electronics

MDPI

Article

D^{2}Former: Dual-Domain Transformer for Change Detection in VHR Remote Sensing Images

Huanhuan Zheng [1,2], Hui Liu [1,*], Lei Lu [2], Shiyin Li [1] and Jiyan Lin [2]

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; zhenghuan@yulinu.edu.cn (H.Z.); lishiyin@cumt.edu.cn (S.L.)
[2] School of Information Engineering, Yulin University, Yulin 719000, China; lulei@yulinu.edu.cn (L.L.); linjiyan@yulinu.edu.cn (J.L.)
* Correspondence: hui.liu@cumt.edu.cn

Abstract: Computational intelligence technologies have been extensively applied for the interpretation of remote sensing imagery. Recently, the computational-intelligence-based Transformer change detection (CD) approach has attracted increasing attention. However, the current Transformer-based CD method can better capture global features, but there is no good solution for the loss of local detail information. For this reason, introducing semantic and frequency information from the perspective of a dual-domain can be beneficial for improving the representation of detailed features to improve CD performance. To overcome this limitation, a dual-domain Transformer (D^{2}Former) is proposed for CD. Firstly, we adopt a semantic tokenizer to capture the semantic information, which promotes the enrichment and refinement of semantic change information in the Transformer. Secondly, a frequency tokenizer is introduced to acquire the frequency information of the features, which offers the proposed D^{2}Former another aspect and dimension to enhance the ability to detect change information. Therefore, the proposed D^{2}Former employs dual-domain tokenizers to acquire and fuse the feature representation with rich semantic and frequency information, which can refine the features to acquire more fine-grained CD ability. Extensive experiments on three CD benchmark datasets demonstrate that the proposed D^{2}Former obviously outperforms some other existing approaches. The results present the competitive performance of our method on the WHU-CD, LEVIR-CD, and GZ-CD datasets, for which it achieved *F1-Score* metrics of 92.85%, 90.60%, and 87.02%, respectively.

Keywords: computational intelligence; change detection; transformer; frequency domain; remote sensing images

check for updates

Citation: Zheng, H.; Liu, H.; Lu, L.; Li, S.; Lin, J. D^{2}Former: Dual-Domain Transformer for Change Detection in VHR Remote Sensing Images. *Electronics* **2024**, *13*, 2204. https://doi.org/10.3390/electronics13112204

Academic Editor: Silvia Liberata Ullo

Received: 7 May 2024
Revised: 1 June 2024
Accepted: 3 June 2024
Published: 5 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As multi-temporal remotely sensed imagery is continuously obtained [1,2], computational intelligence and remote sensing technologies have been widely used to understand remote sensing imagery [3,4], such as for land cover change detection (CD) [5,6], scene classification [7], object segmentation [8], and point cloud registration [9,10]. To this end, computational-intelligence-based technologies have been one of the hot technologies in the field of remote sensing [4]. In particular, technology based on computational intelligence is one of the hot technologies that CD has focused on in recent years. Yet detecting meaningful differences in bi-temporal remote sensing images remains one of the most difficult challenges in remote sensing image processing [11,12]. CD identifies changes in land cover by analyzing and comparing remote sensing images from different periods for the same geographical location [13]. For this reason, CD has been widely used in urban development planning [14], landslide detection and mapping [15–17], earthquake detection [18], etc.

In the past few decades, CD has continued to develop and has made certain progress. In the early stage, CD was implemented through the following two steps: image difference measurement and binary segmentation. Image difference measurement refers to obtaining

the change intensity by comparing the similarity between bi-temporal remote sensing images, such as image differences [19], change vector analysis [20], other improved difference measurements [21,22], and so on. Binary segmentation divides the image differences into changed and unchanged categories, such as: the Otsu threshold [23], k-means clustering [24], fuzzy C-means clustering [25], etc. Based on this workflow, many techniques have been proposed and utilized in CD. These approaches focus on improving CD performance through three aspects in the above-mentioned workflow: hand-crafted features [26], similarity metrics [27], and appropriate segmentation or classification algorithms [28]. However, these traditional methods usually have limited performance and insufficient generalization. This is because these traditional methods mostly rely on basic spectral–spatial and other hand-crafted feature extraction methods to measure the differences between bi-temporal remotely sensed imagery, which is unstable. After all, the imaging conditions between bi-temporal remote sensing images are different, such as season, weather, and humidity.

In recent years, with the rapid development of artificial intelligence technologies such as computational intelligence and deep learning, CD has made great progress. Many CD methods based on deep learning (DL) have been proposed and applied in practical applications [12]. Different from traditional CD methods, the advantage of these deep-learning-based CD methods is that they can effectively extract deep semantic features of remote sensing images by convolution neural network (CNNs) in a supervised or unsupervised learning manner, which is beneficial for identifying land cover changes. In [29], three basic and popular fully convolutional networks are proposed for CD, i.e., fully convolutional early fusion network (FC-EF), fully convolutional Siamese-concatenation network (FC-Siam-C), and fully convolutional Siamese-difference network (FC-Siam-D). Based on these three frameworks, some studies have further proposed many technologies to improve the performance of CD, such as pyramid structures, attention mechanisms, etc. For instance, Chen et al. developed a spatial–temporal attention-based network (STANet) for CD [30] that utilizes a spatial pyramid module with a self-attention mechanism to capture multi-scale spatial–temporal dependencies, which can be helpful for CD. In [31], Liu et al. proposed a local–global pyramid network with cross-task transfer learning for building CD that composes a local feature pyramid module and a global spatial pyramid module to acquire discriminative spatial and contextual features to enhance the performance of CD. In addition, recently, CD methods have included edge information or frequency information to further explicitly increase the representation ability of the network's detailed information from the perspective of the frequency domain, thereby improving the detection accuracy of CD. For example, Zheng et al. proposed a high-frequency attention Siamese network for building CD that can enhance high-frequency information for building feature extraction by exploiting a high-frequency attention block [32]. Further, an attention-guided edge refinement network was proposed for building CD in [33]; it can augment the network's ability to detect the edges of changed regions. These CNN-based methods have made certain progress in CD.

In the past three years, the Transformer architecture has achieved great successes in the natural language processing and computer vision fields [34,35]. Recently, researchers have begun to focus on Transformer-based networks in remote sensing image CD tasks. Chen et al. proposed a prevalent bi-temporal image transformer (BIT) for CD [36]. BIT can model the contexts within the spatial–temporal domain by a Transformer encoder and decoder for improving CD performance. Additionally, in [37], a novel Transformer-based Siamese network named ChangeFormer was designed for CD. ChangeFormer constructs a Siamese CD framework that combines a hierarchical Transformer encoder with a multi-layer perception decoder to consider multi-scale long-range relationships. These Transformer-based CD methods have achieved better performance than many pure CNNs.

Although current methods have achieved good detection accuracy, there are still some limitations to a certain extent. Transformer-based methods tend to have better global feature representation capabilities, but this may also lead to more missed detection pairs for pixel-level annotation tasks. Moreover, Transformer-based methods are still limited in their

ability to capture detailed information, such as edges. At the same time, Transformer-based methods may be more prone to missed detections when facing small ground targets. To this end, the motivation of this paper lies in the following two aspects. On the one hand, a pure Transformer has better global feature representation capabilities, but its ability to extract local detail information is relatively weak. On the contrary, a CNN has better local feature representation capabilities, but its ability to extract global information is weaker than that of a Transformer. Therefore, if the advantages of the CNN and the Transformer are combined in an effective manner, the feature representation capabilities for remote sensing images can be significantly elevated as a whole and can achieve higher-accuracy CD. On the other hand, CNN-based CD methods based on frequency domain information such as edges have been proven to effectively enhance the detailed information representation ability of the CD network and improve detection performance. However, there is a lack of research on the verification of the role of frequency domain information on Transformer-based CD networks. Therefore, if frequency domain information is introduced into a Transformer-based CD network, it may improve the network's ability to mine detailed features.

With the above motivations, the paper proposes a dual-domain Transformer (D^2Former) for CD. The contributions of this paper are summarized as follows:

- We propose a novel D^2Former for CD that comprehensively considers semantic information and frequency information based on the CNN and Transformer structures. Overall, our D^2Former demonstrates that feature representation capabilities can be further improved by introducing frequency information in the frequency domain, thereby elevating CD accuracy.
- In the proposed D^2Former, we introduce a semantic tokenizer and a frequency tokenizer into the Transformer network to enhance feature extraction ability from both semantic and frequency perspectives.
- The proposed D^2Former achieves competitive performance on three sets of public CD benchmark datasets (WHU-CD, LEVIR-CD, and GZ-CD). Moreover, sufficient ablation experiments have also verified the effectiveness of the semantic tokenizer and the frequency tokenizer introduced by our D^2Former.

The rest of this article is constructed as follows. Section 2 gives a summary of related works. The proposed D^2Former is introduced in detail in Section 3. And Sections 4 and 5 present and analyze the experimental results, respectively. Finally, the conclusion and future works are summarized and elicited in Section 6.

2. Related Works

In recent years, DL-based CD approaches have achieved great development and results. In this section, we roughly summarize these methods into the following two categories: CNN-based and Transformer-based CD approaches.

2.1. CNN-Based Change Detection Methods

In the initial stage, researchers handled CD as a semantic segmentation issue in order to achieve change information extraction [31], such as by using classic semantic segmentation networks: FCN [38], U-Net [39], DeepLab [40], etc. Subsequently, more CNN-based methods were proposed to specifically solve CD tasks. Classically, Daudt et al. constructed three frameworks dedicated to CD: namely, early fusion, Siamese-difference, and Siamese-concatenation in [29]. Since then, the majority of DL-based CD methods have been continued or extended based on these three frameworks. In particular, the proposed structures, such as pyramid and attention mechanisms, have been a boon for remote sensing image interpretation [41,42], and pyramid- and attention-based CNNs have further improved the detection accuracy of CD.

The pyramid structure has been widely used in CD and effectively improves the detection ability of changed targets at different scales. For example, Yang et al. devised a pyramid feature learning network for homogeneous and heterogeneous CD in [43]; it learns hierarchical features and fuses image differences to improve the ability of CD. In [44], a CD

network is promoted based on UNet++, called DifUnet++. DifUnet++ designs a differential pyramid for multi-scale feature extraction of changed regions. Zhang et al. construct a self-structured pyramid network for CD in [45]; it integrates features between different layers and different scales, benefiting multi-scale feature representation. In [46], a pyramid pooling CD approach is developed for synthetic aperture radar (SAR) image CD. Other related pyramid-based CD methods can be referred to in [47,48].

In addition, various attention mechanisms have been introduced into networks for CD. For instance, in [49], an attention-guided network was proposed for building CD; it employ spatial and channel attention to improve feature extraction for buildings. Shi et al. devised a network based on a built-in attention metric module for CD in a deeply supervised manner [50]. In [51], a Siamese U-Net-based CD network was established for CD. The approach employs multi-scale residual block to capture multi-scale spatial and spectral features and utilizes an attention-gate module to emphasize regions of interest while suppressing irrelevant regions. Ma et al. promoted a multi-attention guided feature fusion network for CD in [52]; it was composed of a coordinate attention-based feature enhancement module and a refinement residual block to refine the feature representation. Similar attention-based CD networks can be found in [42,53–55]. In addition to these pyramid- or attention-based methods, more recent methods focus on designing hybrid pyramids and attention mechanisms to optimize CD design solutions further [56–58].

2.2. Transformer-Based Change Detection Methods

Transformer has achieved amazing success in remote sensing image processing and has been widely used in various tasks [59], including scene classification, semantic segmentation, object detection, etc. BIT [36] is one of the representative works for Transformer-based CD networks. Afterwards, Transformer-based CD networks have received continued attention from researchers, and various Transformer-based CD methods have been continuously proposed. For instance, a pure Transformer was proposed for CD in [60]: SwinSUNet is a Siamese U-shaped structure network composed of a series of Swin-Transformer blocks. Different from the locality of the convolution operation, the Transformer-based CD method can effectively capture global information better than a CNN-based approach. In [61], a full Transformer network is developed for CD. This approach adopts a Transformer network and a progressive attention module to model the long-range dependency of remote sensing images for more discriminative global-level features, thereby increasing the performance of CD. Analogously, Liu et al. adopted an end-to-end PA-Former for building CD [62] that acquires the prior structural information by prior-aware features. More Transformer-based CD methods can be found in [63–65]. Although the above Transformer-based CD methods achieve effective detection performance, it is also found that Transformer-based CD methods have certain limitations in extracting local detail information. In order to alleviate this problem, many researchers have proposed hybrid CNN-Transformer CD methods. These hybrid CNN-Transformer CD approaches achieve the complementarity of local features and global features by designing effective frameworks that simultaneously combine the advantages of CNNs and Transformers, thereby ameliorating the feature extraction capabilities of the CD network. Wang et al. combined a CNN extraction backbone and a visual Transformer to acquire high-level semantic information and change intensity tokens [66]. In [67], a geometric representation Transformer is devised for CD; it adopts a convolutional feature to guide the Transformer by using the geometric prior. In addition, Yin et al. built a hybrid CNN-Transformer network for CD. The approach can enhance the local and global feature representation ability by comprehensively utilizing the CNN, attention mechanisms, and Transformer to increase CD accuracy [68]. A so-called Hybrid-TransCD network was developed for CD in [69]. Hybrid-TransCD employs a hybrid multi-scale Transformer module to model multi-scale long-range dependencies and exploits token aggregation to represent differential features by a hybrid difference Transformer decoder. Further, recent CD networks that hybridize CNNs and Transformers focus on introducing novel attention mechanisms to enhance the interaction between semantic tokens, thereby

elevating the perception of changed areas [70–72]. More CD networks based on hybrid CNN-Transformer can be found in [73–76].

3. Methodology

In this section, the overall architecture of the proposed dual-domain Transformer (D^2Former) are first outlined in Section 3.1, which briefly illustrates the overview of the CD process over D^2Former. Additionally, in Sections 3.2 and 3.3, we describe the structures of the proposed semantic tokenizer and frequency tokenizer in detail, respectively.

3.1. Overview

To better extract valuable change information, a CNN-based encoder and decoder connected by skip connections are utilized as the backbone of the proposed D^2Former, as depicted in Figure 1. According to several recent CD methods [32,77], the encoder–decoder architecture with skip connections, which was designed in [39], has been proven efficient in CD and similar image segmentation tasks. With this backbone, the detailed change information from bi-temporal remotely sensed imagery can be better acquired, which promotes the recognition of changed objects for the utilized Transformer architecture. Further, during the decoding stage, the skip connections can help with the recovery of spatial information provided by the features from the CNN encoder, which promotes the fine-grained annotation of changed land-cover objects.

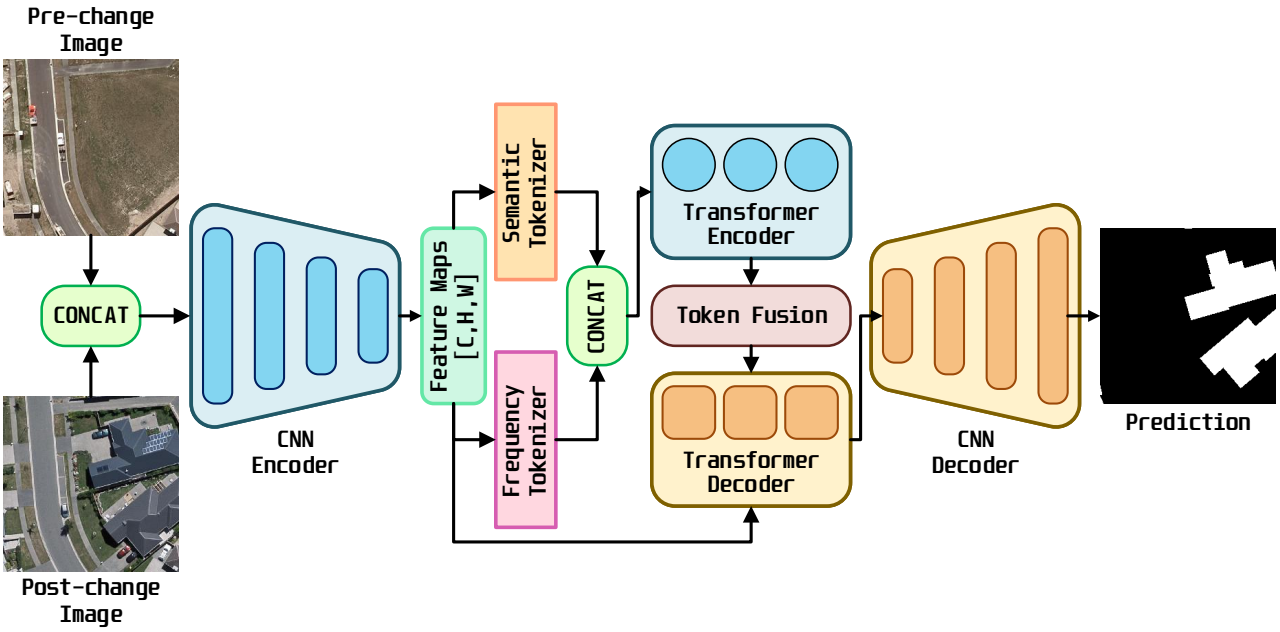

Figure 1. Overview of the proposed D^2Former. The pre-change image and post-change image denote the T1 and T2 images, respectively.

In more detail, the CNN encoder and decoder comprise four down-sampling 3×3 convolutional layers and four up-sampling 3×3 convolutional layers, respectively. Each convolutional layer consists of two 3×3 convolutional blocks with batch normalization (BN) and a rectified linear unit (ReLU) activation function and a 2×2 max pooling layer or a 2×2 bilinear interpolation layer to acquire multi-scale feature maps. When conducting the CD procedure, the pre-change and post-change remotely sensed images are fused in the channel dimension at the early stage and input into the CNN encoder; this has been validated as beneficial for CD tasks [29]. Then, the multilayered feature maps are obtained by the CNN encoder for the subsequent change information extraction and refinement. Among these multi-level features, the deepest features with the smallest spatial size and the largest channel size are input into the semantic and frequency tokenizers to acquire

semantic tokens and frequency tokens, which comprise rich semantic information and frequency information, respectively. The acquired semantic tokens and frequency tokens are concatenated at the channel dimension and input into the Transformer encoder with positional embeddings, which can integrate the location information, semantic information, and frequency information to better extract the change information. Moreover, the encoded tokens will be further fused by a 1×1 convolutional layer to acquire more fine-grained change feature representation. Afterwards, the fused tokens with rich semantic and frequency information and the features from the deepest layer of the CNN encoder are input into the Transformer decoder to acquire both the semantic-rich and frequency-rich features, which promote the accurate concentration of change information during the decoding stage in the CNN decoder. Then, the changed land cover is annotated in the pixel-wise prediction after being decoded by the CNN decoder. To increase clarity, the overall CD pipeline of the proposed D^{2}Former is shown in Algorithm 1.

To sum up, the proposed D^{2}Former employs dual-domain tokenizers to acquire a feature representation with rich semantic and frequency information; the tokenizers can refine the features to obtain more fine-grained CD performance. The utilized Transformer helps to build accurate recognition of meaningful land-cover changes with dual-domain token sets. The detailed structures of the semantic tokenizer and the frequency tokenizer will be introduced below.

Algorithm 1 CD Process of D^{2}Former

Input:
Bi-temporal pre-change and post-change images T^1, T^2;
Output:
Pixel-wise change map CM;

Obtain multi-level features F^i $(i = 1, 2, 3, 4, 5)$ with the CNN encoder as follows:
F^i = CNN_Encoder($[T^1, T^2]$);
Acquire semantic and frequency tokens with the semantic and frequency tokenizers as follows:
T^S = Semantic_Tokenizer(F^5);
T^F = Frequency_Tokenizer(F^5);
Acquire rich semantic and frequency feature representation as follows:
T^m = Transformer_Encoder($[T^S, T^F]$);
T^M = token_fusion(T^m);
F^M = Transformer_Encoder(F^5,T^M);
Acquire final change map as follows:
CM = CNN_Decoder(F^M,F^4,F^3,F^2,F^1);
return Output final binary change map CM.

3.2. Semantic Tokenizer

Semantic information is significant for neural-network-based CD method to identify changed land-cover objects. Based on this fact, we employ a semantic tokenizer to tokenize and serialize the semantic information extracted by the CNN encoder; the tokenizer can facilitate the enrichment and refinement of semantic information in the Transformer. The semantic tokenizer shares a similar concept with the semantic tokenizer in [36] and can extract compact semantic tokens from the feature maps, as presented in Figure 2. The mathematical representation of the semantic tokenizer can be shown as follows:

In the first place, let the input features be $F_{input} \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the channel size, height, and width, respectively, of the input features of the semantic tokenizer. To concentrate meaningful semantic information in the input feature maps, a 1×1 convolutional layer is firstly applied, which can be represented as:

$$F_{st} = conv1\left(F_{input}\right) \tag{1}$$

where $conv1(\cdot)$ denotes the aforementioned convolutional layer with a kernel size of 1×1. The convolutional layer greatly lowers the channel size C to the token length T^L, which, with the help of learnable parameters, makes it possible to collect useful semantic information and reduce irrelevant activations in the feature maps. Then, we reshape the processed features $F_{st} \in \mathbb{R}^{T^L \times H \times W}$ and use a softmax function to recalibrate these features in the spatial dimension; reshaping and recalibration further concentrate the semantic information in the spatial dimension. This step can be represented as:

$$F_{sl} = softmax(reshape(F_{st})) \tag{2}$$

where $reshape(\cdot)$ and $softmax(\cdot)$ respectively flatten and recalibrate the spatial dimension of F_{st}. To obtain the semantic tokens, we utilize $F_{sl} \in \mathbb{R}^{T^L \times HW}$ as attention scores to reweight the input features as follows:

$$T^S = F_{sl} \times reshape\left(F_{input}\right)^T \tag{3}$$

where $\times$ denotes matrix multiplication. Through the semantic tokenizer, each channel of the feature maps is represented as a token with a length of T^L and contains rich semantic information.

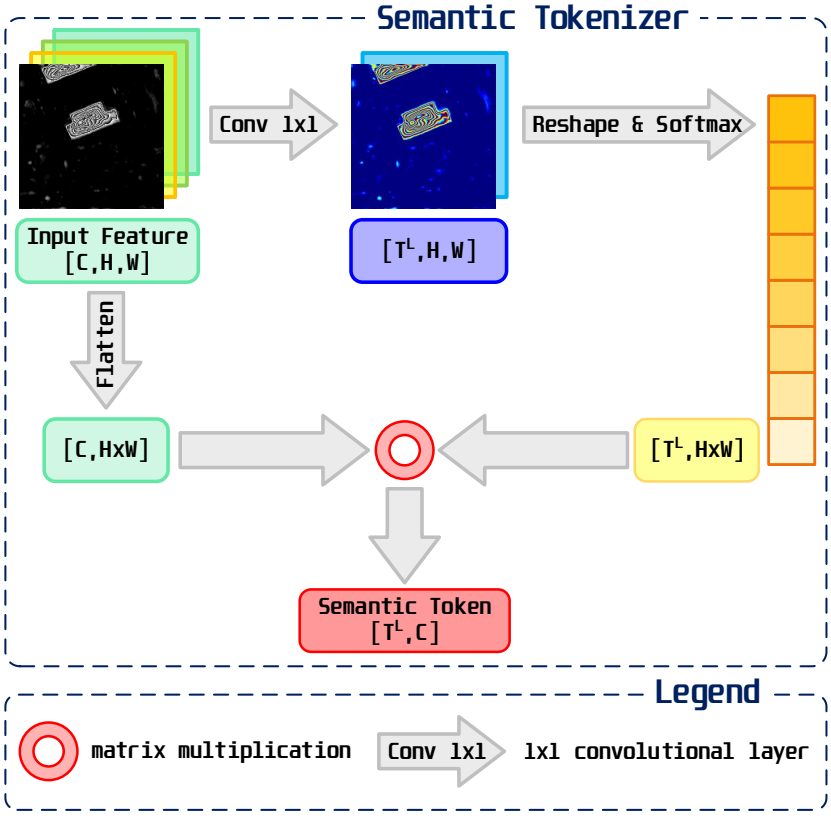

Figure 2. The structure of the semantic tokenizer. T^L indicates the token length.

3.3. Frequency Tokenizer

We consider frequency information as it has been validated as useful and helpful for CD tasks [32]. This is because the neural network is relatively weak in learning different frequency components. Therefore, the network's ability to extract ground features can be effectively improved by explicitly introducing the different frequencies of remote sensing images, thereby improving change detection performance. Therefore, we propose another feature tokenizer, a frequency tokenizer, to feed frequency information to the Transformer for better CD performance. Different from the semantic tokenizer, the proposed frequency

tokenizer utilizes a discrete cosine transform (DCT) to extract the frequency information of the feature maps from the CNN encoder; this offers the proposed D^{2}Former another aspect and dimension to extract change information. Afterwards, the proposed frequency tokenizer employs a similar structure compared to the semantic tokenizer, as shown in Figure 3. And the mathematical style of the FT can be denoted as follows:

Firstly, let the input features be $F_{input} \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the channel size, height, and width, respectively, of the input features of the frequency tokenizer. To focus on essential frequency features within the input feature maps, we first apply a convolutional layer with a kernel size of 1×1; this can be depicted as:

$$F_{ft} = conv1(F_{input}) \tag{4}$$

where $conv1(\cdot)$ represents the previously mentioned 1×1 convolutional layer. This layer can effectively reduce the channel size C to the token length T^L; we gather essential frequency information and minimize irrelevant frequency information in the feature maps through trainable parameters. Then, the features $F_{ft} \in \mathbb{R}^{T^L \times H \times W}$ are reshaped to $T^L \times HW$ and are recalibrated with a softmax function in the spatial dimension; the softmax function further integrates the frequency information in the spatial dimension as follows:

$$F_{fl} = softmax\Big(reshape\Big(F_{ft}\Big)\Big) \tag{5}$$

where the functions $reshape(\cdot)$ and $softmax(\cdot)$ respectively flatten and recalibrate the spatial dimensions of F_{ft}. To derive the frequency tokens, we leverage $F_{fl} \in \mathbb{R}^{T^L \times HW}$ as attention scores to adjust the input features in the following manner:

$$T^F = F_{fl} \times reshape\big(F_{input}\big)^T \tag{6}$$

where $\times$ represents matrix multiplication. To sum up, with the help of the frequency tokenizer and the semantic tokenizer, the frequency information and the semantic information are tokenized to promote the identification of changed land-cover objects. In this way, the proposed D^{2}Former can better analyze and distinguish variations in land cover over time, leading to more accurate and insightful results in land-cover monitoring and analysis.

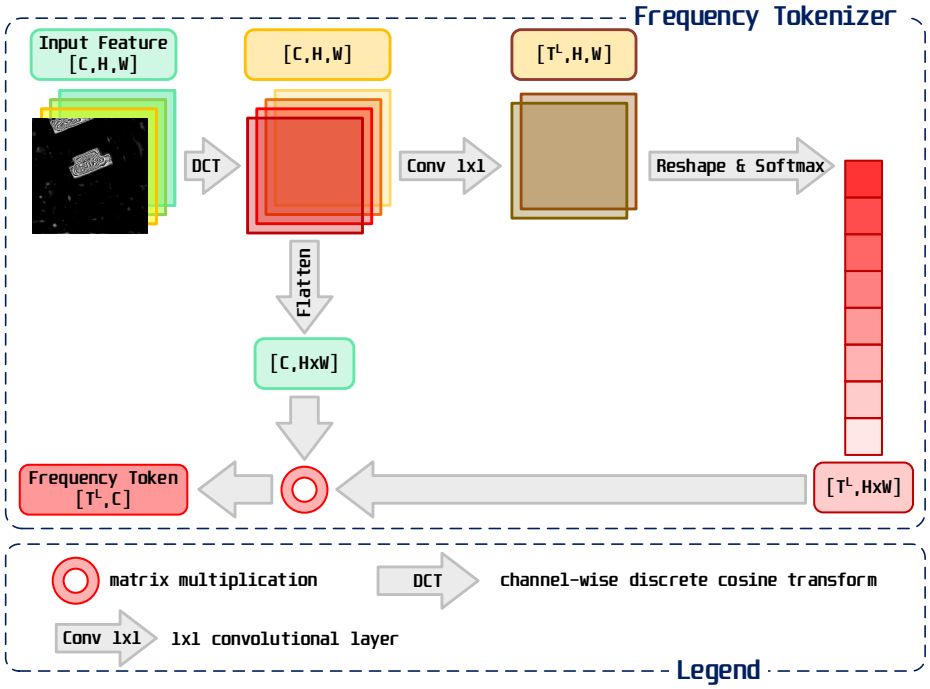

Figure 3. Illustration of the proposed frequency tokenizer. T^L indicates the token length.

4. Results

In order to verify the performance of the proposed D^{2}Former, we conduct experiments on three datasets. In this section, we first describe in detail the three datasets of the experiment: namely, WHU-CD [78], LEVIR-CD [30], and GZ-CD [79]; some example images are shown in Figure 4. Subsequently, benchmark methods and evaluation metrics are introduced for experimental comparisons. Moreover, we provide the implementation details of the proposed D^{2}Former and the other benchmark methods. Finally, the experimental results are obtained and analyzed.

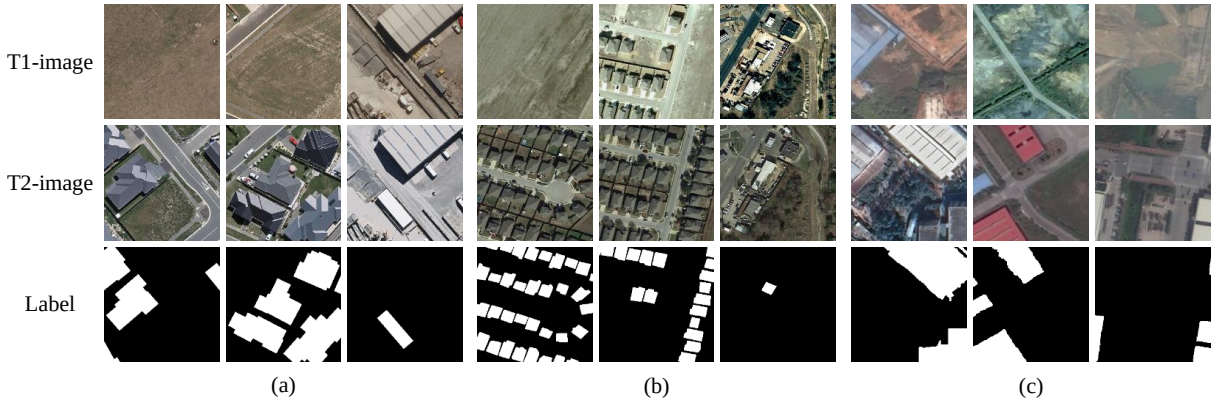

Figure 4. Some examples from the three experimental datasets: (**a**) WHU-CD dataset, (**b**) LEVIR-CD dataset, and (**c**) GZ-CD dataset.

4.1. Experimental Datasets

In our experiments, we selected three extensively used large public change detection datasets for validating the effectiveness of the proposed D^{2}Former, including: the WHU-CD dataset [78], the LEVIR-CD dataset [30], and the GZ-CD dataset [79]. The details of the three datasets are presented in Table 1. For the WHU-CD dataset, the division method of the dataset can be found in [31,32]. For the LEVIR-CD dataset, the division approach of the dataset remains consistent with the originally published version [30]. For the GZ-CD dataset, the partitioning method of this dataset remains consistent with [32,55]. In the experiments, in order to fit into the video memory, all data are divided into 256 × 256 images for training and testing. Some representative examples from the three experimental datasets are provided in Figure 4.

Table 1. Detailed descriptions of the three experimental datasets.

Dataset	Reference	Date	Sensors	Spatial Resolution	Size	Training/Test Samples
WHU-CD	Ji et al. [78]	2012–2016	Aerial	0.075 m/pixel	256 × 256	3480/1084
LEVIR-CD	Chen et al. [30]	2002–2018	Google Earth	0.5 m/pixel	256 × 256	7120/3072
GZ-CD	Peng et al. [79]	2006–2019	Google Earth	0.55 m/pixel	256 × 256	2191/939

4.2. Benchmark Methods and Evaluation Metrics

In this section, we introduce the benchmark methods and quantitative evaluation metrics in our experiments. Details as follow.

4.2.1. Benchmark Methods

- FC-EF [29]: FC-EF is one of the most common strategies in networks specifically designed for CD. The approach adopts an early fusion strategy to achieve CD by processing the CD task into a semantic segmentation task. Our proposed D^{2}Former also adopts an early fusion strategy. Therefore, it is necessary that the benchmark method is selected for comparison in our experiments.

- FC-Siam-C [29]: FC-Siam-C is another popular framework designed specifically for CD. It uses a structured Siamese encoder to extract the features of bi-temporal images; then, it concatenates these features and finally obtains a change map through a decoder.
- FC-Siam-D [29]: FC-Siam-D is also one of the widely employed frameworks that was particularly proposed for CD. This approach adopts a structured Siamese encoder to extract the hierarchical features of bi-temporal images; then, it calculates the difference between each hierarchical feature and gradually decodes the difference features and segments the difference features into binary change maps.
- STANet [30]: Chen et al. developed a fashionable CD approach, named STANet, which is a spatial–temporal attention-based network for CD. STANet exploits a spatial pyramid module and a self-attention mechanism to capture spatial–temporal long-range dependencies at different scales. The approach is an extensively utilized benchmark comparison method. Hence, we select STANet as a comparative CD approach.
- BIT [36]: BIT is currently a mainstream Transformer-based CD network that can capture the contexts within the spatial–temporal domain by a vision Transformer encoder and decoder. It is worth noting that BIT is actually a network that utilizes both a CNN-based encoder and a Transformer-based network. Therefore, since it is one of the excellent and popular methods based on a Transformer and has high similarity with our proposed D^{2}Former, we choose BIT as a comparison method in our experiments.
- ChangeFormer [37]: ChangeFormer is a Siamese CD network based on a Transformer and multi-layer perception to improve detection accuracy by capturing multi-scale long-range information. ChangeFormer is a convincing pure-Transformer-based CD network; therefore, we choose this method for comparison to verify the superiority of our proposed D^{2}Former.

4.2.2. Evaluation Metrics

To quantitatively evaluate the accuracy of the proposed method and other comparison methods, we select four extensively used accuracy evaluation metrics: namely, *Precision*, *Recall*, *F1-Score*, and mean intersection over union (*mIoU*). To calculate these indicators, we constructed a confusion matrix, which contains true positive (*TP*) pixels, true negative (*TN*) pixels, false positive (*FP*) pixels, and false negative (*FN*) pixels. Based on this, these four evaluation metrics can be defined as follows.

- *Precision*: This metric represents the proportion of correctly detected changed pixels to real changed pixels. It can be calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- *Recall*: This metric represents the proportion of correctly detected changed pixels to predicted changed pixels. It can be computed as:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

- *F1-Score*: This metric represents the weighted harmonic average of *Precision* and *Recall* and is used to comprehensively evaluate the performance of prediction results. It can be measured as:

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

- *mIoU*: This indicator represents the average of the changed intersection over union (*cIoU*) and the unchanged intersection over union (*ucIoU*) to comprehensively evaluate

the accuracy of the entirety of the prediction results from both the changed and unchanged perspectives. It can be calculated as:

$$mIoU = \frac{1}{2} \times (cIoU + ucIoU) \tag{10}$$

where *cIoU* and *ucIoU* can be defined as:

$$cIoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$ucIoU = \frac{TN}{TN + FP + FN} \tag{12}$$

4.3. Implementation Details

In the experiments, we selected six CD approaches based on CNNs and Transformers as comparative benchmark methods, as presented in Section 4.2.1. To ensure the fairness of the comparison, we re-executed all comparison methods through the source codes and the optimal parameter settings. In addition, our proposed D^{2}Former was deployed on the PyTorch framework with CUDA version 11.3. Our approach was trained on an NVIDIA 3090 graphics card with 24 GB video memory. In the proposed D^{2}Former, Adam was employed as the optimizer for training, and the initial learning rate was set to 0.0001. In addition, the batch size was set to 8 and 4 in the training stage and testing stage, respectively.

4.4. Comparison with Other Methods

4.4.1. Results on WHU-CD Dataset

According to the above configuration, we performed the proposed D^{2}Former and other comparison methods on the WHU-CD dataset and obtained corresponding quantitative accuracy and visual results. Here, we compare and analyze the quantitative accuracy and visual results for different methods on the WHU-CD dataset in detail.

Table 2 lists the quantitative evaluation metrics (*Precision*, *Recall*, *F*1-*Score*, and *mIoU*) for all methods for the WHU-CD dataset in our experiments. To visually display the performances of different methods, we highlight the best results for each evaluation metric. From Table 2, our proposed method obtains significant performance improvements compared with other popular CNN-based and Transformer-based benchmark methods. For instance, compared with the fashionable BIT [36], our proposed D^{2}Former improves the two comprehensive indicators of *F*1-*Score* and *mIoU* by 4.72% and 4.51%, respectively. Compared with ChangeFormer [37], our proposed approach boosts the *F*1-*Score* and *mIoU* by 6.20% and 5.86%, respectively. Our proposed method achieved the best performance for three indicators: namely, *Precision*, *F*1-*Score*, and *mIoU* at 94.36%, 92.85%, and 92.51% accuracy, respectively. However, although STANet [30] achieved the best *Recall* (91.40%), the *Precision* was the lowest (79.20%). Overall, our method effectively introduces and integrates semantic tokens and frequency tokens, which obviously improves the feature representation ability of bi-temporal remote sensing images and helps to improve the ability to identify changes. Similar conclusions can be acquired when comparing with other methods.

Table 2. Quantitative accuracies of different networks on the WHU-CD dataset.

Method	*Precision* (%)	*Recall* (%)	*F1-Score* (%)	*mIoU* (%)
FC-EF [29]	87.99	81.83	84.80	85.12
FC-Siam-C [29]	48.33	89.72	62.82	66.75
FC-Siam-D [29]	57.99	90.14	70.57	72.91
STANet [30]	81.99	**93.43**	87.34	87.19
BIT [36]	86.94	89.36	88.13	88.00
ChangeFormer [37]	85.07	88.28	86.65	86.65
Proposed D^{2}Former	**94.36**	91.39	**92.85**	**92.51**

In addition to quantitative comparisons, visual results are also compared to further validate the superiority of our method. Figure 5 presents some representative examples on the WHU-CD dataset for the change maps of various approaches. Observing Figure 5, we can see that each method can effectively detect changed buildings, and there are few missed detections. However, many methods inevitably suffer from a wide range of false detections. For example, as shown in the third row of Figure 5, almost all comparison methods present large areas of red false-detected pixels. The reason is that the characteristics of ground objects with similar materials, structures and spectra as buildings are difficult to effectively distinguish from the features of buildings, which interferes with the recognition of building targets and leads to large areas of erroneously detected pixels. A similar situation occurs in the last row of Figure 5. Notably, compared with the selected comparison methods, the proposed D^{2}Former demonstrates better results in these examples and shows very few red misdetected pixels and green missed pixels in large-scale or small-scale target scenes. On the whole, the proposed approach also exhibits obvious advantages in visual results: providing the same conclusion as the quantitative comparison.

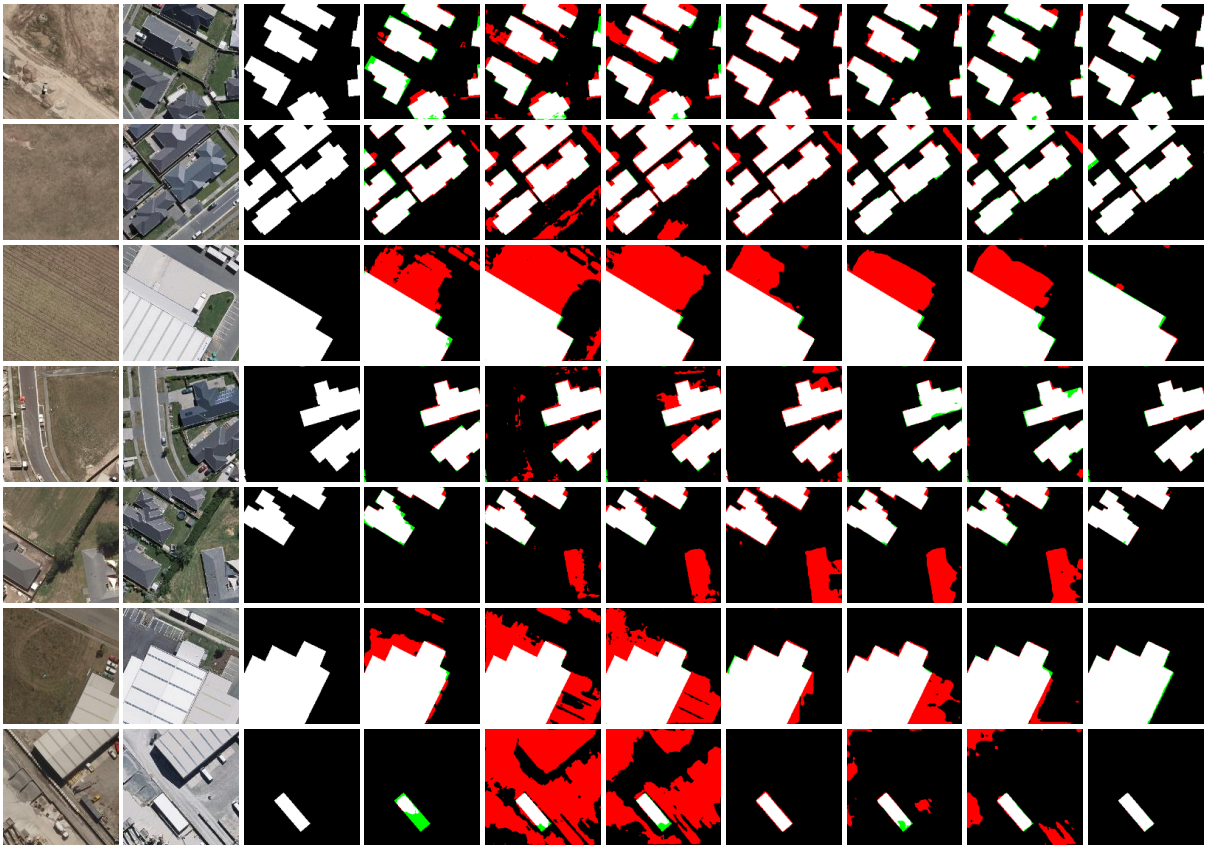

Figure 5. The visual results for different methods on WHU-CD dataset: (**a**) T1-time image, (**b**) T2-time image, (**c**) ground truth image, (**d**) FC-EF [29], (**e**) FC-Siam-C [29], (**f**) FC-Siam-D [29], (**g**) STANet [30], (**h**) BIT [36], (**i**) ChangeFormer [37], and (**j**) proposed D^{2}Former. Note: white, black, red, and green denote true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively.

4.4.2. Results on LEVIR-CD Dataset

Based on the above settings, we deployed the proposed D^{2}Former and other benchmark approaches on the LEVIR-CD dataset and acquired corresponding quantitative accuracy and visualization results. Herein, we illustrate the detailed comparison and analysis of the quantitative accuracy and visualization results of different methods on the LEVIR-CD dataset.

As shown in Table 3, the quantitative evaluation metrics of all methods on the LEVIR-CD dataset are provided in our experiment. For the LEVIR-CD dataset, FC-Siam-D [29] and STANet [30] reached the best accuracies (93.16% and 91.40%) in terms of *Precision* and *Recall*, respectively. The proposed method achieved the second-best results for the *Precision* and *Recall* indicators: namely, 92.32% and 88.84%, respectively. Nonetheless, our method achieves the highest accuracy for the two comprehensive evaluation metrics of *F*1-*Score* and *mIoU* (90.60% and 90.94%, respectively). For instance, compared to the second-best BIT [36], the proposed method is significantly improved for four evaluation indicators: that is, *Precision*, *Recall*, *F*1-*Score*, and *mIoU* increased by 0.89%, 2.55%, 1.76%, and 1.52%, respectively. Compared with the recent ChangeFormer [37], the proposed approach also obtains better performance on the four evaluation metrics for the LEVIR-CD dataset. Therefore, our proposed D^{2}Former also presents an obvious advantage on the LEVIR-CD dataset.

Table 3. Quantitative accuracies of different networks on the LEVIR-CD dataset.

Method	*Precision* (%)	*Recall* (%)	*F*1-*Score* (%)	*mIoU* (%)
FC-EF [29]	86.63	82.84	84.69	85.98
FC-Siam-C [29]	92.11	82.36	86.96	87.85
FC-Siam-D [29]	**93.16**	79.19	85.61	86.75
STANet [30]	79.20	**91.40**	84.86	86.03
BIT [36]	91.43	86.39	88.84	89.42
ChangeFormer [37]	89.44	87.18	88.30	88.94
Proposed D^{2}Former	92.32	88.94	**90.60**	**90.94**

To further observe the effectiveness of the proposed method, we provide visual change plots for the proposed method and other benchmark methods, as shown in Figure 6. Observing Figure 6, we can find that all benchmark networks achieved good detection results on the LEVIR-CD dataset. However, the detection capabilities of different models for different target scales are significantly different for the LEVIR-CD dataset. For instance, as shown in the second row of Figure 6, FC-EF [29], FC-Siam-D [29], STANet [30], and BIT [36] all have difficulty with accurately detecting single small building changes. In addition, FC-Siam-C [29] and ChangeFormer [37] are able to detect some changed areas but still cannot avoid some missed detection pixels. Moreover, other comparison methods perform poorly at detecting detailed information, such as building outlines; these include STANet [30] and ChangeFormer [37]. Overall, compared with these benchmark methods, our method not only shows good detection accuracy on buildings at various scales, but it also reaches good performance for detecting building contours. That is because the proposed D^{2}Former introduces frequency tokens, which can help enhance the detection ability for building contours. Hence, the visual results also demonstrate the advantages of the proposed D^{2}Former.

4.4.3. Results on GZ-CD Dataset

To further establish the robustness and superiority of the proposed D^{2}Former, we conducted experiments on the GZ-CD dataset. In this part, we compare and analyze the quantitative accuracy and visual results of different methods on the GZ-CD dataset.

Table 4 provides the quantitative results of different networks on the GZ-CD dataset. For the GZ-CD dataset, FC-Siam-D [29] presented poorer performance in terms of overall performance indicators *F*1-*Score* and *mIoU*. The main reason is that this method has a small number of model parameters, making it difficult to effectively capture more discriminative features in complex scenarios and thus limiting the accuracy for CD. Among CNN-based methods, STANet [30] presents the best accuracy in terms of *F*1-*Score* and *mIoU* (79.35% and 80.62%, respectively). Compared with CNN-based networks, Transformer-based models achieved better CD results. For instance, compared with STANet [30], the evaluation indicators *F*1-*Score* and *mIoU* of BIT [36] are improved by 2.17% and 1.79%, respectively, and the evaluation indicators *F*1-*Score* and *mIoU* of ChangeFormer [37] are improved by

3.09% and 2.45%, respectively. It is worth noting that our proposed D^{2}Former obtains the best performance on the four evaluation metrics (91.07%, 83.32%, 87.02%, and 87.08%). The convincing results on the GZ-CD dataset once again verify the robustness and superiority of the proposed method. This demonstrates that our proposed D^{2}Former can elevate CD performance by aggregating semantic tokens and frequency tokens from dual-domain information. As presented in Figure 7, similar conclusions can be drawn from the comparison and analysis of visual results.

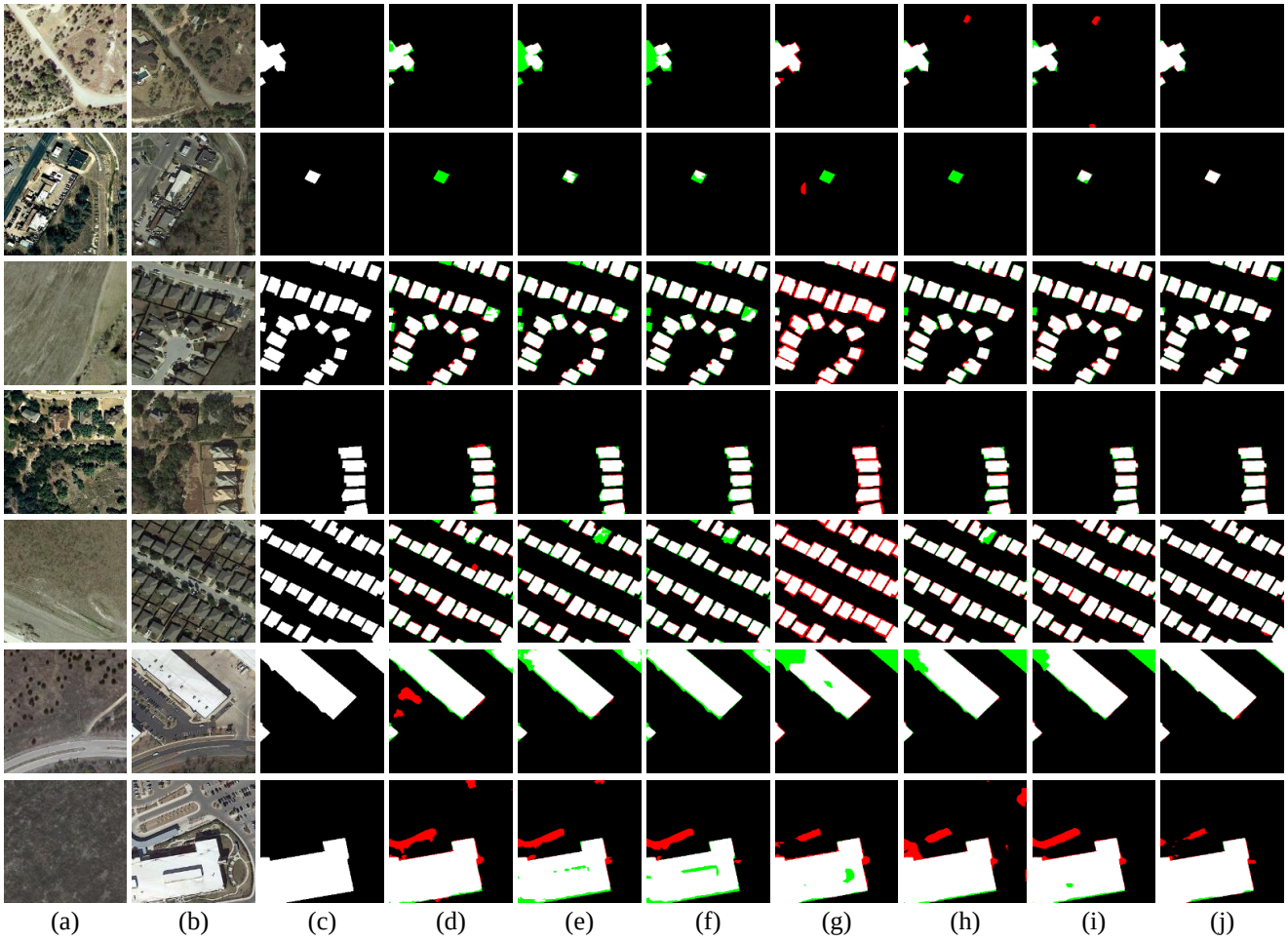

Figure 6. The visual results for different methods on the LEVIR-CD dataset: (**a**) T1-time image, (**b**) T2-time image, (**c**) ground truth image, (**d**) FC-EF [29], (**e**) FC-Siam-C [29], (**f**) FC-Siam-D [29], (**g**) STANet [30], (**h**) BIT [36], (**i**) ChangeFormer [37], and (**j**) proposed D^{2}Former. Note: white, black, red, and green denote true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively.

Table 4. Quantitative accuracies of different networks on the GZ-CD dataset.

Method	*Precision* (%)	*Recall* (%)	*F1-Score* (%)	*mIoU* (%)
FC-EF [29]	81.58	58.13	67.88	72.61
FC-Siam-C [29]	79.63	57.44	66.74	71.84
FC-Siam-D [29]	72.10	51.65	60.19	67.73
STANet [30]	82.98	76.02	79.35	80.62
BIT [36]	86.74	76.89	81.52	82.41
ChangeFormer [37]	83.66	81.25	82.44	83.07
Proposed D^{2}Former	**91.07**	**83.32**	**87.02**	**87.08**

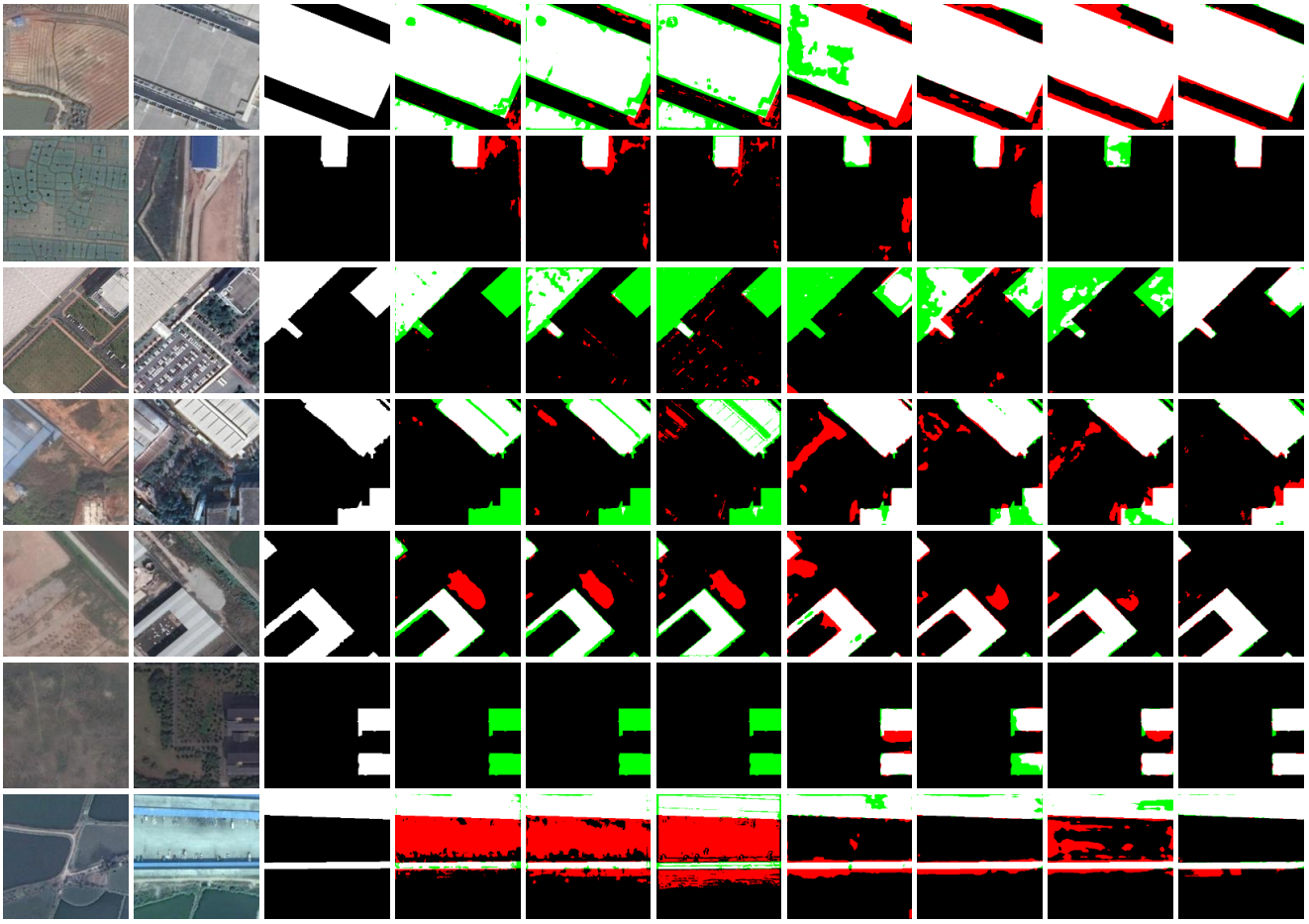

Figure 7. The visual results of different methods on the GZ-CD dataset: (**a**) T1-time image, (**b**) T2-time image, (**c**) ground truth image, (**d**) FC-EF [29], (**e**) FC-Siam-C [29], (**f**) FC-Siam-D [29], (**g**) STANet [30], (**h**) BIT [36], (**i**) ChangeFormer [37], and (**j**) proposed D^{2}Former. Note: white, black, red, and green denote true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively.

5. Ablation Studies and Discussion

In this section, to further explore the effectiveness of the proposed D^{2}Former, we performed sufficient ablation experiments using three experimental datasets: WHU-CD, LEVIR-CD, and GZ-CD. First, we carried out experiments using different components of the proposed D^{2}Former on these three datasets. Then, we visualized the features when using different components on the WHU-CD dataset to intuitively analyze the improvement of the feature representation ability of the proposed D^{2}Former. Ablation experimental details are as follows.

5.1. Ablation Studies on Different Datasets

In order to investigate the effects of different components of the proposed method, we decomposed the proposed method into three core components, including the backbone network, the semantic tokenizer, and the frequency tokenizer. Based on these components, we conducted ablation studies on three experimental datasets.

5.1.1. Ablation Studies on WHU-CD Dataset

Table 5 and Figure 8 present the quantitative accuracy and visual results for the ablation experiments on the WHU-CD dataset. From the perspective of quantitative accuracy, only using the backbone network can also perform well on the WHU-CD dataset:

obtaining an *F*1-*Score* of 90.16% and an *mIoU* of 89.96%. When we introduced the semantic tokenizer, the accuracy of the four evaluations was significantly improved, among which precision achieved the best performance, and the other three indicators—*Recall*, *F*1-*Score*, and *mIoU*—also obtained improvements of 2.00%, 1.10%, and 1.03%, respectively. In addition, compared with only using the backbone network, when the frequency tokenizer was deployed, three evaluation indicators—*Recall*, *F*1-*Score*, and *mIoU*—were also promoted (1.70%, 0.92%, and 0.85%, respectively). Observing Table 5, the introduction of the semantic tokenizer and the frequency tokenizer showed similar performance improvements. More importantly, after both the semantic tokenizer and frequency tokenizer were added to the backbone network, the accuracy of the overall indicators (*F*1-*Score* and *mIoU*) reached the best scores (92.85% and 92.51%, respectively). From Table 5, the proposed D^{2}Former can effectively enhance CD accuracy by employing and aggregating the semantic tokenizer and the frequency tokenizer. Further, visual results of networks based on different components also present similar conclusions, as shown in the first three rows of Figure 8. For example, compared with the result of the backbone network (d), the change maps based on the semantic tokenizer (e) and the frequency tokenizer (f) obtain fewer red pixels. And the visual change map of the proposed D^{2}Former is closer to the ground truth image. Ablation experiments on the WHU-CD dataset have verified the effects of each component in the proposed D^{2}Former.

Table 5. Quantitative comparison of different branches on WHU-CD dataset for ablation studies. The best detection accuracy for each indicator is marked in bold.

Method	*Precision* (%)	*Recall* (%)	*F1-Score* (%)	*mIoU* (%)
Backbone	94.55	86.17	90.16	89.96
w/Semantic Tokenizer	**94.58**	88.17	91.26	90.99
w/Frequency Tokenizer	94.53	87.87	91.08	90.81
Full (Proposed D^{2}Former)	94.36	**91.39**	**92.85**	**92.51**

5.1.2. Ablation Studies on LEVIR-CD Dataset

As presented in Table 6 and Figure 8, we obtained quantitative accuracy and visual change maps on the LEVIR-CD dataset to further test the effectiveness of each component in the proposed D^{2}Former. For the LEVIR-CD dataset, introducing the semantic tokenizer or the frequency tokenizer provides a slight improvement in performance. Compared with the results of the backbone network, our proposed D^{2}Former realized the best accuracy for three indicators: *Precision*, *F1-Score*, and *mIoU*, with increases of 1.58%, 0.62%, and 0.55%, respectively. As can be seen from the results in the middle three rows of Figure 2, the proposed D^{2}Former can exhibit more accurate change maps for land-cover targets at different scales, including details such as contours, through fusing the semantic tokenizer and the frequency tokenizer. On the whole, each component of the proposed D^{2}Former is beneficial for enhancing the CD detection performance in complex scenes.

Table 6. Quantitative comparison of different branches on LEVIR-CD dataset for ablation studies. The best detection accuracy for each indicator is marked in bold.

Method	*Precision* (%)	*Recall* (%)	*F1-Score* (%)	*mIoU* (%)
Backbone	90.74	89.23	89.98	90.39
w/Semantic Tokenizer	91.67	**89.39**	90.51	90.87
w/Frequency Tokenizer	92.24	88.92	90.55	90.90
Full (Proposed D^{2}Former)	**92.32**	88.94	**90.60**	**90.94**

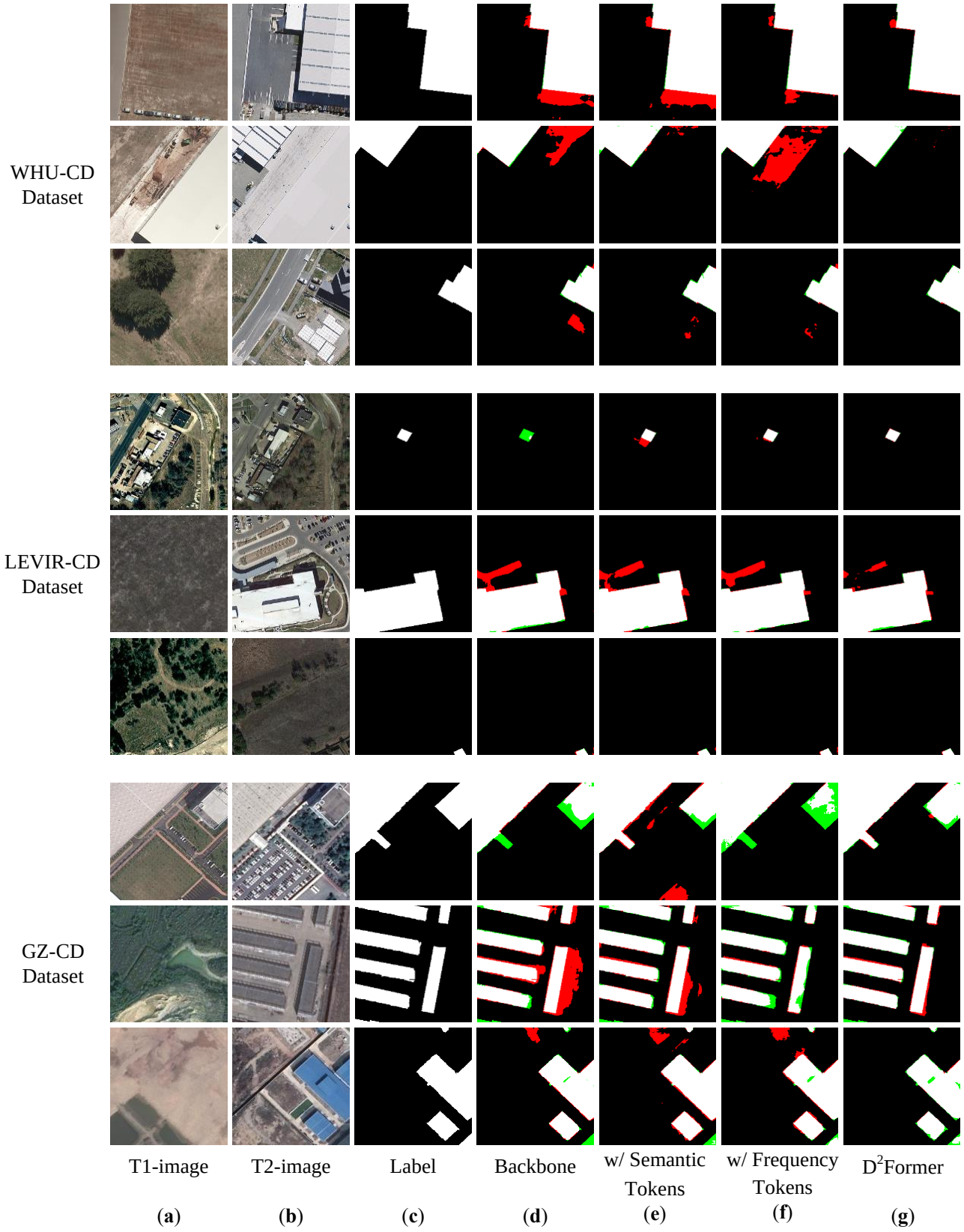

Figure 8. The results on different datasets for ablation studies: (**a**) T1-time image, (**b**) T2-time image, (**c**) ground truth image, (**d**) backbone, (**e**) w/semantic tokenizer, (**f**) w/frequency tokenizer, (**g**) full (proposed D^{2}Former). Note: white, black, red, and green denote true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively.

5.1.3. Ablation Studies On GZ-CD Dataset

To comprehensively illustrate the impacts of the different components in the proposed method, we also implemented experiments on the more challenging GZ-CD dataset. Table 7 and Figure 8 provide quantitative results and visual change maps, respectively, for combinations of different components. The GZ-CD dataset is more challenging than the WHU-CD dataset and the LEVIR-CD dataset. Therefore, the GZ-CD dataset can more adequately validate the impacts of different components in the proposed D^{2}Former. Observing the quantitative results of different component combinations, the introduction of the semantic tokenizer or the frequency tokenizer in the proposed approach can improve the detection performance for CD on the GZ-CD dataset. For instance, compared with the backbone network, after using the semantic tokenizer, the *F*1-*Score* and *mIoU* increased by 5.42% and 4.16%, respectively, and after exploiting the frequency tokenizer, the *F*1-*Score* and *mIoU* increased by 1.49% and 1.22%, respectively. In addition, when we deployed the semantic tokenizer and the frequency tokenizer simultaneously, the *F*1-*Score* and *mIoU* increased by 9.44% and 7.77%, respectively. Similarly, we can find from the comparison of the visual change maps that the results of the proposed D^{2}Former display fewer false detection pixels and missed detection pixels when the semantic tokenizer and the frequency tokenizer are adopted for the proposed approach. In summary, we introduced and combined the semantic tokenizer and the frequency tokenizer to refine and enhance the features of bi-temporal remote sensing images, thereby effectively improving CD performance.

Table 7. Quantitative comparison of different branches on GZ-CD dataset for ablation studies. The best detection accuracy for each indicator is marked in bold.

Method	*Precision* (%)	*Recall* (%)	*F*1-*Score* (%)	*mIoU* (%)
Backbone	83.81	72.20	77.58	79.31
w/Semantic Tokenizer	81.35	**84.73**	83.00	83.47
w/Frequency Tokenizer	87.68	72.00	79.07	80.53
Full (Proposed D^{2}Former)	**91.07**	83.32	**87.02**	**87.08**

5.2. Visualization Analysis

To visually display and analyze the performance for different components in the proposed method, we also visualized the feature maps of the model composed of different components on the WHU-CD dataset, as shown in Figure 9. For scenes without changes, it is difficult for the backbone network to avoid areas with many non-changing targets, as shown in the first and second rows of Figure 9. Moreover, the utilization of the semantic tokenizer or the frequency tokenizer can alleviate interference due to uninteresting changed targets to a certain extent. In addition, by observing the feature maps of the last three rows of data in Figure 9, we can find that the use of the semantic tokenizer can obtain a more complete feature map of the changed area, but it also highlights more erroneous areas. In the proposed D^{2}Former, the frequency tokenizer is introduced to significantly alleviate the problem of emphasizing wrong areas. On the whole, the proposed D^{2}Former more accurately highlights changed areas. In this subsection, the effectiveness of the proposed approach is demonstrated again through the above visual analysis.

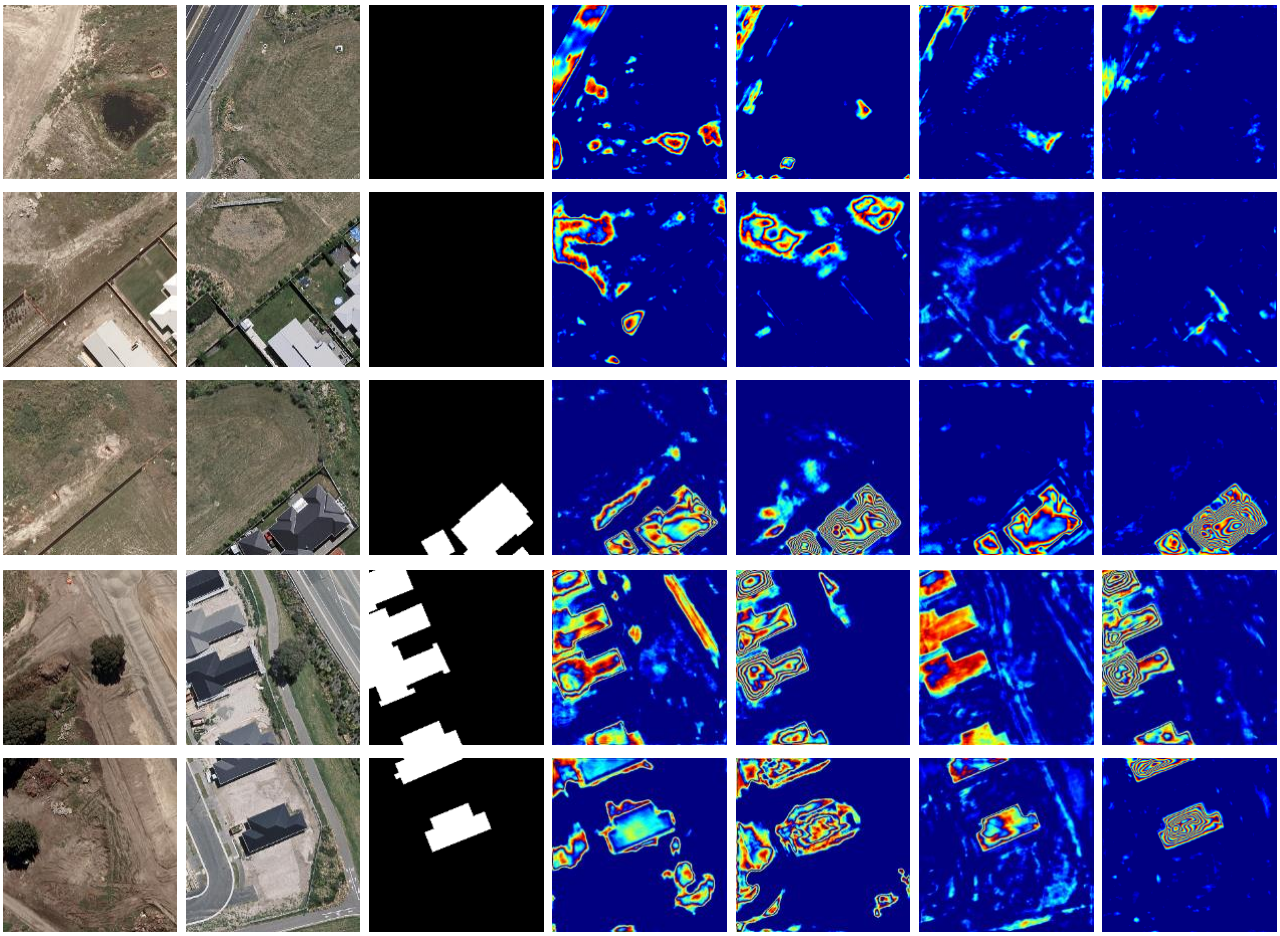

Figure 9. The feature maps for the combinations of different components: (**a**) T1-time image, (**b**) T2-time image, (**c**) ground truth image, (**d**) backbone, (**e**) w/semantic tokenizer, (**f**) w/frequency tokenizer, and (**g**) full (proposed D^{2}Former).

6. Conclusions

In this paper, a novel dual-domain Transformer (D^{2}Former) is proposed for CD. The D^{2}Former approach mainly consists of two novel components: the semantic tokenizer branch and the frequency tokenizer branch. The semantic tokenizer is employed to enhance semantic contextual feature representation, and the frequency tokenizer is introduced to consider critical frequency information from the frequency domain. Moreover, we also adopt a classic visual Transformer to aggregate the semantic tokenizer and the frequency tokenizer to further obtain more discriminative features for CD. Based on these configurations, the proposed D^{2}Former is able to provide more reliable discriminative information for feature extraction of bi-temporal remote sensing images, including semantic and frequency information, and to simultaneously reduce interference due to pseudo-variations triggered by other land-cover targets, thus improving CD performance. Extensive comparison experiments have been performed and have shown that the proposed D^{2}Former realizes better accuracy on three CD benchmark datasets. In addition, sufficient ablation studies have validated the effectiveness of different components in the proposed D^{2}Former. Similar conclusions were obtained from the comparison and analysis of visual features. Despite the proposed D^{2}Former showing some advantages on the three CD benchmark datasets, the method still relies on large-scale training sets for supervised learning. Hence, follow-up work will focus on extending the proposed D^{2}Former to a semi-supervised learning pipeline that can effectively alleviate the above limitation.

Author Contributions: Conceptualization, H.Z. and H.L.; methodology, H.Z., H.L., L.L. and S.L.; validation, J.L., H.Z. and L.L.; investigation, S.L., J.L. and L.L.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z., H.L., J.L., S.L. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Yulin Science and Technology Bureau Industry University Research Project under Grant No. CXY-2022-61 and the National Natural Science Foundation Regional Fund Project under Grant No. 12364057.

Data Availability Statement: The original WHU-CD dataset presented in the study is openly available from [78] at WHU-CD Dataset (http://gpcv.whu.edu.cn/data/building_dataset.html, accessed on 7 January 2024). The original LEVIR-CD dataset presented in the study is openly available from [30] at LEVIR-CD Dataset (https://chenhao.in/LEVIR/, accessed on 7 January 2024). The original GZ-CD dataset presented in the study is openly available from [79] at GZ-CD Dataset (https://github.com/daifeng2016/Change-Detection-Dataset-for-High-Resolution-Satellite-Imagery, accessed on 7 January 2024).

Acknowledgments: The authors would like to appreciate the researchers of the opening datasets and codes for their contributions.

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

References

1. Zhang, L.; Zhang, L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]
2. Wang, S.; Han, W.; Huang, X.; Zhang, X.; Wang, L.; Li, J. Trustworthy remote sensing interpretation: Concepts, technologies, and applications. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 150–172. [CrossRef]
3. Chen, Q.; Chen, Y.; Jiang, W. Genetic particle swarm optimization–based feature selection for very-high-resolution remotely sensed imagery object change detection. *Sensors* **2016**, *16*, 1204. [CrossRef] [PubMed]
4. Wu, Y.; Gong, M.; Miao, Q.; Qin, K. Computational Intelligence in Remote Sensing. *Remote Sens.* **2023**, *15*, 5325. [CrossRef]
5. Lv, Z.Y.; Liu, T.F.; Zhang, P.; Benediktsson, J.A.; Lei, T.; Zhang, X. Novel adaptive histogram trend similarity approach for land cover change detection by using bitemporal very-high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9554–9574. [CrossRef]
6. Wu, Y.; Li, J.; Yuan, Y.; Qin, A.K.; Miao, Q.G.; Gong, M.G. Commonality autoencoder: Learning common features for change detection from heterogeneous images. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4257–4270. [CrossRef] [PubMed]
7. Li, J.; Gong, M.; Liu, H.; Zhang, Y.; Zhang, M.; Wu, Y. Multiform ensemble self-supervised learning for few-shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
8. Lu, L.; Liu, T.; Jiang, F.; Han, B.; Zhao, P.; Wang, G. DFANet: Denoising Frequency Attention Network for Building Footprint Extraction in Very-High-Resolution Remote Sensing Images. *Electronics* **2023**, *12*, 4592. [CrossRef]
9. Wu, Y.; Ding, H.; Gong, M.; Qin, A.; Ma, W.; Miao, Q.; Tan, K.C. Evolutionary multiform optimization with two-stage bidirectional knowledge transfer strategy for point cloud registration. *IEEE Trans. Evol. Comput.* **2022**, *28*, 62–76. [CrossRef]
10. Wu, Y.; Gong, P.; Gong, M.; Ding, H.; Tang, Z.; Liu, Y.; Ma, W.; Miao, Q. Evolutionary multitasking with solution space cutting for point cloud registration. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 110–125. [CrossRef]
11. Asokan, A.; Anitha, J. Change detection techniques for remote sensing applications: A survey. *Earth Sci. Inform.* **2019**, *12*, 143–160. [CrossRef]
12. 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]
13. Lv, Z.; Liu, T.; Benediktsson, J.A.; Falco, N. Land cover change detection techniques: Very-high-resolution optical images: A review. *IEEE Geosci. Remote Sens. Mag.* **2021**, *10*, 44–63. [CrossRef]
14. Saha, S.; Bovolo, F.; Bruzzone, L. Building change detection in VHR SAR images via unsupervised deep transcoding. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 1917–1929. [CrossRef]
15. Liu, T.; Gong, M.; Jiang, F.; Zhang, Y.; Li, H. Landslide inventory mapping method based on adaptive histogram-mean distance with bitemporal VHR aerial images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
16. Ghorbanzadeh, O.; Xu, Y.; Ghamisi, P.; Kopp, M.; Kreil, D. Landslide4Sense: Reference Benchmark Data and Deep Learning Models for Landslide Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3215209. [CrossRef]
17. Ghorbanzadeh, O.; Gholamnia, K.; Ghamisi, P. The application of ResU-net and OBIA for landslide detection from multi-temporal sentinel-2 images. *Big Earth Data* **2023**, *7*, 961–985. [CrossRef]
18. Anniballe, R.; Noto, F.; Scalia, T.; Bignami, C.; Stramondo, S.; Chini, M.; Pierdicca, N. Earthquake damage mapping: An overall assessment of ground surveys and VHR image change detection after L'Aquila 2009 earthquake. *Remote Sens. Environ.* **2018**, *210*, 166–178. [CrossRef]

19. 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]
20. ZhiYong, L.; Wang, F.; Xie, L.; Sun, W.; Falco, N.; Benediktsson, J.A.; You, Z. Diagnostic analysis on change vector analysis methods for LCCD using remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10199–10212. [CrossRef]
21. Lv, Z.; Zhang, M.; Sun, W.; Benediktsson, J.A.; Lei, T.; Falco, N. Spatial-contextual information utilization framework for land cover change detection with hyperspectral remote sensed images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [CrossRef]
22. Lv, Z.; Lei, Z.; Xie, L.; Falco, N.; Shi, C.; You, Z. Novel Distribution Distance based on Inconsistent Adaptive Region for Change Detection Using Hyperspectral Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12. [CrossRef]
23. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [CrossRef]
24. Lv, Z.; Liu, T.; Shi, C.; Benediktsson, J.A.; Du, H. Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images. *IEEE Access* **2019**, *7*, 34425–34437. [CrossRef]
25. Gong, M.; Liang, Y.; Shi, J.; Ma, W.; Ma, J. Fuzzy c-means clustering with local information and kernel metric for image segmentation. *IEEE Trans. Image Process.* **2012**, *22*, 573–584. [CrossRef] [PubMed]
26. Lv, Z.; Liu, T.; Benediktsson, J.A. Object-oriented key point vector distance for binary land cover change detection using VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6524–6533. [CrossRef]
27. Lv, Z.; Yang, T.; Lei, T.; Zhou, W.; Zhang, Z.; You, Z. Spatial-Spectral Similarity based on Adaptive Region For Landslide Inventory Mapping with Remote Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4405111. [CrossRef]
28. Gong, M.; Zhou, Z.; Ma, J. Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering. *IEEE Trans. Image Process.* **2011**, *21*, 2141–2151. [CrossRef] [PubMed]
29. Daudt, R.C.; Le Saux, B.; 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; IEEE: Piscataway, NJ, USA, 2018; pp. 4063–4067.
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. Liu, T.; Gong, M.; Lu, D.; Zhang, Q.; Zheng, H.; Jiang, F.; Zhang, M. Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–17. [CrossRef]
32. Zheng, H.; Gong, M.; Liu, T.; Jiang, F.; Zhan, T.; Lu, D.; Zhang, M. HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images. *Pattern Recognit.* **2022**, *129*, 108717. [CrossRef]
33. 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*, 1–16. [CrossRef]
34. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
35. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
36. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
37. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 207–210.
38. 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.
39. 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; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
40. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
41. Gong, M.; Jiang, F.; Qin, A.K.; Liu, T.; Zhan, T.; Lu, D.; Zheng, H.; Zhang, M. A spectral and spatial attention network for change detection in hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
42. Lv, Z.; Liu, J.; Sun, W.; Lei, T.; Benediktsson, J.A.; Jia, X. Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
43. Yang, M.; Jiao, L.; Liu, F.; Hou, B.; Yang, S.; Jian, M. DPFL-Nets: Deep pyramid feature learning networks for multiscale change detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6402–6416. [CrossRef] [PubMed]
44. 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]
45. Zhang, M.; Zheng, H.; Gong, M.; Wu, Y.; Li, H.; Jiang, X. Self-structured pyramid network with parallel spatial-channel attention for change detection in VHR remote sensed imagery. *Pattern Recognit.* **2023**, *138*, 109354. [CrossRef]

46. Wang, R.; Ding, F.; Chen, J.W.; Liu, B.; Zhang, J.; Jiao, L. SAR image change detection method via a pyramid pooling convolutional neural network. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Virtual, 26 September–2 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 312–315.
47. Panda, M.K.; Subudhi, B.N.; Veerakumar, T.; Jakhetiya, V. Modified ResNet-152 Network with Hybrid Pyramidal Pooling for Local Change Detection. *IEEE Trans. Artif. Intell.* **2023** *5*, 1599–1612. [CrossRef]
48. Vudattu, V.S.; Pati, U.C. Change Detection in Synthetic Aperture Radar Images based on a Spatial Pyramid Pooling Attention Network (SPPANet). *Remote Sens. Lett.* **2023**, *14*, 1139–1149. [CrossRef]
49. Song, K.; Jiang, J. AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4816–4831. [CrossRef]
50. 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*, 1–16. [CrossRef]
51. Yang, X.; Hu, L.; Zhang, Y.; Li, Y. MRA-SNet: Siamese networks of multiscale residual and attention for change detection in high-resolution remote sensing images. *Remote Sens.* **2021**, *13*, 4528. [CrossRef]
52. Ma, J.; Shi, G.; Li, Y.; Zhao, Z. MAFF-Net: Multi-attention guided feature fusion network for change detection in remote sensing images. *Sensors* **2022**, *22*, 888. [CrossRef] [PubMed]
53. Li, J.; Zhu, S.; Gao, Y.; Zhang, G.; Xu, Y. Change detection for high-resolution remote sensing images based on a multi-scale attention Siamese network. *Remote Sens.* **2022**, *14*, 3464. [CrossRef]
54. Wang, L.; Li, H. Hmcnet: Hybrid efficient remote sensing images change detection network based on cross-axis attention mlp and cnn. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
55. Zhu, Y.; Jin, G.; Liu, T.; Zheng, H.; Zhang, M.; Liang, S.; Liu, J.; Li, L. Self-Attention and Convolution Fusion Network for Land Cover Change Detection over a New Data Set in Wenzhou, China. *Remote Sens.* **2022**, *14*, 5969. [CrossRef]
56. Jiang, H.; Hu, X.; Li, K.; Zhang, J.; Gong, J.; Zhang, M. PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection. *Remote Sens.* **2020**, *12*, 484. [CrossRef]
57. Xia, Y.; Xu, X.; Pu, F. Pcba-net: Pyramidal convolutional block attention network for synthetic aperture radar image change detection. *Remote Sens.* **2022**, *14*, 5762. [CrossRef]
58. Yang, Y.; Qu, J.; Xiao, S.; Dong, W.; Li, Y.; Du, Q. A deep multiscale pyramid network enhanced with spatial–spectral residual attention for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
59. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.S.; Khan, F.S. Transformers in remote sensing: A survey. *Remote Sens.* **2023**, *15*, 1860. [CrossRef]
60. 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*, 1–13.
61. 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.
62. 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*, 1–5.
63. Xu, X.; Li, J.; Chen, Z. TCIANet: Transformer-based context information aggregation network for remote sensing image change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1951–1971. [CrossRef]
64. Ma, J.; Duan, J.; Tang, X.; Zhang, X.; Jiao, L. Eatder: Edge-assisted adaptive transformer detector for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 1–15. [CrossRef]
65. Zhou, Y.; Huo, C.; Zhu, J.; Huo, L.; Pan, C. DCAT: Dual cross-attention-based transformer for change detection. *Remote Sens.* **2023**, *15*, 2395. [CrossRef]
66. Wang, G.; Li, B.; Zhang, T.; Zhang, S. A network combining a transformer and a convolutional neural network for remote sensing image change detection. *Remote Sens.* **2022**, *14*, 2228. [CrossRef]
67. Zhao, J.; Jiao, L.; Wang, C.; Liu, X.; Liu, F.; Li, L.; Yang, S. GeoFormer: A Geometric Representation Transformer for Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
68. Yin, M.; Chen, Z.; Zhang, C. A CNN-Transformer Network Combining CBAM for Change Detection in High-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 2406. [CrossRef]
69. Ke, Q.; Zhang, P. Hybrid-transcd: A hybrid transformer remote sensing image change detection network via token aggregation. *ISPRS Int. J. -Geo-Inf.* **2022**, *11*, 263. [CrossRef]
70. 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*, 1–16. [CrossRef]
71. 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*, 1–15. [CrossRef]
72. Liu, W.; Lin, Y.; Liu, W.; Yu, Y.; Li, J. An attention-based multiscale transformer network for remote sensing image change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 599–609. [CrossRef]
73. Deng, Y.; Meng, Y.; Chen, J.; Yue, A.; Liu, D.; Chen, J. TChange: A hybrid transformer-CNN change detection network. *Remote Sens.* **2023**, *15*, 1219. [CrossRef]
74. 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]

75. Li, W.; Xue, L.; Wang, X.; Li, G. ConvTransNet: A CNN-transformer network for change detection with multi-scale global-local representations. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
76. Fan, Z.; Wang, S.; Pu, X.; Wei, H.; Liu, Y.; Sui, X.; Chen, Q. Fusion-Former: Fusion Features across Transformer and Convolution for Building Change Detection. *Electronics* **2023**, *12*, 4823. [CrossRef]
77. Zheng, H.; Li, D.; Zhang, M.; Gong, M.; Qin, A.K.; Liu, T.; Jiang, F. Spectral Knowledge Transfer for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
78. 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]
79. 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.* **2020**, *59*, 5891–5906. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.