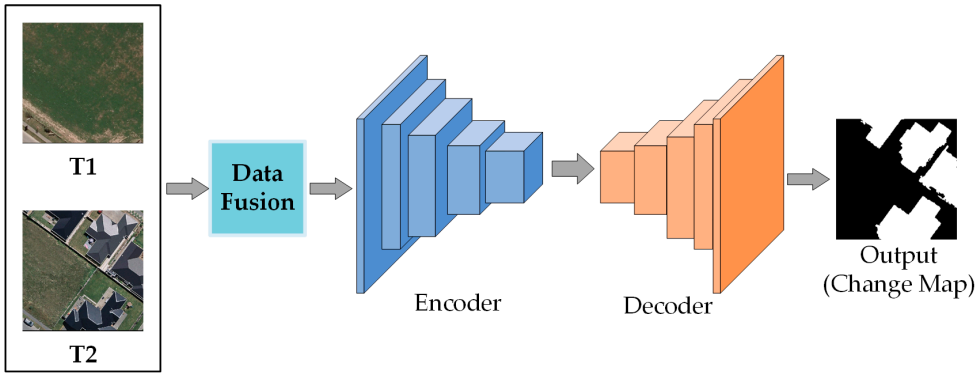

Figure 7. Schematic diagram of an Encoder–Decoder structure.

Consequently, current mainstream methods based on CNN typically rely on Siamese network encoder–decoder structures, combined with attention mechanisms at the core of feature transfer and fusion modules. Figure 8 showcases a basic network structure, building upon this network with tailored encoder–decoder structures, feature transfer and fusion modules, and other improvements can achieve higher accuracy in CD and exhibit stronger applicability and effectiveness in practical applications. For example, Chen et al. [30] designed a spatio-temporal attention neural network based on a Siamese network, inputting bi-temporal images into two branches of a shared-weight Siamese network to capture rich spatio-temporal features using the correlation between spatio-temporal pixels. They integrated the attention module into a pyramid structure to capture spatio-temporal dependencies at various scales and generate feature maps of bi-temporal images, subsequently achieving refined CD results through deep metric learning. Shi et al. [31] first used a Siamese network to learn the nonlinear transformation from input images to embedding space, then applied convolutional attention mechanisms to extract more discriminative features, employing a metric module to learn the change map and a deep supervision module [91] to enhance the feature extractor's learning capability. They also used a contrastive loss function to encourage smaller distances between unchanged pixels and greater distances between changed ones. Fang et al. [32] proposed a densely connected Siamese network, stacking feature maps after feature extraction through the network, and utilizing attention modules to capture relationships between pixels at different times and locations, thus generating more distinctive features. Li et al. [33] introduced a novel lightweight network, A2Net, using a shared-weight MobileNetV2 [92] to extract deep features from images. They incorporated a neighborhood aggregation module (NAM) to fuse features from adjacent stages of the backbone network, enhancing the representation of temporal features. The progressive change identification module (PCIM) was proposed to extract temporal difference information from bi-temporal features, and the supervised attention module was used for reweighting features, effectively aggregating multi-level features from high to low levels. Similarly, Zhu et al. [79] used an encoder–decoder-based Siamese network to extract features from bi-temporal images and introduced a global hierarchical sampling mechanism for balanced training sample selection. Additionally, they incorporated a binary change mask into the decoder to reduce the influence of unchanged background areas on changed foreground areas, further enhancing detection accuracy.

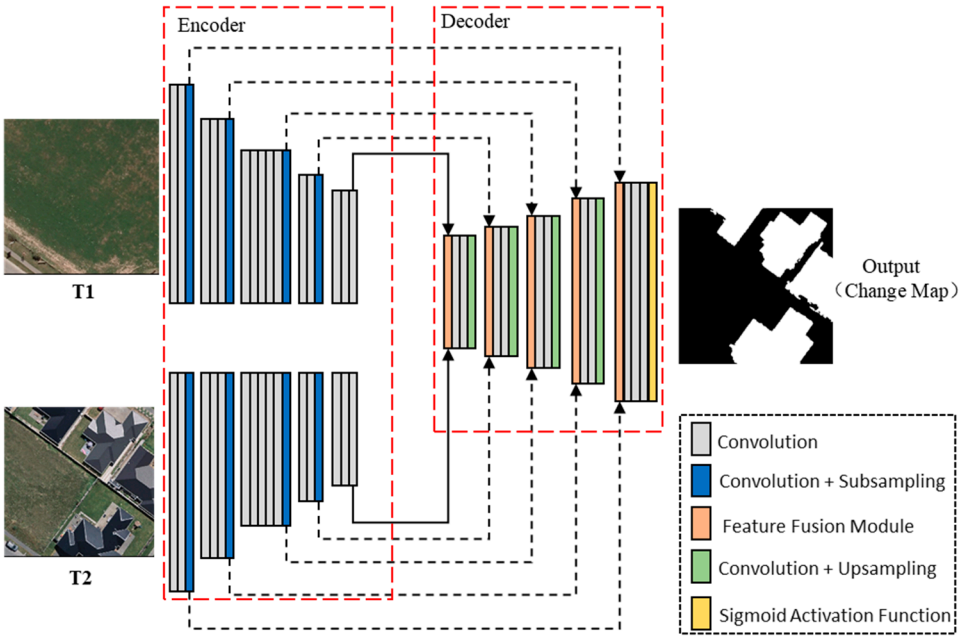

Figure 8. Schematic diagram of mainstream CNN-based CD basic network.

3.1.2. Fully Supervised CD Methods Based on Transformer

In 2020, Google's research team introduced the Vision Transformer (ViT) model, pioneering the application of the Transformer architecture to computer vision tasks. They demonstrated, through extensive experiments on large-scale image datasets, that ViT could achieve performance on par with CNN in image classification tasks, marking the official entry of Transformers into the visual domain. This milestone sparked a wave of development in Transformer-based visual models, such as DeiT [50], Swin Transformer [51], Twins [93], PVT [49], Cait [94], TNT [95], and SETR [28]. These advancements have positioned Transformers as one of the focal points in the research into visual tasks.

In CD tasks, fully supervised methods based on Transformers have achieved significant results. Typically combined with CNN, these methods leverage the CNN's prowess in extracting local features and image details while utilizing the Transformer's ability to capture global dependencies and contextual information. This dual approach enables a comprehensive understanding of complex changes in remote sensing imagery, leading to superior performance in CD. Common model structures, similar to those used in semantic segmentation tasks within computer vision, involve using the Transformer as a feature extractor. Post-extraction, a decoder, either CNN- or Transformer-based, maps the features back to the size of the input image to produce the CD output. Most methods are built on a basic architecture, as shown in Figure 9, and can be categorized into two types based on the decoder used:

- Transformer Encoder + Transformer Decoder [61–63,96–100]. This design fully exploits the Transformer's self-attention mechanism in both encoding and decoding phases, effectively integrating global information during up-sampling in the decoding process. Additionally, this all-attention architecture maintains efficiency in handling long-distance dependencies and large-scale contextual information, especially in parsing complex remote sensing data structures. For example, Cui et al. [61] proposed SwinSUNet, a pure Transformer network with a Siamese U-shaped structure, comprising encoders, fusers, and decoders, all based on Swin Transformer blocks. The encoder uses hierarchical Swin Transformers to extract multi-scale features, while the fuser primarily merges bi-temporal features generated by the encoders. Similar to the encoder, the decoder, also based on hierarchical Swin Transformers, uses up-sampling to restore the feature map to the original input image size and employs linear projection for dimensionality reduction to generate the CD map. Chen et al. [98] introduced

an RS image CD framework based on bi-temporal image Transformers. This uses Siamese CNN to extract high-level semantic features and spatial attention to convert each temporal feature map into a compact processing unit (token) sequence. The Transformer encoder then models the context of these two token sequences, generating context-rich tokens. An improved Transformer decoder reprojects these back into pixel space, enhancing the original pixel-level features. Finally, a feature difference map is computed from the two refined feature maps and input into a shallow CNN to produce the CD map.

- Transformer Encoder + CNN Decoder [101–107]: In this configuration, the Transformer encoder acts as the feature extractor, capturing the global contextual information of the input data. The extracted features are then passed to a CNN decoder for more refined image segmentation and reconstruction. For instance, Li et al. [102] proposed TransUNetCD, an end-to-end CD model combining Transformer and UNet. The Transformer encoder, based on the UNet architecture, encodes feature maps obtained from Siamese CNN, models the context, and extracts rich global contextual information. The CNN-based decoder up-samples the encoded features and integrates them with high-resolution, multi-scale features through skip connections. This process learns local–global semantic features, restoring the feature map to the original input image size to generate the CD map.

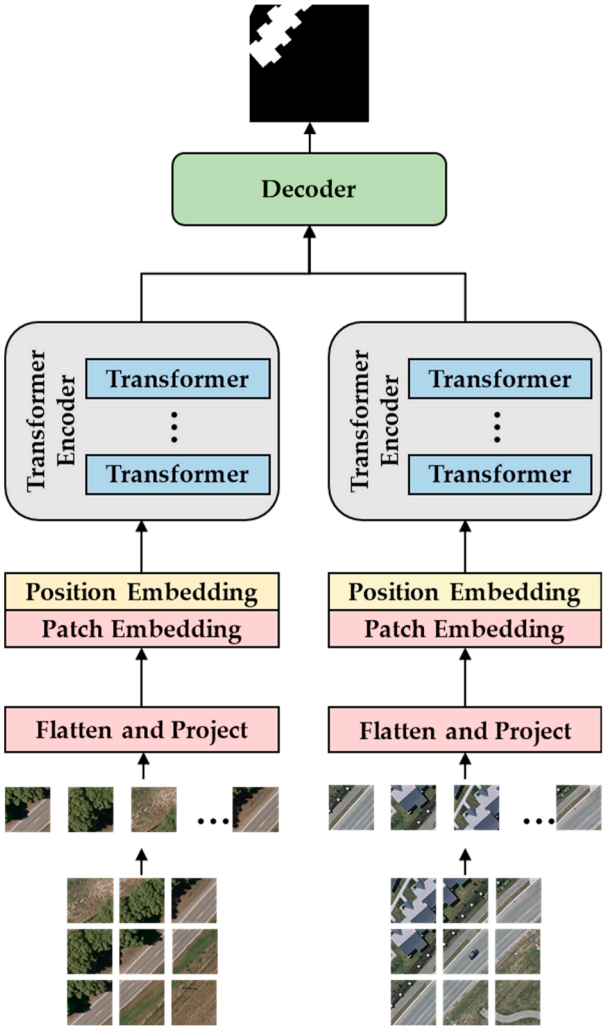

Figure 9. Schematic diagram of mainstream Transformer -based CD basic network.

Beyond the common method previously mentioned, there are additional Transformer-based model architectures. For instance, Bandara et al. [108] utilized a hierarchical Transformer solely as an encoder to extract features from bi-temporal images, followed by employing a lightweight Multi-Layer Perceptron (MLP) as the decoder. Additionally, another method [109,110] employs the Transformer as a key metric module positioned between the CNN encoder and CNN decoder, optimizing the depth features.

3.2. Semi-Supervised Learning

Semi-supervised learning sits at the intersection of supervised and unsupervised learning paradigms. It trains models using both labeled and unlabeled samples, extracting features in a supervised manner from labeled samples while also employing various strategies to expand features from unlabeled ones. This approach facilitates the construction of more effective models while reducing reliance on extensive labeled data, making it highly practical in real-world scenarios, where acquiring a large volume of fully labeled data is often challenging.

In the context of rapid advancements in RS technology and the accumulation of vast amounts of unlabeled multi-temporal RS imagery, semi-supervised learning methods have emerged as a viable and promising research approach for CD tasks. Currently, semi-supervised CD methods can be categorized into three types: those based on adversarial learning, self-training, and consistency regularization.

3.2.1. Semi-Supervised CD Methods Based on Adversarial Learning

These methods are developed based on generative adversarial networks (GANs) [111]. The training process of GANs is an optimization problem where the generator aims to create increasingly realistic samples to deceive the discriminator, which strives to differentiate between real and generated samples. The model's performance is enhanced by minimizing the adversarial loss between the generator and the discriminator.

In semi-supervised CD tasks, the key to these methods lies in using the discriminator to distinguish between actual change maps and those generated by the CD network. Specifically, the discriminator is adversarially trained alongside the CD segmentation model, with the objective of accurately differentiating between true and predicted labels. During this process, the discriminator can produce prediction confidence maps for unlabeled samples. By selecting highly confident unlabeled samples and incorporating them into the training, the model gains more information, thus enhancing its predictive capabilities for changes. The basic framework of this approach is illustrated in Figure 10. In the field of CD, semi-supervised learning methods based on adversarial learning are still in their nascent stage. Jiang et al. [112] initially trained a GAN model, then connected two identically trained discriminators in parallel to extract features from bi-temporal images. The outputs of these discriminators were concatenated into a vector as the final output, which was then fine-tuned using a subset of labeled data to derive the CD model. Yang et al. [113] followed the basic framework shown in Figure 10 but did not adopt a confidence strategy; instead, they directly incorporated all unlabeled samples into the training process. Peng et al. [114] introduced SemiCDNet, which inputs labeled and unlabeled samples into a CD segmentation network to generate initial predictions and entropy maps. It then employs two discriminators to reinforce the consistency of feature distribution between change segmentation maps and entropy maps. The final model is trained by combining supervised loss, segmentation adversarial loss, and entropy adversarial loss.

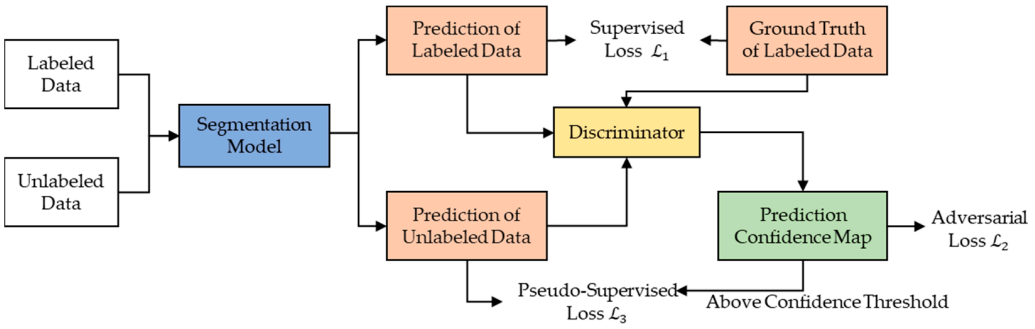

Figure 10. Basic framework of semi-supervised learning based on GANs.

3.2.2. Semi-Supervised CD Methods Based on Self-Training

The core concept of these methods is as follows: initially, model parameters trained on labeled samples are used to predict unlabeled samples, treating these predictions as pseudo-labels for the unlabeled samples. Subsequently, the training dataset is augmented with these unlabeled samples and their corresponding pseudo-labels, and the model is retrained on this expanded dataset. The general workflow of self-training methods includes:

1. Initialization: Train the initial model using the available labeled dataset.
2. Pseudo-Label Generation: Predict unlabeled samples using the initial model, select those with high prediction confidence, and assign the prediction results as their pseudo-labels.
3. Model Re-Training: Merge unlabeled samples with pseudo-labels into the labeled dataset to form an expanded training set and retrain the model using this dataset.

This process is iteratively repeated, generating new training data with pseudo-labels from unlabeled samples in each iteration. However, in practical applications, the self-training process can introduce noise, particularly regarding the reliability of pseudo-label generation. If the pseudo-labels are not reliable, they might adversely affect model training. Hence, additional strategies are often employed, such as implementing effective confidence strategies [115] to filter unlabeled samples and enhance the stability and effectiveness of self-training. For instance, Wang et al. [116] select reliable unlabeled samples based on their prediction stability across different training checkpoints and the stability between class activation maps and prediction results within the model. Yang et al. [117] proposed using checkpoints set in the early, middle, and late stages of training, selecting reliable unlabeled samples for self-training based on the stability of predictions from different checkpoints. Wang et al. [118] built upon this stability with different checkpoints and designed a positive–negative pixel contrast loss to enhance the model's ability to extract change features. Sun et al. [119] utilized the confidence threshold filtering from FixMatch [115] to select reliable unlabeled samples for self-training, further enhancing model performance and robustness by enforcing consistency between CD results from distorted images and pseudo-labels during the self-training phase. In addition to these strategies, many other methods [120–123] have been proposed to improve the effectiveness of self-training, and their efficacy in the domain of semi-supervised CD is a subject worthy of deeper exploration.

3.2.3. Semi-Supervised CD Methods Based on Consistency Regularization

The essence of consistency regularization methods lies in encouraging the model to produce similar outputs for the same sample subjected to different perturbations or transformations. Based on two key hypotheses, smoothness and clustering, consistency regularization operates as follows. The smoothness hypothesis posits that closely situated samples are likely to share the same label. As illustrated in Figure 11, in a feature space, samples of the same category are usually closer to each other than to samples of different categories. This implies that models should offer similar predictions for neighboring samples. The clustering hypothesis suggests that decision boundaries should lie in low-density regions. An effective decision boundary (like the solid line in Figure 11) should ideally

pass through the sparsest areas of the sample space, reducing the model's sensitivity to noise and irrelevant features, thereby enhancing stability and accuracy. On these foundations, consistency regularization methods incorporate unlabeled samples into the training process, expanding the model's feature space by constraining the consistency of various perturbations or transformations of these samples. This leads to more generalized feature representations and improved model performance.

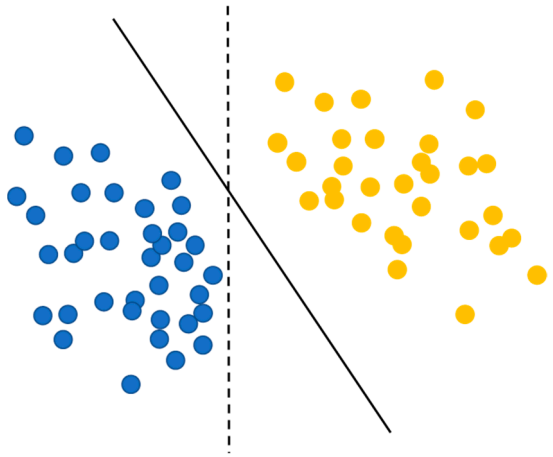

Figure 11. The fundamental hypotheses of Consistency Regularization.

The critical aspect of consistency regularization methods is how to obtain the perturbation space of unlabeled samples and, once obtained, how to train a model that is insensitive to these perturbations, ensuring consistency in predictions for the original image and its various perturbed spaces. Currently, consistency regularization methods, categorized into three types based on the perturbation space, are illustrated in terms of their basic framework for different perturbation spaces in Figure 12.

- Image Perturbation Space [116,119,124,125]. This approach involves applying operations, like rotation, scaling, and color transformations, to images, generating a series of perturbed images. For example, Sun et al. [124] proposed a semi-supervised CD method using data augmentation strategies to access image perturbation space and generate pseudo bi-temporal images to further expand this space. The method then minimizes the differences between the change maps obtained from the image perturbation space and the original images.
- Feature Perturbation (FP) Space [126,127]. This involves perturbing the internal feature space of the image within the model, rather than directly manipulating the image itself. This can be achieved through operations like dropout on features. For instance, Bandara et al. [126] introduced a semi-supervised CD method based on feature consistency regularization. The method perturbs the deep feature space of bi-temporal difference features of unlabeled image pairs, minimizing the differences between change maps derived from various feature perturbation spaces and the original space as a consistency loss.
- Model Perturbation Space [128,129]. This approach involves altering the model itself to create pseudo-labels for unlabeled samples using different models and then supervising them mutually. For example, Chen et al. [129] used two networks with the same structure but different initializations during model training. They added a loss function to ensure that both networks produce similar outputs for the same sample.
- Combined Perturbation Space [130]. This approach synergizes elements from Image Perturbation Space, Feature Perturbation Space, and Model Perturbation Space. Yang et al. [130] effectively merged image perturbation techniques with feature perturbation strategies, an integration that led to the exploration of a broader perturbation space and yielded models with superior performance and enhanced generalization

capabilities. Notably, their method demonstrated commendable results in CD datasets, underscoring the benefits of this integrated perturbation approach.

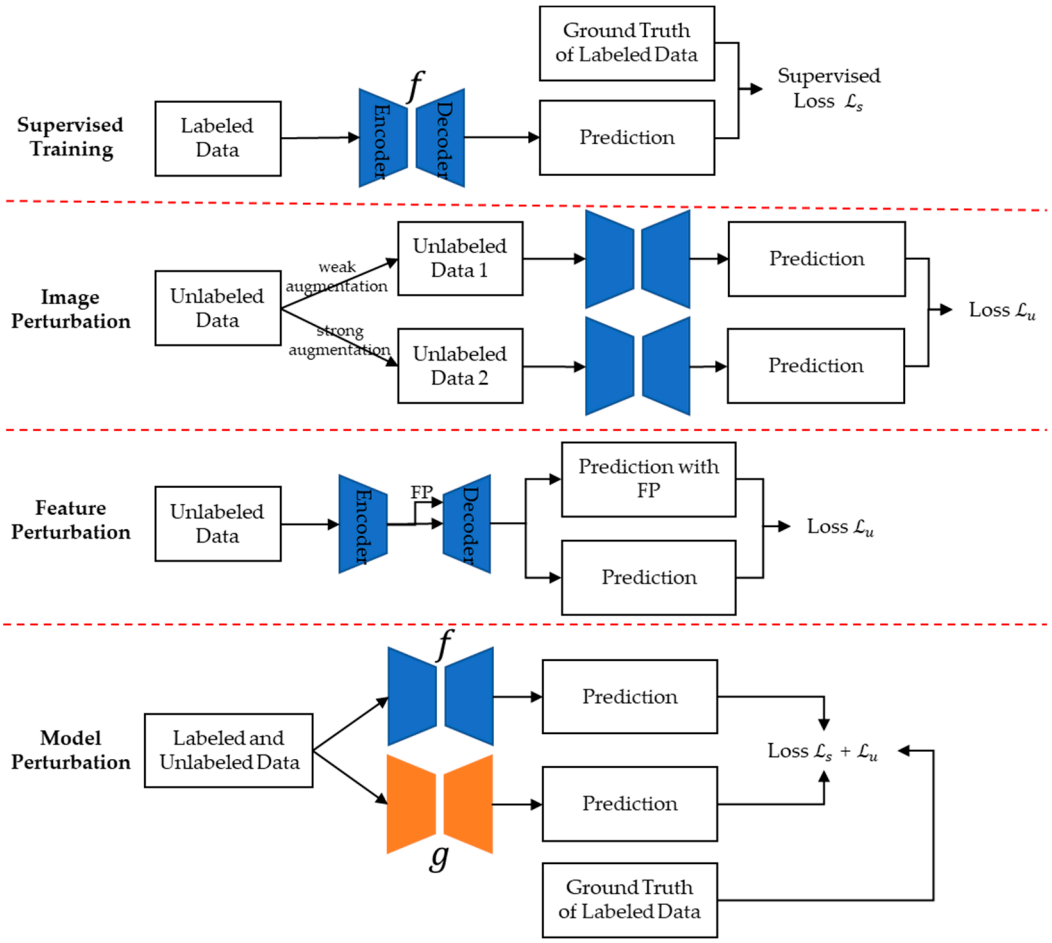

Figure 12. Basic framework of semi-supervised learning based on Consistency Regularization across different perturbation spaces.

3.3. Weakly Supervised Learning

Weakly supervised learning methods involve training models with incomplete or imprecise labeling information, subsequently employing these models to achieve pixel-level predictions for unlabeled samples. The labels in these methods are usually in a "weaker" form, such as image-level, bounding box, or scribble labels, rather than detailed, pixel-level labels, which is often prohibitively expensive or unattainable due to limitations in domain-specific knowledge. Compared to fully supervised learning, weakly supervised learning provides greater flexibility in acquiring labels, making it highly applicable and promising across diverse real-world scenarios.

Weakly supervised learning typically follows two core steps, as shown in Figure 13:

- Step 1: Extract information from incomplete or imprecise labels to generate pixel-level pseudo-labels.
- Step 2: Utilize these pseudo-labels to train a pixel-level CD model.

The central challenge in these steps is the generation of high-quality pseudo-labels. Among all weak labels, image-level labeling is a relatively cost-effective option. It requires only semantic labeling of each image pair as changed or unchanged, without the need for pixel, region, or boundary labels. The process of generating pseudo-labels includes:

1. Firstly, creating initial change areas from the image-level labels for each image pair.

2. Then, propagating semantic information from these initial areas across the entire image pair to generate pixel-level pseudo-labels.

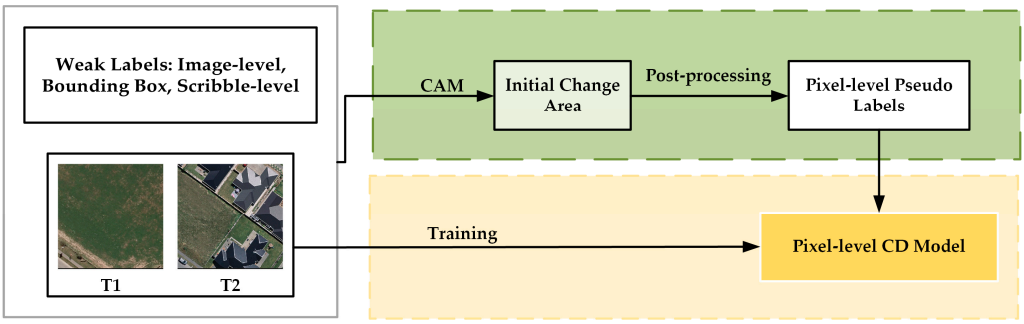

Figure 13. General steps of weakly supervised learning.

The process of developing initial change areas involves evolving image-level labels into scribble-level or bounding box-level labels. Hence, this paper will focus in detail on image-level weakly supervised learning.

The method for generating initial change areas generally involves training an image-level classifier with image-level labels and extracting information from the classifier's deep features to create initial change areas. As Shen [131] suggests, this step often embodies a concept similar to Class Activation Mapping (CAM) [132]. CAM can locate specific areas in an image associated with changes, serving as the initial change areas for weakly supervised methods. The basic workflow of CAM, as shown in Figure 14, assumes that the feature map of the last convolutional layer is F with N channels. Let the weights of the model's final fully connected layer be w, then the process of generating CAM can be represented by the following formula:

$$L_c = \sum_i w_i F_i \tag{7}$$

where L_c represents the Class Activation Maps (CAMs). There are many variants of CAM, such as Grad-CAM [133], Grad-CAM++ [134], Score-CAM [135], LayerCAM [136], and EigenCAM [137], offering more flexible ways to generate initial change areas.

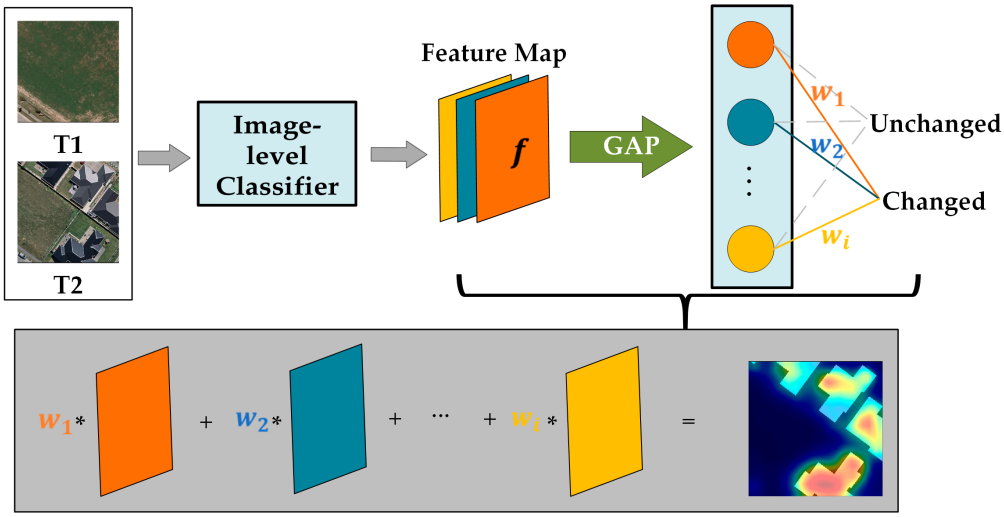

Figure 14. Basic workflow for generating CAMs.

In the field of weakly supervised CD, current methods for propagating initial change areas across the entire image pair to generate pixel-level pseudo-labels predominantly rely on relatively traditional post-processing techniques such as PCA [138], K-Means [139], and conditional random fields (CRF) [140]. For instance, Kalita et al. [141] trained a Siamese

CNN classification network using image-level labels to obtain deep features of image pairs and generate change localization maps. They then applied PCA and K-Means methods to segment these maps for pixel-level CD results. Jiang et al. [142] trained an image-level CNN model with weighted global average pooling, also obtaining change localization maps, and then used CRF to refine the boundaries of these maps for CD. Andermatt et al. [143] proposed a weakly supervised convolutional network that utilizes a feature comparator to obtain change features, ultimately forming pixel-level CD results through a change segmentation module composed of residual blocks and CRF–RNN.

In the image segmentation domain, we have witnessed the emergence of novel post-processing strategies for initial areas, such as cross-consistency [144,145], pixel relationships [146,147], and affinity learning [148–150]. Huang et al. [151] attempted to extend some of these new methods from weakly supervised semantic segmentation to weakly supervised CD, including SGCD [152] and AFA [150], achieving promising results. However, whether these new strategies are suitable for RS imagery, or if further development of more fitting post-processing methods for initial change areas in RS imagery is required, remains an area for further research.

Beyond the mainstream methods, scholars have also explored the application of other approaches to weakly supervised CD. For example, Wu et al. [153] proposed a GANs-based weakly supervised CD framework, which deceives discriminators into predicting image pairs with masked change areas as unchanged. They use these masked areas as pseudo-labels to train segmentation networks, forming a robust CD segmentation network through iterative adversarial learning. Additionally, Zhang et al. [154] introduced a novel neural network that combines CD with multiple instance learning for landslide detection.

3.4. Unsupervised Learning

Unsupervised DL methods for CD employ deep neural networks to autonomously learn image features, facilitating CD without prior knowledge or manual intervention. These unsupervised methods, not requiring any labeled data, leverage raw imagery for training, offering higher automation and broader adaptability.

Unsupervised CD methods generally combine DL networks with traditional CD techniques. The primary concept involves extracting effective feature representations using DL networks and applying traditional CD methods for post-processing to obtain CD maps. For instance, LV et al. [155] first used linear iterative techniques to obtain super-pixels, then proposed a feature extraction network based on stacked contractive AE (sCAE) [156] to learn advanced encoded features, utilizing the k-means method for binary classification of these high-level encoded features to achieve final CD results. Luppino et al. [157] introduced an unsupervised heterogeneous data CD method that uses local information extracted from input imagery to align two AEs for CD. Bergamasco et al. [158] proposed an unsupervised deep neural network method based on multi-layer convolutional AE (CAE) [159], using single-temporal image blocks to train the CAE. The feature representations are learned by minimizing the reconstruction error between inputs and outputs. The trained CAE is then used to extract multi-scale features from pre- and post-change images, and these features are fused using a detail-preserving, scale-driven approach to generate CD maps. Saha et al. [160] described a deep change vector analysis (DCVA) for Very High Resolution (VHR) image CD, initially extracting deep features from a pre-trained multi-layer CNN. By combining features from different CNN layers to form a deep feature hyper-vector, spatial contextual information of images is captured. Deep change hyper-vectors of bi-temporal images are computed using CVA with threshold constraints to produce CD maps. Wu et al. [161] applied kernel PCA (KPCA) [162] convolution as a basic module in a Siamese structure to extract deep-level features of images. Channel differences are used to obtain feature differential maps, which are then mapped to a two-dimensional polar domain, with unsupervised clustering techniques employed to obtain CD results. Du et al. [163] initially used CVA for pre-detection, treating invariant pixel pairs as training input for the deep network. Upon network convergence, transformed features are input into a slow

feature analysis for difference calculation, followed by Chi-square distance computation for change intensity mapping and, finally, thresholding methods are applied for final CD results. Gong et al. [164] used stacked AE (SAE) to transform differential images into feature space, subsequently establishing pseudo-labels through clustering methods for training a CNN-based CD network. Zhang et al. [165] captured information of change and invariant areas through deep belief networks to generate feature spaces, followed by a feature change analysis network to identify changes.

Moreover, GANs have found applications in unsupervised CD. They employ adversarial learning between generators and discriminators to facilitate image transformation, enhancement, or reconstruction, while assessing differences between images using discriminator outputs or feature distances. For instance, Gong et al. [166] initially utilize conventional methods like CVA, PCA, and IR-MAD [167] for initial CD, subsequently training the discriminator to learn the distribution and the correlation of change pixels from both initial CD results and generator-produced CD maps. This process enables the generator to create more refined CD outputs. Gong et al. [168] start with CVA and Otsu methods for initial CD, followed by generating additional training samples using a generator. These samples, along with the initial detection results, are fed into a discriminative classification network (DCN) [169] to learn the concept of changed and unchanged pixels. The adversarially trained generated data approaches real labels, allowing the well-trained DCN to categorize original image data into changed and unchanged pixels, completing the CD process. Noh et al. [170] introduced an unsupervised CD method based on image reconstruction loss, which is trained solely on single-temporal images. It inputs both source and optically transformed images into an encoder–decoder-based GAN, training the model to reconstruct the original source image. During inference, the model receives bi-temporal images, where areas of change exhibit higher reconstruction loss. Wu et al. [153] developed an approach based on the assumption that unchanged landscapes exhibit certain spectral, spatial, and semantic similarities across multi-temporal images. They transformed the CD task into identifying a minimal region on an image that, once having masked this region, allows a GAN's generator to predict it accurately as another image.

4. Discussion of Different Learning Paradigms for CD

In this section, we embark on a comprehensive and progressive discussion of CD across various learning paradigms. The discussion is structured into four key aspects. First, we delve into publicly available fully supervised datasets, exploring how they can be adapted to generate datasets suitable for other learning paradigms. Next, we provide a comparative analysis of state-of-the-art (SOTA) methods within these paradigms. Thirdly, we summarize the advantages and disadvantages inherent to each learning paradigm, offering a balanced perspective that evaluates their applicability in the context of CD. Finally, we discuss the specific application scenarios for each paradigm. This multifaceted exploration aims to offer a deeper understanding of the current landscape in CD methodologies and to inspire future directions in this dynamic field.

4.1. Adaptation of Datasets for Various Learning Paradigms

In this subsection, we focus on several widely-used publicly available datasets that play a pivotal role in the field of CD. These datasets provide researchers with abundant experimental materials to validate and assess the performance of various CD methods. Through analysis and application of these datasets, researchers can better understand the problems and challenges in remote sensing image CD and develop more effective solutions. Table 1 presents some representative open datasets, primarily sourced from GitHub (https://github.com/wenhwu/awesome-remote-sensing-change-detection, accessed on 23 January 2024), along with additional datasets from various other sources. This paper compiles and summarizes information about these datasets, including their image types, image resolutions, number of image pairs, acquisition years, coverage areas, and data sources. Additionally, the datasets have been categorized based on their data types.

Table 1. A list of publicly available datasets for CD.

Dataset Name	Image Type	Resolution	Number of Image Pair	Acquisition Year	Coverage Area	Image Source
HRCUS-CD [171]	RGB	0.5 m	11,388 pairs of 256 × 256 pixels	2010 to 2022	Zhuhai, China	-
GVLM [172]	RGB	0.59 m	17 pairs of varying sizes	2010 to 2021	Global	Google Earth
EGY-BCD [173]	RGB	0.25 m	6091 pairs of 256 × 256 pixels	2015 to 2022	Egypt	Google Earth
SI-BU [174]	RGB	0.5–0.8 m	4932 pairs of 512 × 512 pixels	2019 to 2021	Guiyang, China	Google Earth
BANDON [175]	RGB	0.6 m	2283 pairs of 2048 × 2048 pixels	-	Some cities in China	Google Earth, Microsoft Virtual Earth, ArcGIS
DynamicEarthNet [176]	RGB	3 m	730 pairs of 1024 × 1024 pixels	2018 to 2019	75 regions worldwide	Planet Labs
CLCD [177]	RGB	0.5–2 m	600 pairs of 512 × 512 pixels	2017 to 2019	Guangdong Province, China	GF-2
S2Looking [178]	RGB	0.5–0.8 m	5000 pairs of 1024 × 1024 pixels	Spanning 1–3 years	Global	-
SYSU-CD [31]	RGB	0.5 m	20,000 pairs of 256 × 256 pixels	2007 to 2014	Hong Kong, China	-
DSIFN [78]	RGB	-	3940 pairs of 512 × 512 pixels	-	Six cities in China	Google Earth
SenseEarth2020	RGB	0.5–3 m	4662 pairs of 512 × 512 pixels	-	-	-
Google Dataset [114]	RGB	0.55 m	1067 pairs of 256 × 256 pixels	2006 to 2019	Guangzhou, China	Google Earth
LEVIR-CD [30]	RGB	0.5 m	637 pairs of 1024 × 1024 pixels	Spanning 5–14 years	Texas, USA	Google Earth
HRSCD [179]	RGB	0.5 m	291 pairs of 10,000 × 10,000 pixels	2005 to 2012	France	IGN
WHU-CD [180]	RGB	0.075 m	One pair of 15,354 × 32,507 pixels	2012 to 2016	New Zealand	Aerial
CDD [181]	RGB	3–100 cm	16,000 pairs of 256 × 256 pixels	-	-	Google Earth
SZTAKI [182]	RGB	1.5 m	13 pairs of 952 × 640 pixels	Spanning 5–23 years	-	-
Hyperspectral CDD [183]	Hyperspectral	-	Three pairs of varying sizes	2004 to 2014	USA	AVIRIS
River dataset [184]	Hyperspectral	30 m	One pair of 463 × 241 pixels	2013.5–2013.12	Jiangsu Province, China	EO-1 Hyperion
MtS-WH [185]	Multispectral	1 m	One pair of 7200 × 6000 pixels	2002 to 2009	Wuhan, China	IKONOS
OSCD [186]	Multispectral	10–60 m	24 pairs	2015 to 2018	Global	Sentinel-2
SMARS [187]	RGB, DSM	0.3 m/ 0.5 m	Two pairs of 5600 × 5600 pixels, one pair of 4500 × 3560 pixels	-	Paris and Venice	Synthetic
LEVIR-CC [188]	RGB, Natural Language	-	10,077 pairs of 512 × 512 pixels, 50,385 natural language statements	Spanning 5–14 years	Texas, USA	Google Earth
MSBC [189]	RGB, Multispectral, SAR	2 m	3769 pairs of 256 × 256 pixels	2018 to 2019	Guigang, Guangxi, China	GF-2, Sentinel-1, Sentinel-2A
MSOSCD [189]	RGB, Multispectral, SAR	-	5107 pairs of 256 × 256 pixels	2015 to 2018	Global	Sentinel-1, Sentinel-2

The datasets presented in Table 1 are characterized by their dense labeling, rendering them particularly suitable for fully supervised CD methods. However, it is essential to recognize that the use of these datasets extends beyond fully supervised learning. Adaptations of these fully labeled datasets form the basis for semi-supervised, weakly supervised, unsupervised, and self-supervised learning paradigms, each undergoing specific modifications to meet their unique requirements:

- In the case of semi-supervised learning, a subset of the data (typically around 5% to 10%) is used as labeled data, with the remainder serving as unlabeled data.
- In the case of weakly supervised learning, weak labels are generated from these precise labels. The transition from dense to weak labels is made by transforming the detailed annotations into more generalized or less informative labels.
- For unsupervised learning, the original labels of the dataset are completely disregarded.
- Furthermore, in self-supervised learning, the focus is on exploiting the unlabeled dataset for primary model training. This is followed by a fine-tuning phase, wherein a minimal subset of the data (approximately 1%) with labels is employed to refine the model's performance.

4.2. Analysis of SOTA Methods for Different Learning Paradigms

In this subsection, our focus shifts to an in-depth analysis of the highest achievable accuracy across various learning paradigms. To conduct this evaluation, we utilize the widely recognized WHU–CD dataset [180], a benchmark in the field that facilitates a comprehensive assessment. The significance of selecting the WHU–CD dataset lies in its extensive use in current research encompassing fully supervised, semi-supervised, weakly supervised, and unsupervised learning paradigms. It provides a common ground for comparing the efficacy of different learning approaches, ensuring consistency and reliability in the evaluation of accuracy metrics. The accuracies achieved by each learning paradigm are detailed in Table 2, offering a direct comparison and quantitative understanding of their respective performances. The fully supervised method, represented by A2Net [33], demonstrates superior performance with precision, recall, F1 score, and IoU metrics, all indicating high accuracy. This underscores the effectiveness of fully supervised learning in scenarios where detailed and accurate labeling is available, as seen in its highest F1 score of 0.9536 and IoU of 0.9113. In contrast, the semi-supervised approach, exemplified by STCRNet (10% labeled) [130], shows a noteworthy IoU of 0.8191, despite not having full label availability. This highlights the efficacy of semi-supervised methods in situations where only a limited amount of labeled data is accessible, leveraging the vast amount of unlabeled data to achieve considerable accuracy. The weakly supervised paradigm, as demonstrated by CS-WSCDNet (Image-level labels) [190], presents a different picture. With an IoU of 0.5729, it reflects the challenges inherent in relying on less detailed, image-level labels, which tend to yield lower precision value. Lastly, the unsupervised method, CDRL [170], shows an IoU of 0.5000, indicating its potential in scenarios where no labeled data is available. Despite the lower accuracy compared to supervised methods, its recall of 0.9300 is notably high, suggesting effectiveness in identifying relevant changes, albeit with less precision. These results collectively illustrate the trade-offs and decision-making considerations when selecting a learning paradigm for CD tasks, depending on the availability of labeled data and the required accuracy level.

Table 2. Comparative accuracy metrics across learning paradigms on the WHU–CD dataset.

Method	Paradigm	Pre.	Rec.	F1	IoU
A2Net [33]	Fully Supervised	0.9430	0.9644	0.9536	0.9113
STCRNet (10% labeled) [116]	Semi-Supervised	-	-	0.9006	0.8191
CS-WSCDNet [190]	Weakly Supervised	0.6457	0.8356	0.7284	0.5729
CDRL [170]	Unsupervised	0.5200	0.9300	-	0.5000

4.3. Pros and Cons of Different Learning Paradigms in CD

Building upon our earlier analysis of the frameworks of various learning paradigms, their specific data requirements, and a comparative assessment of accuracy across these paradigms, this subsection provides a critical evaluation of the various learning paradigms used in CD, shedding light on their advantages and disadvantages, as shown in Table 3. By examining the intrinsic characteristics and operational efficiencies of each paradigm, we aim to present a balanced perspective that can guide researchers and practitioners in selecting the most appropriate method for their specific CD tasks.

Table 3. Advantages and disadvantages across learning paradigms.

Paradigm	Advantages	Disadvantages
Fully Supervised	High accuracy; Reliable performance with well-defined ground truth	Time-consuming and costly data annotation process; Less adaptable to new data scenarios
Semi-Supervised	Utilizes both labeled and unlabeled data; Balances performance with data availability	Performance depending on the quality and amount of label; Requires careful tuning; Less effective when label is not representative
Weakly Supervised	Reduces annotation burden with coarse labels; Suitable for rapid response	Limited performance; Struggles with complex scenarios; Dependent on the quality and relevance of weak labels
Unsupervised	No need for labeled data; Suitable for exploratory and large-scale monitoring	Lower performance; Challenging objective evaluation

4.4. Application Scenarios for Different Learning Paradigms

In this subsection, we delve into the practical deployment of different learning paradigms in the realm of CD, highlighting their optimally suited application contexts. It is imperative to recognize that each learning approach, while exhibiting a particular affinity for certain applications, possesses the inherent flexibility to be adapted to a multitude of scenarios. This subsection, therefore, focuses on elucidating the most congruent application scenarios for each learning paradigm, based on their intrinsic characteristics and efficacy in addressing the unique challenges posed by these contexts.

- Fully Supervised Learning: This approach is most apt for the detailed monitoring of urban expansion and land use changes, such as tracking the growth of urban buildings or the development of roadways. These scenarios often demand highly accurate CD, as they directly impact urban planning and management. Moreover, in these contexts, there are usually sufficient resources available to acquire a large amount of precise ground truth data.
- Semi-Supervised Learning: This is suitable for the monitoring of natural resources, such as assessing deforestation or degradation. Given the vast coverage of forest areas, often only a portion of these regions may have detailed annotated data, with the majority remaining unlabeled. In such cases, the limited annotated data, in conjunction with extensive unlabeled data, can be utilized to monitor the health of forests over large areas, thus efficiently evaluating environmental impacts.
- Weakly Supervised Learning: This paradigm is ideal for rapid disaster response, such as quick assessment of changes following floods or fire disasters. In these instances, rapidly acquiring a general understanding of the disaster-affected areas through limited and coarse annotated data is of paramount importance.
- Unsupervised Learning: This method is suitable for monitoring global environmental changes, such as glacier retreat or desertification. The long-term nature of these changes often makes it challenging to obtain a large quantity of precise annotated data.

5. Opportunities and Challenges for DL-based CD

While DL technology has made significant progress in the field of CD, its rapid evolution has introduced new challenges and opportunities, urgently calling for further research and innovation. This section focuses on these emerging aspects, including the continued development of incompletely supervised CD in scenarios with scarce data, the potential applications of self-supervised learning in RS image processing, the exploration of Foundation Models' adaptability in CD tasks, and the challenges of multimodal CD in integrating heterogeneous data sources. These points not only highlight key issues currently awaiting solutions but also suggest possible future research directions, charting a course for the evolution of DL in the field of CD.

5.1. Incomplete Supervised CD

In Section 3, we comprehensively reviewed the methods of incomplete supervision in CD, with a particular focus on semi-supervised and weakly supervised methods. These methods have shown significant potential both in theoretical research and practical applications, especially in scenarios with limited labeled data or coarse-grained labels. By effectively leveraging unlabeled data or coarse labels, these methods offer new perspectives for addressing CD tasks, reducing reliance on costly fine-grained labeled data. Consequently, further research and development in incomplete supervision techniques for CD represent a crucial trend and opportunity in the field. However, these methods are still in their nascent stages and face several challenges:

- Model Performance: In CD tasks, the performance of models is crucial, directly impacting their practical efficacy. Weakly supervised methods, which rely on vague or incomplete labels (image-level, bounding box, scribble-level), may struggle with recognition in complex scenarios. Additionally, sensitivity to subtle changes poses a challenge, particularly in applications sensitive to fine-grained variations.
- Uncertainty Management: Incompleteness, imprecision, or vagueness in annotations can lead to uncertainty in weakly supervised learning predictions, affecting reliability and trust in practical applications. Managing this uncertainty—accurately representing and quantifying it in predictions—is key to enhancing the effectiveness of weakly supervised models. Current strategies include integrating Bayesian methods and confidence assessments into the training process to explicitly account for uncertainties and achieve more reliable model outcomes.
- Severe Sample Imbalance: Existing semi-supervised CD studies typically select 5% to 40% of samples from supervised datasets to simulate a semi-supervised scenario. In real-world contexts, this ratio is often more skewed, with labeled samples possibly comprising less than 1% of a much larger total sample size. Thus, developing robust semi-supervised learning algorithms that utilize a minimal amount of labeled data and learn from a large pool of unlabeled data is a significant challenge.

The rapid advancement in DL within the field of computer vision has brought about many new technologies and methods, offering potential solutions to these challenges:

- Existing incomplete supervision methods in CD primarily utilize CNN as the backbone. The swift evolution of DL has introduced more powerful and flexible network architectures capable of handling complex and high-dimensional data more effectively, thereby enhancing the accuracy and efficiency of CD. For instance, the ViT has become a popular model in image processing, recently applied to supervised CD with satisfying results. Exploring its application in incomplete supervision CD is one of the most promising future research directions.
- Emerging learning paradigms, like self-supervised learning, not only provide effective solutions for handling severely imbalanced datasets but also offer new approaches for rapid model adaptation and generalization. Self-supervised learning will be further discussed in Section 5.2.

- Additionally, the emergence of Visual Foundation Models opens new possibilities. Their exceptional transferability offers novel tools and innovative potential for incomplete supervision in CD, which will be further discussed in Section 5.3.

5.2. Self-Supervised Learning

Self-supervised Learning leverages the intrinsic structure of unlabeled data as a learning signal, learning effective feature representations from the data itself through pretext tasks. It has garnered immense attention in DL and achieved remarkable success in computer vision. Numerous self-supervised methods, like MoCo [191], BYOL [192], SwAV [193], SimCLR [194], MAE [195], and DINO [196], have been extensively applied in image classification, object detection, image reconstruction, and image semantic segmentation, demonstrating performance comparable to traditional supervised training methods with minimal fine-tuning on a small set of samples. However, the potential of self-supervised in RS image CD is yet to be fully exploited. Given the significant strides made in CD due to DL, yet with data annotation remaining a major challenge, the self-supervised approach is seen as a promising in CD research.

Existing studies have attempted to apply the self-supervised concept to CD, as seen in references [197–200], focusing mainly on medium-resolution imagery and combining self-supervised with transfer learning or traditional CD techniques, operating in an unsupervised manner without relying on labeled samples. While these methods demonstrate the potential application of self-supervised methods in CD tasks, their performance lags behind fully supervised methods due to their unsupervised nature. On the other hand, self-supervised methods for high-resolution image CD have been proposed, as in references [201–204], but they still rely on a substantial amount of supervised data during fine-tuning, not fully addressing the challenges of data annotation. To date, only reference [205] has experimented with fine-tuning self-supervised models on a minimal dataset (1%), but there remains significant room for performance improvement.

A key future research direction for CD is exploring self-supervised methods under conditions of few-shot and even one-shot learning. This direction is crucial for understanding and enhancing self-supervised applications in CD and also offers a new perspective to address the challenge of scarce labeled data. Several challenges might emerge in this process: first, ensuring that features extracted by self-supervised learning from unlabeled samples are sufficiently representative; second, avoiding overfitting in few-shot or one-shot learning scenarios; and third, optimizing and adjusting self-supervised strategies for specific CD tasks. These challenges boil down to a core question: how to maintain the generalizability of self-supervised models while adapting quickly to specific CD tasks with minimal or single sample learning. Addressing these challenges necessitates in-depth exploration of innovative self-supervised methods and how they can be effectively integrated with CD tasks.

5.3. Visual Foundation Models

The concept of "Foundation Models [206]," proposed by The Stanford Institute for Human-Centered Artificial Intelligence's (HAI) Center for Research on Foundation Models (CRFM) in August 2021, is defined as "models trained on broad data (generally using large-scale self-supervised learning) that can be adapted (e.g., via fine-tuning) to a wide range of downstream tasks." This definition highlights the Foundation Models' characteristics of leveraging extensive data for pre-training and their wide applicability across various scenarios. Initially achieving breakthrough success in natural language processing, particularly through the development of Large Language Models (LLMs), like GPT series [207], PaLM [208], T5 [209], LLaMa [210], and ERNIE [211], these models, with their DL of language semantics and syntax from massive textual data, can perform a wide range of complex linguistic tasks, including text generation, translation, sentiment analysis, and question-answering systems, marking a new era in AI research and applications. Exploration of Foundation Models has also been conducted in the field of vision. Models

like CLIP [212] and ALIGN [213], trained on a vast number of image–text pairs, demonstrate the capability to understand and link image content with textual descriptions. They map images and text into a shared representational space, enabling potent cross-modal capabilities. Derived models, such as Florence [214], RegionCLIP [215], CLIP2Video [216], and CLIP-ViL [217], along with integration with modules like DALL-E [218], have shown adaptability to various computer vision tasks, like image classification, object detection, visual question answering, and image generation.

In RS image processing, a related task in computer vision, the application of Visual Foundation Models has shown significant potential and broad prospects [219–222]. As CD in RS imagery is essentially a semantic segmentation task, models specifically developed for segmentation, like CLIPSeg [223], SegGPT [224], Segment Anything Model (SAM) [225], and SEEM [226], are closer to CD tasks and exhibit immense potential in precisely identifying and tracking terrestrial changes. Recent research [227–230] efforts have begun to explore the applicability of these Visual Foundation Models in the field of RS image segmentation, offering innovative perspectives and methodologies. For more specific CD tasks, Ding et al. [231] integrated FastSAM [232] as an encoder in a supervised learning model for feature extraction in RS imagery, exploring its potential advantages in semi-supervised CD tasks. Wang et al. [190] combined the localization capability of CAM with SAM's zero-shot segmentation ability, establishing a weakly supervised CD framework that achieves precise pixel-level CD on VHR RS images using only image-level labels. These explorations demonstrate the opportunities Visual Foundation Models bring to RS image processing, accelerating data processing speed, enhancing task accuracy, and reducing reliance on large-scale annotated datasets, offering new directions for future developments in CD and related tasks in RS image processing. Nevertheless, applying Visual Foundation Models in RS image processing still faces several challenges that need to be overcome through continuous technological innovation and research depth:

- In incomplete supervision CD scenarios, Visual Foundation Models can serve as a powerful auxiliary tool. Researchers can generate high-quality pseudo-labels using Visual Foundation Models combined with appropriate prompts, reducing reliance on extensive accuracy annotations. However, the potential of Visual Foundation Models extends beyond this; developing effective learning algorithms to leverage Foundation Models' advantages in incomplete supervision and integrating them more directly into the main process of CD are crucial areas for further exploration.
- Existing research shows that pre-training datasets for Visual Foundation Models often lack images specific to certain domains, like RS imagery. Further exploration into developing specialized Foundation Models using large-scale RS datasets could enable models to capture the unique features of remote sensing imagery more accurately, facilitating zero-shot transfer to related tasks. However, processing and analyzing large-scale RS datasets require immense computational resources.
- In scenarios with limited computational resources, fine-tuning Visual Foundation Models through open interfaces is a practical solution. Employing a partial weight-locking strategy allows researchers to update the model for specific RS image-related tasks selectively. This method not only conserves computational resources but also ensures the model's ability to adapt quickly to new tasks. Developing more effective fine-tuning strategies to maintain the model's generalizability and ensuring its continuous update and maintenance remain significant challenges.

5.4. Multimodal CD

Traditionally, CD has relied on single-source data, primarily optical imagery. However, with technological advancements and growing application demands, RS has seen significant improvements in data acquisition and sensor technologies. This progress has yielded a wealth of heterogeneous and complex earth observation data for CD, such as optical, SAR, LiDAR, thermal infrared, and satellite video data. Additionally, various data sources. like GIS and ground survey data. provide rich information about geographical environments,

topographical features, and land use, offering multi-dimensional references and support for CD. The fusion of multimodal data not only overcomes the limitations of single data sources, such as temporal and spatial coverage or occlusions, but also leverages the strengths of each modality, presenting immense potential for a more comprehensive and detailed understanding of surface changes.

Multimodal data can enrich the representation of the Earth's surface, detecting changes that may be challenging to discern in single datasets. For instance, SAR data excel in all-weather conditions, penetrating clouds to complement optical imagery; LiDAR data provides detailed information about terrain and surface elevation, enhancing CD accuracy in diverse terrain regions. There are already some DL-based CD studies leveraging multimodal data [157,233–236]. For instance, Li et al. [237] proposed a GAN and CNN-based network for optical and SAR image CD, using GANs to align optical and SAR images into the same feature space, followed by supervised CNN for CD. Zhang et al. [238] applied domain adaptation constraints to align optical and SAR images at a deep feature level within the same feature space, unifying deep heterogeneous feature alignment and CD tasks in an end-to-end framework, thereby avoiding unintended noise introduction.

However, current DL-based multimodal CD approaches mostly focus on bi-modal imagery. Effectively utilizing a broader range of multimodal imagery, as well as integrating data beyond imagery such as GIS and ground survey data, remain challenges in multimodal CD tasks. Furthermore, better aligning multimodal data using DL techniques, such as registering non-homogenous imagery or recognizing relationships among elements across different modalities, remains a direction for future research. Equally important is the design of DL networks that can effectively merge the complementary aspects of multimodal data and eliminate redundancy, thereby achieving improved feature representation. Utilizing these enhanced multimodal feature representations to execute tasks such as CD is also crucial.

6. Conclusions

This review, taking diverse learning paradigms as perspectives, reports on and analyzes the latest methods and challenges in the field of DL-based CD. Firstly, it introduces the fundamental network architectures utilized in DL for CD, laying a solid foundation for understanding core technologies in the field. Subsequently, the review comprehensively summarizes and analyzes DL-based CD methods under different learning paradigms, meticulously sorting out their commonalities and characteristics, and summarizing commonly used frameworks, thus providing essential references for designing CD methods. Following this, the review highlights a range of publicly available datasets for CD, underscoring the importance of diverse data sources in advancing research. Finally, the review explores forward-looking prospects and challenges in CD, focusing on the roles of incomplete supervision, self-supervised learning, visual Foundation Models, and multimodal CD. These insights pave the way for future research directions, emphasizing the need for continuous innovation and adaptation in the rapidly evolving field of DL-based CD. Through this review, researchers gain a comprehensive understanding of current methods, challenges, and future trajectories in CD, benefiting both newcomers and seasoned professionals in the field.

Author Contributions: All of the authors made significant contributions to the manuscript. The review was created and written by L.W. guided by the oversight of M.Z. and W.S., and X.G. helped with the review of the related literature. All authors discussed the basic structure of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Otto Poon Charitable Foundation Smart Cities Research Institute, the Hong Kong Polytechnic University (Work Program: CD03); The Hong Kong Polytechnic University (1-ZVN6; ZVU1; U-ZECR).

Data Availability Statement: Not applicable.

Acknowledgments: The authors express their sincere gratitude to the academic editors and reviewers for their valuable comments and constructive suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Singh, A. Review Article Digital Change Detection Techniques Using Remotely-Sensed Data. *Int. J. Remote Sens.* **1989**, *10*, 989–1003. [CrossRef]
2. Zhu, Z.; Woodcock, C.E. Continuous Change Detection and Classification of Land Cover Using All Available Landsat Data. *Remote Sens. Environ.* **2014**, *144*, 152–171. [CrossRef]
3. Wang, L.; Zhang, M.; Shen, X.; Shi, W. Landslide Mapping Using Multilevel-Feature-Enhancement Change Detection Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3599–3610. [CrossRef]
4. Luo, H.; Liu, C.; Wu, C.; Guo, X. Urban Change Detection Based on Dempster–Shafer Theory for Multitemporal Very High-Resolution Imagery. *Remote Sens.* **2018**, *10*, 980. [CrossRef]
5. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sens.* **2020**, *12*, 1688. [CrossRef]
6. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sens.* **2022**, *14*, 871. [CrossRef]
7. Khelifi, L.; Mignotte, M. Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. *IEEE Access* **2020**, *8*, 126385–126400. [CrossRef]
8. Deng, J.S.; Wang, K.; Deng, Y.H.; Qi, G.J. PCA-Based Land-Use Change Detection and Analysis Using Multitemporal and Multisensor Satellite Data. *Int. J. Remote Sens.* **2008**, *29*, 4823–4838. [CrossRef]
9. Chen, J.; Gong, P.; He, C.; Pu, R.; Shi, P. Land-Use/Land-Cover Change Detection Using Improved Change-Vector Analysis. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 369–379. [CrossRef]
10. Bruzzone, L.; Prieto, D.F. Automatic Analysis of the Difference Image for Unsupervised Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 1171–1182. [CrossRef]
11. Celik, T. Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
12. National Academies of Sciences, Engineering, and Medicine. *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*; National Academies Press: Washington, DC, USA, 2019.
13. Zhao, Q.; Yu, L.; Du, Z.; Peng, D.; Hao, P.; Zhang, Y.; Gong, P. An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends. *Remote Sens.* **2022**, *14*, 1863. [CrossRef]
14. Ma, Y.; Wu, H.; Wang, L.; Huang, B.; Ranjan, R.; Zomaya, A.; Jie, W. Remote Sensing Big Data Computing: Challenges and Opportunities. *Future Gener. Comput. Syst.* **2015**, *51*, 47–60. [CrossRef]
15. Chi, M.; Plaza, A.; Benediktsson, J.A.; Sun, Z.; Shen, J.; Zhu, Y. Big Data for Remote Sensing: Challenges and Opportunities. *Proc. IEEE* **2016**, *104*, 2207–2219. [CrossRef]
16. Asokan, A.; Anitha, J. Change Detection Techniques for Remote Sensing Applications: A Survey. *Earth Sci. Inform.* **2019**, *12*, 143–160. [CrossRef]
17. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
18. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
20. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
23. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
24. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
26. Girshick, R. Fast R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

27. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
28. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
29. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
30. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
31. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 3085870. [CrossRef]
32. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3056416. [CrossRef]
33. Li, Z.; Tang, C.; Liu, X.; Zhang, W.; Dou, J.; Wang, L.; Zomaya, A.Y. Lightweight Remote Sensing Change Detection with Progressive Feature Aggregation and Supervised Attention. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3241436. [CrossRef]
34. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
35. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
36. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
37. Mikolov, T.; Kombrink, S.; Burget, L.; Černocký, J.; Khudanpur, S. Extensions of Recurrent Neural Network Language Model. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 5528–5531.
38. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
39. Rangapuram, S.S.; Seeger, M.; Gasthaus, J.; Stella, L.; Wang, Y.; Januschowski, T. Deep State Space Models for Time Series Forecasting. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 7796–7805.
40. Lyu, H.; Lu, H.; Mou, L.; Li, W.; Wright, J.; Li, X.; Li, X.; Zhu, X.X.; Wang, J.; Yu, L.; et al. Long-Term Annual Mapping of Four Cities on Different Continents by Applying a Deep Information Learning Method to Landsat Data. *Remote Sens.* **2018**, *10*, 471. [CrossRef]
41. Lyu, H.; Lu, H.; Mou, L. Learning a Transferable Change Rule from a Recurrent Neural Network for Land Cover Change Detection. *Remote Sens.* **2016**, *8*, 506. [CrossRef]
42. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
43. Kramer, M.A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE J.* **1991**, *37*, 233–243. [CrossRef]
44. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]
45. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
46. Ng, A. Sparse Autoencoder. *CS294A Lect. Notes* **2011**, *72*, 1–19.
47. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and Composing Robust Features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
49. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
50. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation through Attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10347–10357.
51. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
52. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.

53. Dong, L.; Xu, S.; Xu, B. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5884–5888.
54. Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-Augmented Transformer for Speech Recognition. *arXiv* **2020**, arXiv:2005.08100.
55. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-Xl: Attentive Language Models beyond a Fixed-Length Context. *arXiv* **2019**, arXiv:1901.02860.
56. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
57. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable Detr: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
58. Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; Chen, L.-C. Max-Deeplab: End-to-End Panoptic Segmentation with Mask Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 5463–5474.
59. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
60. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
61. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3160007. [CrossRef]
62. Song, X.; Hua, Z.; Li, J. Remote Sensing Image Change Detection Transformer Network Based on Dual-Feature Mixed Attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3209972. [CrossRef]
63. Liu, M.; Shi, Q.; Chai, Z.; Li, J. PA-Former: Learning Prior-Aware Transformer for Remote Sensing Building Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3200396. [CrossRef]
64. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A Nested u-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: DLMIA 2018, Granada, Spain, 20 September 2018; pp. 3–11.
65. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
66. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
67. Daudt, R.C.; Saux, B.L.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
68. Peng, D.; Zhang, Y.; Guan, H. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
69. Zhang, X.; Yue, Y.; Gao, W.; Yun, S.; Su, Q.; Yin, H.; Zhang, Y. DifUnet++: A Satellite Images Change Detection Network Based on UNet++ and Differential Pyramid. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
70. Lei, T.; Zhang, Q.; Xue, D.; Chen, T.; Meng, H.; Nandi, A.K. End-to-End Change Detection Using a Symmetric Fully Convolutional Network for Landslide Mapping. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3027–3031.
71. Liu, J.; Gong, M.; Qin, K.; Zhang, P. A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *29*, 545–559. [CrossRef] [PubMed]
72. Zhan, T.; Gong, M.; Liu, J.; Zhang, P. Iterative Feature Mapping Network for Detecting Multiple Changes in Multi-Source Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 38–51. [CrossRef]
73. Johnson, R.D.; Kasischke, E.S. Change Vector Analysis: A Technique for the Multispectral Monitoring of Land Cover and Condition. *Int. J. Remote Sens.* **1998**, *19*, 411–426. [CrossRef]
74. Vorovencii, I.; Nir, R. *A Change Vector Analysis Technique for Monitoring Land Cover Changes in Copsa Mica, Romania, in the Period 1985–2011*; Transilvania University of Brasov, Faculty of Silviculture: Brasov, Romania, 2011.
75. Abdi, H.; Williams, L.J. Principal Component Analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
76. Wang, L.; Li, Y.; Zhang, M.; Shen, X.; Peng, W.; Shi, W. MSFF-CDNet: A Multiscale Feature Fusion Change Detection Network for Bi-Temporal High-Resolution Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3305623. [CrossRef]
77. Chen, T.; Lu, Z.; Yang, Y.; Zhang, Y.; Du, B.; Plaza, A. A Siamese Network Based U-Net for Change Detection in High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2357–2369. [CrossRef]
78. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A Deeply Supervised Image Fusion Network for Change Detection in High Resolution Bi-Temporal Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
79. Zhu, Q.; Guo, X.; Deng, W.; Shi, S.; Guan, Q.; Zhong, Y.; Zhang, L.; Li, D. Land-Use/Land-Cover Change Detection Based on a Siamese Global Learning Framework for High Spatial Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [CrossRef]

80. Ding, L.; Guo, H.; Liu, S.; Mou, L.; Zhang, J.; Bruzzone, L. Bi-Temporal Semantic Reasoning for the Semantic Change Detection in HR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3154390. [CrossRef]
81. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change Detection on Remote Sensing Images Using Dual-Branch Multilevel Intertemporal Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3241257. [CrossRef]
82. Lei, T.; Geng, X.; Ning, H.; Lv, Z.; Gong, M.; Jin, Y.; Nandi, A.K. Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3261273. [CrossRef]
83. Xing, Y.; Jiang, J.; Xiang, J.; Yan, E.; Song, Y.; Mo, D. LightCDNet: Lightweight Change Detection Network Based on VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3304309. [CrossRef]
84. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
85. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
86. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
87. Almahairi, A.; Ballas, N.; Cooijmans, T.; Zheng, Y.; Larochelle, H.; Courville, A. Dynamic Capacity Networks. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 2549–2558.
88. Jaderberg, M.; Simonyan, K.; Zisserman, A. Others Spatial Transformer Networks. *Adv. Neural Inf. Process Syst.* **2015**, *28*, 2017–2025.
89. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. Bam: Bottleneck Attention Module. *arXiv* **2018**, arXiv:1807.06514.
90. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
91. Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
92. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
93. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. *Adv. Neural Inf. Process Syst.* **2021**, *34*, 9355–9366.
94. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going Deeper with Image Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 32–42.
95. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *Adv. Neural Inf. Process Syst.* **2021**, *34*, 15908–15919.
96. Zheng, Z.; Zhong, Y.; Tian, S.; Ma, A.; Zhang, L. ChangeMask: Deep Multi-Task Encoder-Transformer-Decoder Architecture for Semantic Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 228–239. [CrossRef]
97. Zhang, X.; Cheng, S.; Wang, L.; Li, H. Asymmetric Cross-Attention Hierarchical Network Based on CNN and Transformer for Bitemporal Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3245674. [CrossRef]
98. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 3095166. [CrossRef]
99. Liu, M.; Shi, Q.; Li, J.; Chai, Z. Learning Token-Aligned Representations with Multimodel Transformers for Different-Resolution Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3200684. [CrossRef]
100. Song, X.; Hua, Z.; Li, J. PSTNet: Progressive Sampling Transformer Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8442–8455. [CrossRef]
101. Yan, T.; Wan, Z.; Zhang, P. Fully Transformer Network for Change Detection of Remote Sensing Images. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022; pp. 1691–1708.
102. Li, Q.; Zhong, R.; Du, X.; Du, Y. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]
103. Jiang, B.; Wang, Z.; Wang, X.; Zhang, Z.; Chen, L.; Wang, X.; Luo, B. VcT: Visual Change Transformer for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3327139. [CrossRef]
104. Yan, T.; Wan, Z.; Zhang, P.; Cheng, G.; Lu, H. TransY-Net: Learning Fully Transformer Networks for Change Detection of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3327253. [CrossRef]
105. Wang, Y.; Hong, D.; Sha, J.; Gao, L.; Liu, L.; Zhang, Y.; Rong, X. Spectral–Spatial–Temporal Transformers for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3203075. [CrossRef]
106. Li, W.; Xue, L.; Wang, X.; Li, G. ConvTransNet: A CNN–Transformer Network for Change Detection with Multiscale Global–Local Representations. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3272694. [CrossRef]
107. Xue, D.; Lei, T.; Yang, S.; Lv, Z.; Liu, T.; Jin, Y.; Nandi, A.K. Triple Change Detection Network via Joint Multi-Frequency and Full-Scale Swin-Transformer for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4408415. [CrossRef]

108. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210.
109. Song, F.; Zhang, S.; Lei, T.; Song, Y.; Peng, Z. MSTDSNet-CD: Multiscale Swin Transformer and Deeply Supervised Network for Change Detection of the Fast-Growing Urban Regions. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3165885. [CrossRef]
110. Mao, Z.; Tong, X.; Luo, Z.; Zhang, H. MFATNet: Multi-Scale Feature Aggregation via Transformer for Remote Sensing Image Change Detection. *Remote Sens.* **2022**, *14*, 5379. [CrossRef]
111. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
112. Jiang, F.; Gong, M.; Zhan, T.; Fan, X. A Semisupervised GAN-Based Multiple Change Detection Framework in Multi-Spectral Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1223–1227. [CrossRef]
113. Yang, S.; Hou, S.; Zhang, Y.; Wang, H.; Ma, X. Change Detection of High-Resolution Remote Sensing Image Based on Semi-Supervised Segmentation and Adversarial Learning. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 1055–1058.
114. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; DIng, H.; Huang, X. SemiCDNet: A Semisupervised Convolutional Neural Network for Change Detection in High Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5891–5906. [CrossRef]
115. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.-L. Fixmatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process Syst.* **2020**, *33*, 596–608.
116. Wang, L.; Zhang, M.; Shi, W. STCRNet A Semi-Supervised Network Based on Self-Training and Consistency Regularization for Change Detection in VHR Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2272–2282. [CrossRef]
117. Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; Gao, Y. St++: Make Self-Training Work Better for Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4268–4277.
118. Wang, J.X.; Li, T.; Chen, S.B.; Tang, J.; Luo, B.; Wilson, R.C. Reliable Contrastive Learning for Semi-Supervised Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3228016. [CrossRef]
119. Sun, C.; Wu, J.; Chen, H.; Du, C. SemiSANet: A Semi-Supervised High-Resolution Remote Sensing Image Change Detection Model Using Siamese Networks with Graph Attention. *Remote Sens.* **2022**, *14*, 2801. [CrossRef]
120. Chen, Z.; Zhang, R.; Zhang, G.; Ma, Z.; Lei, T. Digging into Pseudo Label: A Low-Budget Approach for Semi-Supervised Semantic Segmentation. *IEEE Access* **2020**, *8*, 41830–41837. [CrossRef]
121. Zhu, Y.; Zhang, Z.; Wu, C.; Zhang, Z.; He, T.; Zhang, H.; Manmatha, R.; Li, M.; Smola, A.J. Improving Semantic Segmentation via Efficient Self-Training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *46*, 1589–1602. [CrossRef] [PubMed]
122. He, R.; Yang, J.; Qi, X. Re-Distributing Biased Pseudo Labels for Semi-Supervised Semantic Segmentation: A Baseline Investigation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6930–6940.
123. Yuan, J.; Liu, Y.; Shen, C.; Wang, Z.; Li, H. A Simple Baseline for Semi-Supervised Semantic Segmentation with Strong Data Augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8229–8238.
124. Sun, C.; Chen, H.; Du, C.; Jing, N. SemiBuildingChange: A Semi-Supervised High-Resolution Remote Sensing Image Building Change Detection Method with a Pseudo Bi-Temporal Data Generator. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5622319. [CrossRef]
125. Zhang, X.; Huang, X.; Li, J. Joint Self-Training and Rebalanced Consistency Learning for Semi-Supervised Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3314452. [CrossRef]
126. Bandara, W.G.C.; Patel, V.M. Revisiting Consistency Regularization for Semi-Supervised Change Detection in Remote Sensing Images. *arXiv* **2022**, arXiv:2204.08454.
127. Ouali, Y.; Hudelot, C.; Tami, M. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12674–12684.
128. Shu, Q.; Pan, J.; Zhang, Z.; Wang, M. MTCNet: Multitask Consistency Network with Single Temporal Supervision for Semi-Supervised Building Change Detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103110. [CrossRef]
129. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 2613–2622.
130. Yang, L.; Qi, L.; Feng, L.; Zhang, W.; Shi, Y. Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7236–7246.
131. Shen, W.; Peng, Z.; Wang, X.; Wang, H.; Cen, J.; Jiang, D.; Xie, L.; Yang, X.; Tian, Q. A Survey on Label-Efficient Deep Image Segmentation: Bridging the Gap Between Weak Supervision and Dense Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9284–9305. [CrossRef]
132. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.

133. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
134. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-Cam++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
135. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 24–25.
136. Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; Wei, Y. Layercam: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Trans. Image Process.* **2021**, *30*, 5875–5888. [CrossRef]
137. Muhammad, M.B.; Yeasin, M. Eigen-Cam: Class Activation Map Using Principal Components. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7.
138. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
139. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1967; Volume 1, pp. 281–297.
140. Lafferty, J.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
141. Kalita, I.; Karatsiolis, S.; Kamilaris, A. Land Use Change Detection Using Deep Siamese Neural Networks and Weakly Supervised Learning. In Proceedings of the Computer Analysis of Images and Patterns: 19th International Conference, CAIP 2021, Virtual Event, 28–30 September 2021; pp. 24–35.
142. Jiang, X.; Tang, H. Dense High-Resolution Siamese Network for Weakly-Supervised Change Detection. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; pp. 547–552.
143. Andermatt, P.; Timofte, R. A Weakly Supervised Convolutional Network for Change Segmentation and Classification. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
144. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting Dilated Convolution: A Simple Approach for Weakly-and Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7268–7277.
145. Zhang, F.; Gu, C.; Zhang, C.; Dai, Y. Complementary Patch for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7242–7251.
146. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-Pixel Relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2204–2213. [CrossRef]
147. Lee, J.; Kim, E.; Mok, J.; Yoon, S. Anti-Adversarially Manipulated Attributions for Weakly Supervised Semantic Segmentation and Object Localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *46*, 1618–1634. [CrossRef]
148. Zhang, X.; Peng, Z.; Zhu, P.; Zhang, T.; Li, C.; Zhou, H.; Jiao, L. *Adaptive Affinity Loss and Erroneous Pseudo-Label Refinement for Weakly Supervised Semantic Segmentation*; Association for Computing Machinery: New York, NY, USA, 2021; Volume 1, ISBN 978-1-45038-651-7.
149. Ahn, J.; Kwak, S. Learning Pixel-Level Semantic Affinity with Image-Level Supervision for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4981–4990. [CrossRef]
150. Ru, L.; Zhan, Y.; Yu, B.; Du, B. Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16846–16855.
151. Huang, R.; Wang, R.; Guo, Q.; Wei, J.; Zhang, Y.; Fan, W.; Liu, Y. Background-Mixed Augmentation for Weakly Supervised Change Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 7919–7927.
152. Zhao, W.; Shang, C.; Lu, H. Self-Generated Defocus Blur Detection via Dual Adversarial Discriminators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 6933–6942.
153. Wu, C.; Du, B.; Zhang, L. Fully Convolutional Change Detection Framework with Generative Adversarial Network for Unsupervised, Weakly Supervised and Regional Supervised Change Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9774–9788. [CrossRef] [PubMed]
154. Zhang, M.; Shi, W.; Chen, S.; Zhan, Z.; Shi, Z. Deep Multiple Instance Learning for Landslide Mapping. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1711–1715. [CrossRef]
155. Lv, N.; Chen, C.; Qiu, T.; Sangaiah, A.K. Deep Learning and Superpixel Feature Extraction Based on Contractive Autoencoder for Change Detection in SAR Images. *IEEE Trans. Ind. Inform.* **2018**, *14*, 5530–5538. [CrossRef]
156. Kosiorek, A.; Sabour, S.; Teh, Y.W.; Hinton, G.E. Stacked Capsule Autoencoders. *Adv. Neural Inf. Process Syst.* **2019**, *32*, 15486–15496.

157. Luppino, L.T.; Hansen, M.A.; Kampffmeyer, M.; Bianchi, F.M.; Moser, G.; Jenssen, R.; Anfinsen, S.N. Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 60–72. [CrossRef] [PubMed]
158. Bergamasco, L.; Saha, S.; Bovolo, F.; Bruzzone, L. Unsupervised Change-Detection Based on Convolutional-Autoencoder Feature Extraction. In Proceedings of the Image and Signal Processing for Remote Sensing XXV, Strasbourg, France, 9–11 September 2019; Volume 11155, pp. 325–332.
159. Masci, J.; Meier, U.; Cireşan, D.; Schmidhuber, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, 14–17 June 2011; pp. 52–59.
160. Saha, S.; Bovolo, F.; Bruzzone, L. Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3677–3693. [CrossRef]
161. Wu, C.; Chen, H.; Du, B.; Zhang, L. Unsupervised Change Detection in Multitemporal VHR Images Based on Deep Kernel PCA Convolutional Mapping Network. *IEEE Trans. Cybern.* **2021**, *52*, 12084–12098. [CrossRef] [PubMed]
162. Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299–1319. [CrossRef]
163. Du, B.; Ru, L.; Wu, C.; Zhang, L. Unsupervised Deep Slow Feature Analysis for Change Detection in Multi-Temporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9976–9992. [CrossRef]
164. Gong, M.; Yang, H.; Zhang, P. Feature Learning and Change Feature Classification Based on Deep Learning for Ternary Change Detection in SAR Images. *ISPRS J. Photogramm. Remote Sens.* **2017**, *129*, 212–225. [CrossRef]
165. Zhang, H.; Gong, M.; Zhang, P.; Su, L.; Shi, J. Feature-Level Change Detection Using Deep Representation and Feature Change Analysis for Multispectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1666–1670. [CrossRef]
166. Gong, M.; Niu, X.; Zhang, P.; Li, Z. Generative Adversarial Networks for Change Detection in Multispectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2310–2314. [CrossRef]
167. Nielsen, A.A. The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi-and Hyperspectral Data. *IEEE Trans. Image Process.* **2007**, *16*, 463–478. [CrossRef] [PubMed]
168. Gong, M.; Yang, Y.; Zhan, T.; Niu, X.; Li, S. A Generative Discriminatory Classified Network for Change Detection in Multispectral Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 321–333. [CrossRef]
169. Liu, Y.-H.; Van Nieuwenburg, E.P.L. Discriminative Cooperative Networks for Detecting Phase Transitions. *Phys. Rev. Lett.* **2018**, *120*, 176401. [CrossRef] [PubMed]
170. Noh, H.; Ju, J.; Seo, M.; Park, J.; Choi, D.-G. Unsupervised Change Detection Based on Image Reconstruction Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1352–1361.
171. Zhang, J.; Shao, Z.; Ding, Q.; Huang, X.; Wang, Y.; Zhou, X.; Li, D. AERNet: An Attention-Guided Edge Refinement Network and a Dataset for Remote Sensing Building Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3300533. [CrossRef]
172. Zhang, X.; Yu, W.; Pun, M.-O.; Shi, W. Cross-Domain Landslide Mapping from Large-Scale Remote Sensing Images Using Prototype-Guided Domain-Aware Progressive Representation Learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 1–17. [CrossRef]
173. Holail, S.; Saleh, T.; Xiao, X.; Li, D. AFDE-Net: Building Change Detection Using Attention-Based Feature Differential Enhancement for Satellite Imagery. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3283505. [CrossRef]
174. Liao, C.; Hu, H.; Yuan, X.; Li, H.; Liu, C.; Liu, C.; Fu, G.; Ding, Y.; Zhu, Q. BCE-Net: Reliable Building Footprints Change Extraction Based on Historical Map and up-to-Date Images Using Contrastive Learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *201*, 138–152. [CrossRef]
175. Pang, C.; Wu, J.; Ding, J.; Song, C.; Xia, G.-S. Detecting Building Changes with Off-Nadir Aerial Images. *Sci. China Inf. Sci.* **2023**, *66*, 140306. [CrossRef]
176. Toker, A.; Kondmann, L.; Weber, M.; Eisenberger, M.; Camero, A.; Hu, J.; Hoderlein, A.P.; Şenaras, Ç.; Davis, T.; Cremers, D. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 21158–21167.
177. Liu, M.; Chai, Z.; Deng, H.; Liu, R. A CNN-Transformer Network with Multiscale Context Aggregation for Fine-Grained Cropland Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4297–4306. [CrossRef]
178. Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; Jiang, B. S2Looking: A Satellite Side-Looking Dataset for Building Change Detection. *Remote Sens* **2021**, *13*, 5094. [CrossRef]
179. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Multitask Learning for Large-Scale Semantic Change Detection. *Comput. Vision Image Underst.* **2019**, *187*, 102783. [CrossRef]
180. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
181. Lebedev, M.A.; Vizilter, Y.V.; Vygolov, O.V.; Knyaz, V.A.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 565–571. [CrossRef]
182. Benedek, C.; Sziranyi, T. Change Detection in Optical Aerial Images by a Multilayer Conditional Mixed Markov Model. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3416–3430. [CrossRef]

183. López-Fandiño, J.; Garea, A.S.; Heras, D.B.; Argüello, F. Stacked Autoencoders for Multiclass Change Detection in Hyperspectral Images. In Proceedings of the IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1906–1909.
184. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [CrossRef]
185. Wu, C.; Zhang, L.; Du, B. Kernel Slow Feature Analysis for Scene Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2367–2384. [CrossRef]
186. Daudt, R.C.; Le Saux, B.; Boulch, A.; Gousseau, Y. Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. In Proceedings of the IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2115–2118.
187. Fuentes Reyes, M.; Xie, Y.; Yuan, X.; d'Angelo, P.; Kurz, F.; Cerra, D.; Tian, J. A 2D/3D Multimodal Data Simulation Approach with Applications on Urban Semantic Segmentation, Building Extraction and Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *205*, 74–97. [CrossRef]
188. Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; Shi, Z. Remote Sensing Image Change Captioning with Dual-Branch Transformers: A New Method and a Large Scale Dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3218921. [CrossRef]
189. Li, H.; Zhu, F.; Zheng, X.; Liu, M.; Chen, G. MSCDUNet: A Deep Learning Framework for Built-Up Area Change Detection Integrating Multispectral, SAR, and VHR Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5163–5176. [CrossRef]
190. Wang, L.; Zhang, M.; Shi, W. CS-WSCDNet: Class Activation Mapping and Segment Anything Model-Based Framework for Weakly Supervised Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3330479. [CrossRef]
191. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
192. Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap Your Own Latent—A New Approach to Self-Supervised Learning. *Adv. Neural Inf. Process Syst.* **2020**, *33*, 21271–21284.
193. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Adv. Neural Inf. Process Syst.* **2020**, *33*, 9912–9924.
194. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
195. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
196. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. Dino: Detr with Improved Denoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605.
197. Akiva, P.; Purri, M.; Leotta, M. Self-Supervised Material and Texture Representation Learning for Remote Sensing Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8203–8215.
198. Manas, O.; Lacoste, A.; Giró-i-Nieto, X.; Vazquez, D.; Rodriguez, P. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9414–9423.
199. Chen, Y.; Bruzzone, L. A Self-Supervised Approach to Pixel-Level Change Detection in Bi-Temporal RS Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3203897. [CrossRef]
200. Leenstra, M.; Marcos, D.; Bovolo, F.; Tuia, D. Self-Supervised Pre-Training Enhances Change Detection in Sentinel-2 Imagery. In Proceedings of the Pattern Recognition ICPR International Workshops and Challenges, Virtual Event, 10–15 January 2021; pp. 578–590.
201. Jiang, F.; Gong, M.; Zheng, H.; Liu, T.; Zhang, M.; Liu, J. Self-Supervised Global–Local Contrastive Learning for Fine-Grained Change Detection in VHR Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3238327. [CrossRef]
202. Chen, H.; Li, W.; Chen, S.; Shi, Z. Semantic-Aware Dense Representation Learning for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3203769. [CrossRef]
203. Saha, S.; Ebel, P.; Zhu, X.X. Self-Supervised Multisensor Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3109957. [CrossRef]
204. Chen, Y.; Bruzzone, L. Self-Supervised Change Detection in Multiview Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3089453. [CrossRef]
205. Zhang, Y.; Zhao, Y.; Dong, Y.; Du, B. Self-Supervised Pretraining via Multimodality Images with Transformer for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3271024. [CrossRef]
206. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
207. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process Syst.* **2020**, *33*, 1877–1901.
208. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling Language Modeling with Pathways. *arXiv* **2022**, arXiv:2204.02311.

209. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
210. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
211. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. Ernie: Enhanced Representation through Knowledge Integration. *arXiv* **2019**, arXiv:1904.09223.
212. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
213. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 4904–4916.
214. Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A New Foundation Model for Computer Vision. *arXiv* **2021**, arXiv:2111.11432.
215. Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L.H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. Regionclip: Region-Based Language-Image Pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16793–16803.
216. Fang, H.; Xiong, P.; Xu, L.; Chen, Y. Clip2video: Mastering Video-Text Retrieval via Image Clip. *arXiv* **2021**, arXiv:2106.11097.
217. Shen, S.; Li, L.H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; Keutzer, K. How Much Can Clip Benefit Vision-and-Language Tasks? *arXiv* **2021**, arXiv:2107.06383.
218. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with Clip Latents. *arXiv* **2022**, arXiv:2204.06125.
219. Cha, K.; Seo, J.; Lee, T. A Billion-Scale Foundation Model for Remote Sensing Images. *arXiv* **2023**, arXiv:2304.05215.
220. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Zhou, J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *arXiv* **2023**, arXiv:2306.11029.
221. Zhang, J.; Zhou, Z.; Mai, G.; Mu, L.; Hu, M.; Li, S. Text2Seg: Remote Sensing Image Semantic Segmentation via Text-Guided Visual Foundation Models. *arXiv* **2023**, arXiv:2304.10597.
222. Wen, C.; Hu, Y.; Li, X.; Yuan, Z.; Zhu, X.X. Vision-Language Models in Remote Sensing: Current Progress and Future Trends. *arXiv* **2023**, arXiv:2305.05726.
223. Lüddecke, T.; Ecker, A. Image Segmentation Using Text and Image Prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7086–7096.
224. Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; Huang, T. Seggpt: Segmenting Everything in Context. *arXiv* **2023**, arXiv:2304.03284.
225. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
226. Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; Lee, Y.J. Segment Everything Everywhere All at Once. *arXiv* **2023**, arXiv:2304.06718.
227. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. Rsprompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *arXiv* **2023**, arXiv:2306.16269. [CrossRef]
228. Osco, L.P.; Wu, Q.; de Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.M.; Li, J.; Junior, J.M. The Segment Anything Model (Sam) for Remote Sensing Applications: From Zero to One Shot. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103540. [CrossRef]
229. Ji, W.; Li, J.; Bi, Q.; Li, W.; Cheng, L. Segment Anything Is Not Always Perfect: An Investigation of Sam on Different Real-World Applications. *arXiv* **2023**, arXiv:2304.05750.
230. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, LA, USA, 10–16 December 2023.
231. Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Guo, H. Adapting Segment Anything Model for Change Detection in HR Remote Sensing Images. *arXiv* **2023**, arXiv:2309.01429.
232. Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; Wang, J. Fast Segment Anything. *arXiv* **2023**, arXiv:2306.12156.
233. Chen, H.; Yokoya, N.; Chini, M. Fourier Domain Structural Relationship Analysis for Unsupervised Multimodal Change Detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 99–114. [CrossRef]
234. Hao, F.; Ma, Z.-F.; Tian, H.-P.; Wang, H.; Wu, D. Semi-Supervised Label Propagation for Multi-Source Remote Sensing Image Change Detection. *Comput. Geosci.* **2023**, *170*, 105249. [CrossRef]
235. Chen, H.; Yokoya, N.; Wu, C.; Du, B. Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3229027. [CrossRef]
236. Jin, H.; Mountrakis, G. Fusion of Optical, Radar and Waveform LiDAR Observations for Land Cover Classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 171–190. [CrossRef]

237. Li, X.; Du, Z.; Huang, Y.; Tan, Z. A Deep Translation (GAN) Based Change Detection Network for Optical and SAR Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *179*, 14–34. [CrossRef]
238. Zhang, C.; Feng, Y.; Hu, L.; Tapete, D.; Pan, L.; Liang, Z.; Cigna, F.; Yue, P. A Domain Adaptation Neural Network for Change Detection with Heterogeneous Optical and SAR Remote Sensing Images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *109*, 102769. [CrossRef]