remote sensing

MDPI

Article

Sliding-Window Dissimilarity Cross-Attention for Near-Real-Time Building Change Detection

Wen Lu and Minh Nguyen *

School of Engineering, Computer & Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand; wen.lu@autuni.ac.nz
* Correspondence: minh.nguyen@aut.ac.nz

Abstract: A near-real-time change detection network can consistently identify unauthorized construction activities over a wide area, empowering authorities to enforce regulations efficiently. Furthermore, it can promptly assess building damage, enabling expedited rescue efforts. The extensive adoption of deep learning in change detection has prompted a predominant emphasis on enhancing detection performance, primarily through the expansion of the depth and width of networks, overlooking considerations regarding inference time and computational cost. To accurately represent the spatio-temporal semantic correlations between pre-change and post-change images, we create an innovative transformer attention mechanism named Sliding-Window Dissimilarity Cross-Attention (SWDCA), which detects spatio-temporal semantic discrepancies by explicitly modeling the dissimilarity of bi-temporal tokens, departing from the mono-temporal similarity attention typically used in conventional transformers. In order to fulfill the near-real-time requirement, SWDCA employs a sliding-window scheme to limit the range of the cross-attention mechanism within a predetermined window/dilated window size. This approach not only excludes distant and irrelevant information but also reduces computational cost. Furthermore, we develop a lightweight Siamese backbone for extracting building and environmental features. Subsequently, we integrate an SWDCA module into this backbone, forming an efficient change detection network. Quantitative evaluations and visual analyses of thorough experiments verify that our method achieves top-tier accuracy on two building change detection datasets of remote sensing imagery, while also achieving a real-time inference speed of 33.2 FPS on a mobile GPU.

Keywords: remote sensing; building change detection; transformer; cross-attention

Academic Editors: Wen Liu and Yonas Zewdu Ayele

Received: 6 September 2024
Revised: 14 November 2024
Accepted: 16 November 2024
Published: 2 January 2025

Citation: Lu, W.; Nguyen, M. Sliding-Window Dissimilarity Cross-Attention for Near-Real-Time Building Change Detection. *Remote Sens.* **2025**, *17*, 135. https://doi.org/10.3390/rs17010135

1. Introduction

Building Change Detection (BCD) is the process of discerning and analyzing changes in buildings over time, accomplished by assigning binary labels (unchanged or changed) at the pixel level. This encompasses the detection of changes like new construction, demolition, expansion, or other modifications to the physical structure of buildings. Real-time BCD is crucial for land use management and disaster response. Continuous monitoring of building changes across entire cities or regions enables timely interventions against illegal urban sprawl and encroachments on protected habitats. Following natural disasters, real-time BCD can quickly assess structural damage, facilitating prompt emergency response and resource allocation. BCD is often performed using aligned remote sensing data, such as satellite imagery or aerial photographs, combined with image processing and deep learning techniques to automatically identify and quantify these changes.

Onboard processing of satellite images for BCD offers several advantages over ground processing. First, satellite onboard processing enables preliminary analysis and feature extraction directly in space. This capability reduces the volume of data that needs to be transmitted to the ground by sending only pertinent information such as detected changes and specific features. This advantage is particularly beneficial in scenarios where bandwidth is limited or expensive. Secondly, processing data onboard the satellite allows for the analysis of sensitive information without transmitting raw imagery back to Earth. This approach enhances privacy and security by mitigating the risks associated with interception or unauthorized access to sensitive data. Thirdly, satellite onboard processing promotes more autonomous operation of the satellite system. This autonomy reduces the dependence on ground stations for processing tasks, enabling satellites to function independently and respond promptly to changing conditions or priorities. For instance, onboard processing facilitates adaptive imaging strategies where satellites can dynamically adjust their imaging parameters or focus areas based on preliminary analysis of onboard data. This capability significantly enhances the responsiveness of satellite systems to changing events. The widespread adoption of deep learning in change detection tasks has led many current methods to prioritize improving detection performance through increased network depth and width. However, these methods often overlook the stringent constraints imposed by embedded system platforms on satellites, such as low computational cost, minimal memory consumption, and the need for low-latency processing.

Fully Convolutional Networks (FCNs) [1] revolutionized research on change detection by introducing an end-to-end framework for pixel-level classification. This breakthrough signaled the dawn of a new era in leveraging neural networks for change detection, significantly advancing the capabilities and accuracy of such applications. In recent years, deep learning-based change detection has evolved from an early-fusion approach to a late-fusion approach. The early-fusion approach concatenates bi-temporal images and inputs them into a semantic segmentation network, which is then trained utilizing ground-truth labels. Nevertheless, unlike mono-temporal computer vision tasks, change detection involves extracting not only the semantics contained in mono-temporal images but also the embedded information pertaining to changes within the bi-temporal pairs. The early-fusion approach intertwines semantics in mono-temporal images with change-related information across a bi-temporal pair, resulting in semantic confusion. To address this issue, the late-fusion approach adopts a strategy of separating feature extraction and fusion. It begins with the separate extraction of features from each mono-temporal image, followed by the integration of these features. In their late-fusion methodologies, Weber et al. integrated bi-temporal features through concatenation [2], whereas Gupta et al. employed subtraction for fusion [3]. Concatenation proves effective in preserving building features, despite lacking foresight regarding change. Conversely, subtraction allows for the anticipation of change yet results in the forfeiture of building features and is inadequate for addressing pseudo-changes arising from weather conditions, seasonal variations, disparities in image sources, or illumination discrepancies. Recent research endeavors have shifted towards employing diverse attention mechanisms instead of the aforementioned rudimentary fusion techniques, aiming to assign greater weights to informative segments of feature maps. For example, STANet utilizes self-attention to calculate attention weights between pairs of bi-temporal pixels at different spatial locations [4]. After integrating bi-temporal features and performing subtraction, ADS-Net utilizes a dual-path attention structure for to extract change-related features [5]. LGPNet integrates position and channel attention modules, facilitating adaptive selection and enhancement of high-response features [6]. In DTCDSCN, a dual-attention module is integrated to capture interdependencies among channels and spatial positions, thereby enhancing features [7]. IFNet expands upon the

fundamental concepts of FC-Siam-Conc [8] and integrates both a channel attention module and a spatial attention module subsequent to the merging of bi-temporal representations and upper-level change representations [9]. Nevertheless, these mainstream methods either independently utilize attention to improve features within each mono-temporal image [7,10,11] or adjust the weights of the combined bi-temporal features in either the spatial dimension or channel dimension [4–6,9,12–16] rather than utilizing attention mechanisms to model the spatio-temporal semantic correlations between the aligned pre-change and post-change images.

Lately, there has been a surge in the development of cross-attention mechanisms designed to capture temporal correlations within bi-temporal image pairs. Some CNN mechanisms empower networks to pay attention to pertinent positions in both images and grasp the spatio-temporal correlations among them. For instance, BDANet proposes a cross-directional attention structure aimed at exploring correlations between pre-change and post-change images [17]. Changer integrates multiple interaction layers into the encoder and introduces a flow-based dual-alignment fusion module to enable interactive alignment and feature fusion [18]. FCCDN devises a dense connection-based feature fusion module for the amalgamation of bi-temporal features [19]. PGLF introduces a multi-scale interaction module to augment the spatio-temporal relationships between paired change features, extracting robust change representations while considering bidirectional temporal changes [20].

Following their success in natural language processing, transformers have demonstrated superior performance over CNNs in various mono-temporal vision tasks, where the goal is to understand the content of the input image or frame without considering its temporal context, such as semantic segmentation and object detection. Transformers excel at capturing spatial dependencies within images. In BCD tasks, where the spatial arrangement and relationships between different building structures are crucial, transformers demonstrate the potential to acquire and utilize this spatial context effectively to detect changes. For example, in order to obtain long-range representations, ChangeFormer is configured as a Siamese network comprising transformers [21]. Its encoder adopts a hierarchical transformer architecture, while the decoder utilizes a Multi-Layer Perceptron (MLP). BiT leverages convolutional kernels in the shallow layers for local feature extraction and transformer blocks in the deeper layers to capture long-range dependencies [22]. This design synergizes the strengths of both CNNs and transformers, enhancing the model's capacity to effectively process and understand complex visual information. As a hybrid architecture merging UNet and transformer components, TransUNetCD employs a difference enhancement structure to produce difference features enriched with change-related information [23]. However, these transformer networks have three notable limitations. First, their self-attention mechanisms are optimized for mono-temporal tasks, making them inadequate for capturing temporal changes. Secondly, their global attention mechanisms exhibit computational complexity that increases quadratically in relation to the size of the image, becoming prohibitively expensive for large-sized remote sensing images and real-time dense prediction tasks like BCD. Thirdly, unlike object detection and semantic segmentation, which require a global receptive field, buildings in remote sensing imagery are typically small, and the adjacent areas do not have meaningful contextual relationships with the buildings. As a result, distant information may lack relevance or could potentially mislead, emphasizing the importance of prioritizing middle-range context over long-range dependencies.

To accurately represent the spatio-temporal semantic correlations between pre-change and post-change images, we create an innovative transformer attention mechanism named Sliding-Window Dissimilarity Cross-Attention (SWDCA), which detects spatio-temporal

semantic discrepancies by explicitly modeling the dissimilarity of bi-temporal tokens, departing from the mono-temporal similarity attention typically used in conventional transformers. In order to fulfill the near-real-time requirement, SWDCA employs a sliding-window scheme to limit the range of the cross-attention mechanism within a predetermined window/dilated window size. This approach not only excludes information that is distant or irrelevant but also reduces computational cost, making it more friendly to embedded systems with stringent limitations on computational resources and memory occupation. Figure 1 compares the accuracy, computational complexity, and number of network parameters on the S2Looking Dataset [24], evaluated using bi-temporal image pairs with a resolution of 512 × 512 pixels. Our SWDCA achieves an optimal balance between accuracy and computational complexity.

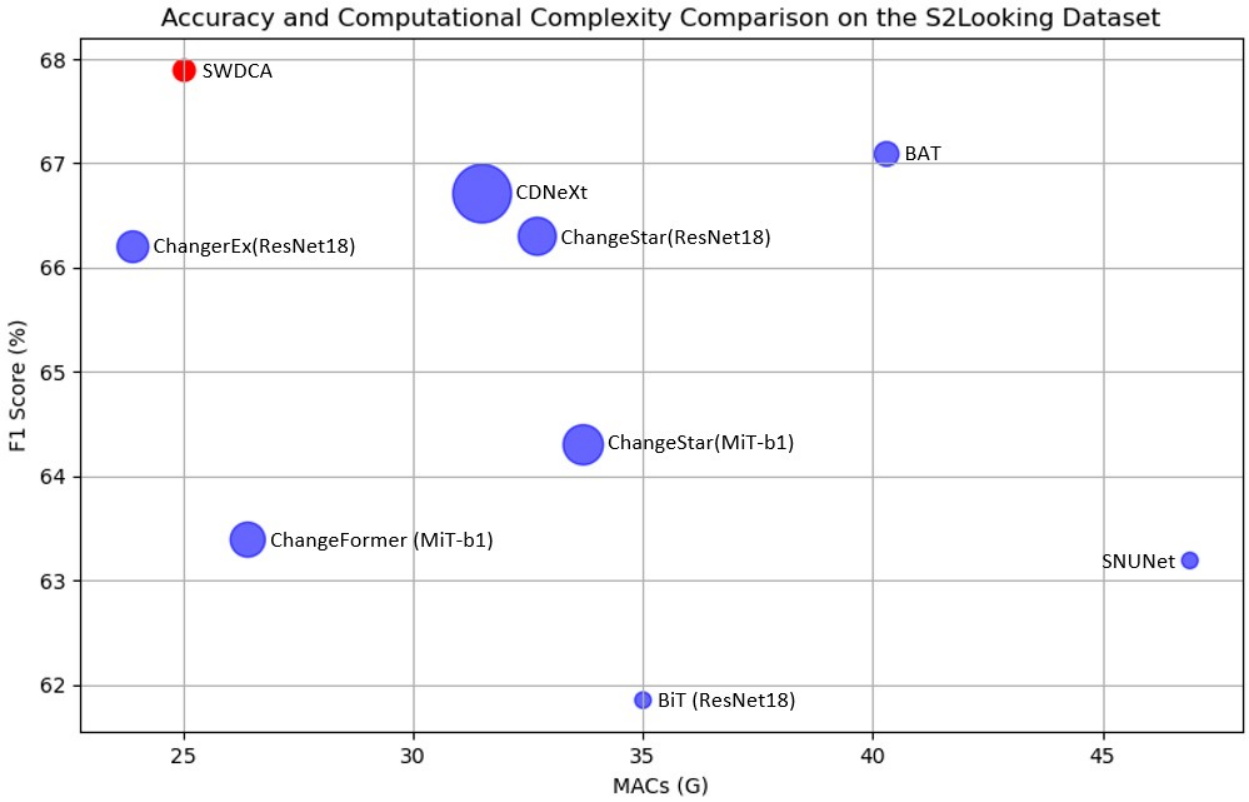

Figure 1. The size of circles represents the number of network parameters; circles positioned closer to the top left indicate better performance. The computational complexity, quantified by Multiply–Accumulate Operations (MACs), was evaluated using bi-temporal image pairs with a resolution of 512 × 512 pixels.

This study's contributions can be summarized in two main aspects:

1. We propose sliding-window dissimilarity cross-attention as an innovative transformer attention mechanism designed to effectively capture spatio-temporal semantic correlations across aligned bi-temporal images.
2. We develop a lightweight Siamese backbone for the extraction of building and environmental features. By integrating an SWDCA module into this backbone, we create an efficient change detection network that achieves top-tier accuracy on two building change detection datasets, with a real-time inference speed of 33.2 FPS on a mobile GPU for bi-temporal image pairs at a resolution of 512 × 512 pixels.

2. Related Work

The past few years have seen a notable emergence of transformer cross-attention mechanisms that are specifically designed to capture the temporal connections between a bi-temporal image pair. CTD-Former employs a cross-attention module that leverages the distinctions present in similarity matrices of paired image features [25]. In Cot-SR's cross-attention mechanism [26], features encoded from one temporal image initially focus on particular areas within itself, producing an attention matrix. Subsequently, the matrix is used to scale the other temporal features through multiplication. Cot-SR's cross-attention mechanism can be denoted as follows:

$$CA(\mathbf{Q1}, \mathbf{K1}, \mathbf{V2}) = \sigma(\mathbf{Q1} \times \mathbf{K1}^{\top} / \sqrt{d} + \mathbf{B}) \times \mathbf{V2} \tag{1}$$

$$CA(\mathbf{Q2}, \mathbf{K2}, \mathbf{V1}) = \sigma(\mathbf{Q2} \times \mathbf{K2}^{\top} / \sqrt{d} + \mathbf{B}) \times \mathbf{V1} \tag{2}$$

where the pre-change query matrix, pre-change key matrix, and pre-change value matrix are respectively represented as $\mathbf{Q1}, \mathbf{K1}, \mathbf{V1} \in \mathbb{R}^{hw \times d}$. Similarly, the post-change query matrix, post-change key matrix, and post-change value matrix are respectively represented as $\mathbf{Q2}, \mathbf{K2}, \mathbf{V2} \in \mathbb{R}^{hw \times d}$. Each row in the six matrices represents a single query, key, or value token; the σ symbol refers to the softmax operation; d denotes the query/key dimension; h and w represent the height and width of the feature map, respectively; and $\mathbf{B}$ indicates the relative position bias.

However, a drawback of this mechanism is its computation of weighted sums of values, which is decided by the relevance inferred from attention scores between queries and keys within the same temporal feature map. As a result, it primarily evaluates feature similarity and spatial correlation within a mono-temporal image, without directly comparing corresponding features across different temporal images.

On the other hand, BAT uses a cross-attention module based on bi-temporal mutual information [27]. In BAT's attention computation, queries are extracted from one temporal image, while keys are derived from other temporal images. By leveraging information from different time points, this process enhances the model's capacity to discern important features and relationships across time. BAT's cross-attention mechanism can be formulated as follows:

$$CA(\mathbf{Q1}, \mathbf{K2}, \mathbf{V2}) = \sigma(\mathbf{Q1} \times \mathbf{K2}^{\top} / \sqrt{d} + \mathbf{B}) \times \mathbf{V2} \tag{3}$$

$$CA(\mathbf{Q2}, \mathbf{K1}, \mathbf{V1}) = \sigma(\mathbf{Q2} \times \mathbf{K1}^{\top} / \sqrt{d} + \mathbf{B}) \times \mathbf{V1} \tag{4}$$

where $\mathbf{Q1}, \mathbf{K1}, \mathbf{V1} \in \mathbb{R}^{m^2 \times d}$ and $\mathbf{Q2}, \mathbf{K2}, \mathbf{V2} \in \mathbb{R}^{m^2 \times d}$, with $m \times m$ representing the size of a window/shifted window partition. $\mathbf{Q1} \times \mathbf{K2}^{\top}$ and $\mathbf{Q2} \times \mathbf{K1}^{\top}$ are regarded as bi-temporal mutual information, while $\mathbf{V1}$ and $\mathbf{V2}$ are regarded as the mono-temporal image features.

Within BAT's cross-attention mechanism, each token in one mono-temporal image identifies the most pertinent tokens in the other mono-temporal image through the computation of relevance based on feature space similarity. However, crucially, in change detection tasks, it is the dissimilar tokens that necessitate identification. Therefore, our proposed cross-attention mechanism detects spatio-temporal semantic discrepancies by explicitly modeling the dissimilarity of bi-temporal tokens.

In terms of the token range involved in cross-attention, Cot-SR employs a global attention mechanism, wherein each token attends to all other tokens within the same image. However, this approach introduces redundancy and quadratic complexity. SCanNet [28] utilizes a cross-shaped window transformer mechanism [29] to facilitate long-range modeling of temporal correlations. This mechanism divides the feature map into vertical and horizontal stripes. Nevertheless, this partitioning strategy preserves remote and irrelevant information while also ignoring essential features in the neighboring diagonal region. BAT

restricts cross-attention to tokens within fixed-sized partitions. It then applies a window-shifting strategy to the same input feature map to facilitate information exchange among tokens in adjacent partitions. However, its drawback becomes evident as the window-shifting strategy fails to ensure that every pixel is positioned at the center of cross-attention. Additionally, it introduces additional computational overhead due to the requirement for iterative processing of overlapping windows, increasing the overall complexity of the model. In contrast to existing methods, our proposed cross-attention mechanism utilizes a sliding-window strategy. Here, the attention range is centered at each pixel, and attention computation is confined to a predetermined window/dilated window size. Consequently, our approach effectively filters out distant and irrelevant information while encompassing all neighboring features.

3. Proposed Method

This section commences with the presentation of the our innovative cross-attention mechanism, followed by the introduction of our efficient change detection network.

3.1. Sliding-Window Dissimilarity Cross-Attention Module

Current cross-attention mechanisms overlook the distinction between the change detection task, which emphasizes dissimilarity, and other mono-temporal tasks, which emphasize relevance. As a result, they adhere to the paradigm of mono-temporal tasks by computing the weighted sum of value tokens based on the relevance inferred from attention scores between query tokens and key tokens. In contrast, our SWDCA module consolidates the query and key tokens into a single token, termed the similarity token, which encapsulates feature semantic likeness. Specifically, the similarity token matrices (**S1** and **S2**) and value token matrices (**V1** and **V2**) are obtained through linear projections of the pre-change and post-change feature maps (**F1**, **F2** $\in \mathbb{R}^{H\times W\times C}$, respectively):

$$\mathbf{S1} = \mathbf{W_s F1} \tag{5}$$

$$\mathbf{S2} = \mathbf{W_s F2} \tag{6}$$

$$\mathbf{V1} = \mathbf{W_v F1} \tag{7}$$

$$\mathbf{V2} = \mathbf{W_v F2} \tag{8}$$

where $\mathbf{W_s}, \mathbf{W_v} \in \mathbb{R}^{C\times C}$ are learnable weights.

As illustrated in Figure 2, the output of the SWDCA module at position (i, j), denoted as y_{ij}, can be expressed as follows:

$$CA(s1_{ij}, \widehat{\mathbf{S2}}, \widehat{\mathbf{V2}}) = \sigma(-s1_{ij} \times \widehat{\mathbf{S2}}^{\top} / \sqrt{d} + \mathbf{B}) \times \widehat{\mathbf{V2}} \tag{9}$$

$$CA(s2_{ij}, \widehat{\mathbf{S1}}, \widehat{\mathbf{V1}}) = \sigma(-s2_{ij} \times \widehat{\mathbf{S1}}^{\top} / \sqrt{d} + \mathbf{B}) \times \widehat{\mathbf{V1}} \tag{10}$$

$$y_{ij} = CA(s1_{ij}, \widehat{\mathbf{S2}}, \widehat{\mathbf{V2}}) \odot CA(s2_{ij}, \widehat{\mathbf{S1}}, \widehat{\mathbf{V1}}) \tag{11}$$

where $s1_{ij}$ and $s2_{ij}$ are the similarity tokens from **S1** and **S2** at position (i, j), respectively, and the $\odot$ symbol refers to the element-wise multiplication operation. For a token matrix (**X**), the $\widehat{\mathbf{X}}$ symbol indicates the stacking of tokens within a window (or dilated window with a dilation rate of r) of size $w \times w$ centered at (i, j) as a matrix and can be represented as follows:

$$\widehat{\mathbf{X}} = \left\{ x_{i'j'} | i' = i + p \times r, j' = j + q \times r \right\}, \tag{12}$$

with $-w/2 \leq p, q \leq w/2$.

The SWDCA module selects post-change similarity and value tokens within a window or dilated window centered at (i, j) for cross-attention computation with the pre-change similarity token located at (i, j) and vice-versa. Notice that within the softmax function, we invert the sign of the matrix multiplication result to emphasize dissimilar token pairs and suppress similar ones, aiming for a more suitable modeling of change detection. Furthermore, we segment the pre-change and post-change feature maps (**F1** and **F2**, respectively) along the channel dimension into several slices. Then, we conduct multi-head SWDCA with varying dilation rates to enable multi-scale cross-attention. Hence, the implementation of a sliding-window/dilated window strategy not only captures the requisite middle-range context but also adapts to misalignments. Moreover, it reduces computational complexity, leading to linear computational complexity in proportion to image size.

In remote sensing images, buildings often appear small. Therefore, a 3×3 sliding window with a dilation rate of 2 is sufficiently large to encompass most buildings on $1/8$ downscaled feature maps. Consequently, in subsequent experiments, we utilize dilation rates of 1 and 2 as network settings.

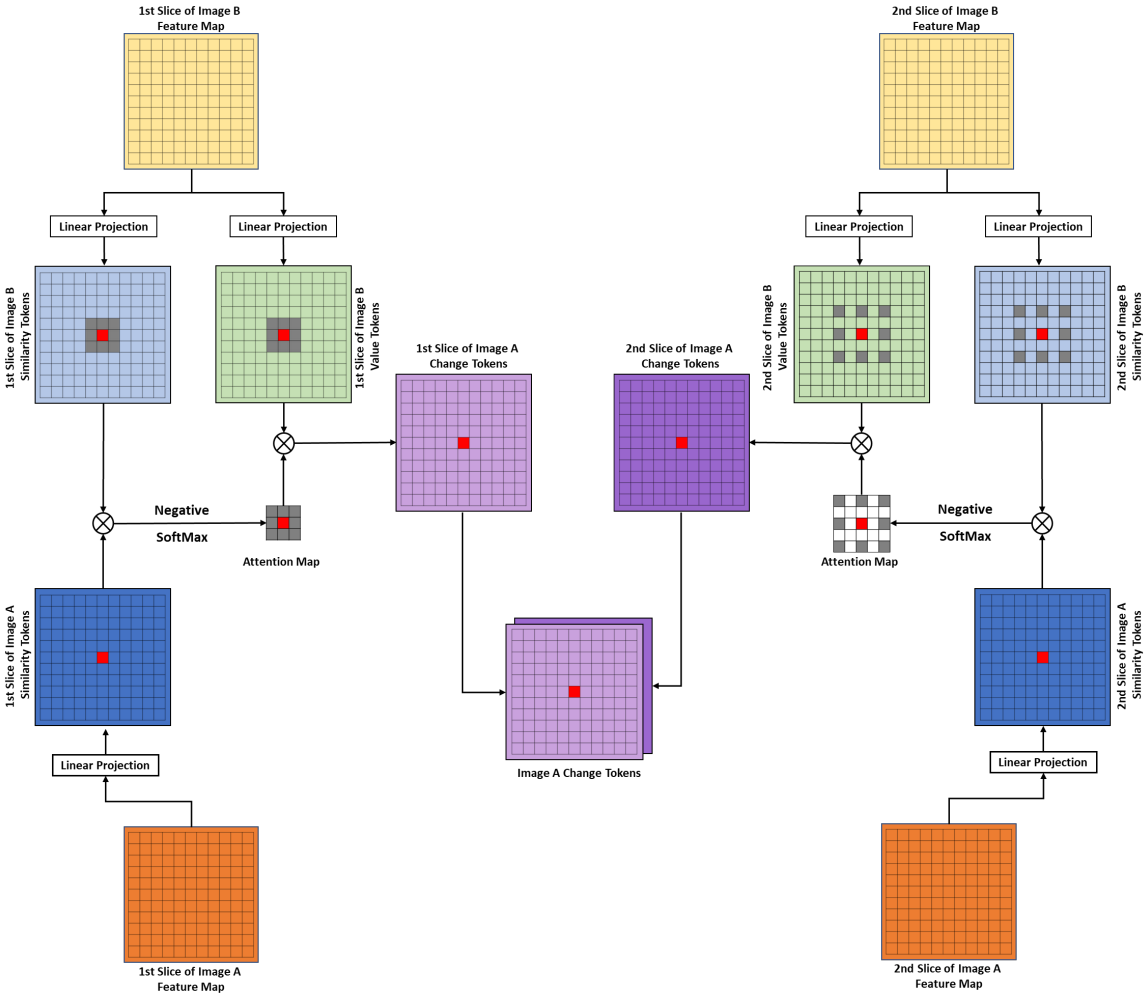

Figure 2. The structure of the sliding-window dissimilarity cross-attention module, where $(A, B) \in \{(1, 2), (2, 1)\}$ denote two time points and $\otimes$ represents matrix multiplication.

3.2. Efficient Change Detection Network

To prevent semantic confusion inherent in early-fusion methods, our method employs a late-fusion strategy that isolates mono-temporal feature extraction from bi-temporal feature fusion. As depicted in the bottom section of Figure 3, we use a weights-sharing Siamese

architecture to extract features equitably from every mono-temporal image. Afterwards, these mono-temporal feature maps are paired as input for the SWDCA module.

Many contemporary methods improve detection performance by utilizing a heavy-weight backbone, which requires the cropping of an image into multiple smaller patches to accommodate GPU memory constraints. However, such an initial processing step often fragments entire buildings, leading to the loss of crucial middle-range contextual knowledge. Furthermore, adjusting the outcomes to original dimensions causes further delays. To address these issues, we design a backbone that optimizes both computational complexity and memory utilization, thereby reducing the need for cropping.

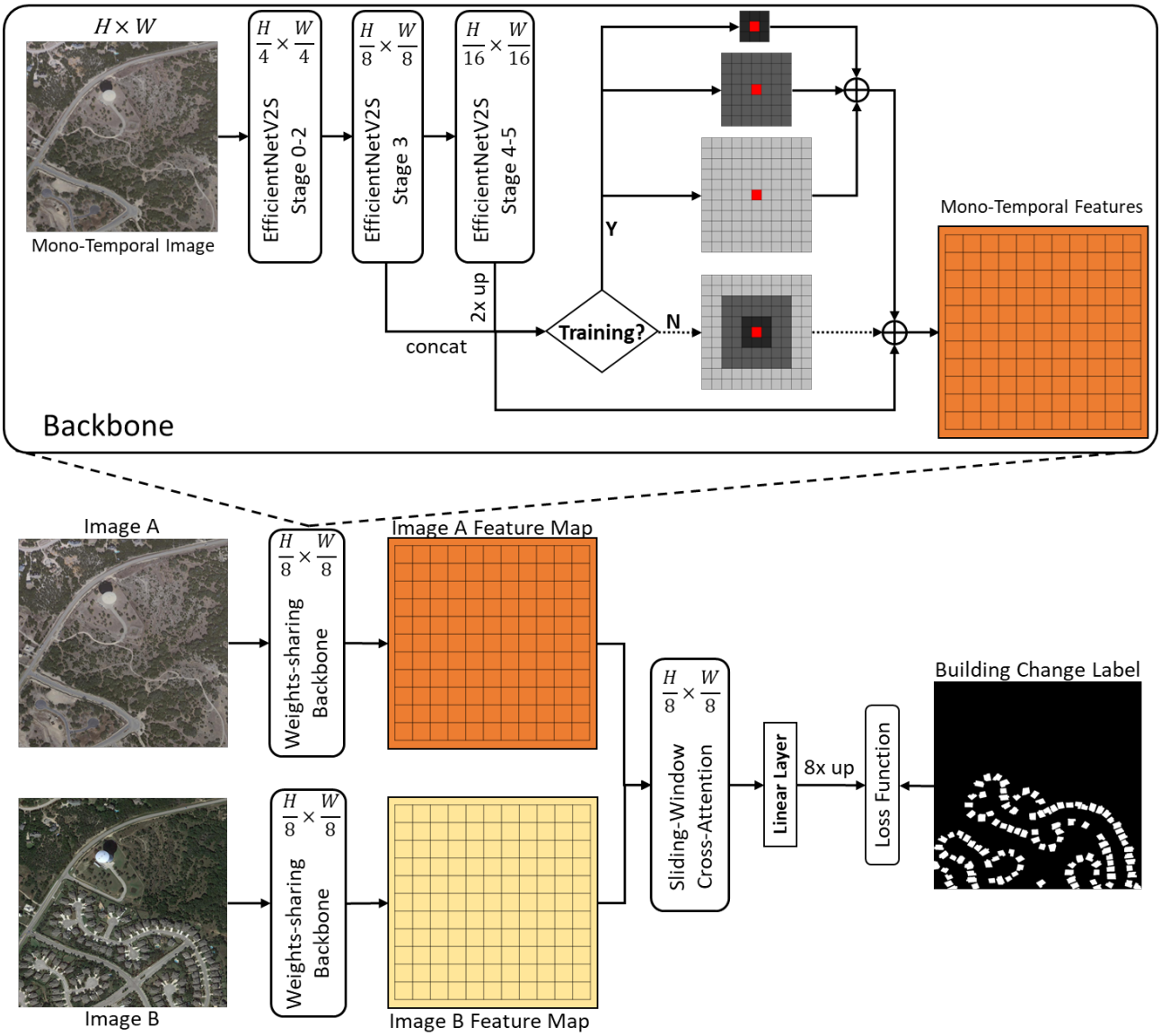

Figure 3. The architecture of our efficient change detection network, where ⊕ represents element-wise addition.

As depicted in the top section of Figure 3, we initially employ stages 0–3 and stage 4–5 of EfficientNetV2-S [30] (Table 1) to produce 1/8 and 1/16 downscaled feature maps, respectively. Subsequently, we upsample the 1/16 feature map by a factor of 2× and concatenate it with the 1/8 feature map to generate multi-scale features. To reduce computational complexity, we use only the shallow layers of EfficientNetV2, resulting in a limited receptive field. To compensate for this limitation, we employ three depthwise convolutions

with kernel sizes of 3 × 3, 7 × 7, and 11 × 11 connected in parallel during the training stage. These convolutions extract small, medium, and large features, respectively. The larger 7 × 7 and 11 × 11 kernels help the network achieve a receptive field comparable to that of deeper networks. During the inference stage, we re-parameterize the smaller 3 × 3 and 7 × 7 kernels into the larger 11 × 11 kernel by fusing batch normalization layers and combining kernel parameters. Consequently, the computational cost of the original three parallel branches is reduced to that of a single 11 × 11 large kernel.

Table 1. Stages 0–5 of the EfficientNetV2-S architecture.

Stage	Operator	Expan	Stride	#Channels	#Layers
0	Conv		2	24	1
1	Fused-MBConv	1	1	24	2
2	Fused-MBConv	4	2	48	4
3	Fused-MBConv	4	2	64	4
4	MBConv,SE0.25	4	2	128	6
5	MBConv,SE0.25	6	1	160	9

MBConv and Fused-MBConv blocks are illustrated in Figure 4. # means the number of.

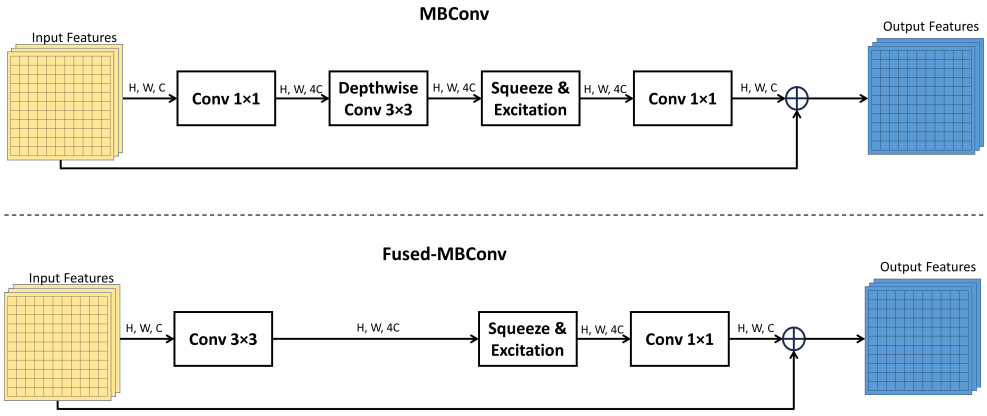

Figure 4. Structures of MBConv and Fused-MBConv [30].

The change detection head consists of a straightforward linear layer. The resulting logits undergo an eight-fold upscaling using bilinear interpolation prior to being fed into the loss function.

Table 2 shows that among the competitive models, our SWDCA network exhibits the second-lowest number of parameters and computational complexity. Although BiT and SNUNet have slightly fewer parameters than the SWDCA network, their computational complexities are significantly higher. The SWDCA network possesses fewer than half the parameters of ChangerEx (ResNet18) while maintaining nearly identical computational complexity. The SWDCA network also consumes less GPU memory during both the training and inference stages. This efficiency enables the training and inference of original images at a resolution of 1024 × 1024 pixels. In contrast, most competitive methods require the cropping of original images of size 1024 × 1024 pixels into smaller patches of 512 × 512 pixels or even 256 × 256 pixels to mitigate GPU memory overflow due to inefficient network design.

When processing bi-temporal image pairs at a resolution of 512 × 512 pixels on an NVIDIA RTX 3050 4 GB Mobile GPU (FP32: 5.5 TFLOPS), which is comparable in computational capacity to the NVIDIA Jetson AGX Orin 64GB embedded system (FP32: 5.3 TFLOPS), the SWDCA network achieved a real-time inference speed of 33.2 FPS. This is twice the inference speed of BAT, which achieved 17.8 FPS. This remarkable efficiency is credited to the implementation of a sliding-window/dilated window strategy, which

confines the cross-attention computation to a compact 3 × 3 pixel range, thereby curbing computational complexity. Consequently, the computational load transitions from quadratic with respect to image size to a linear scaling while still capturing the necessary mid-range context.

Table 2. Comparison of network parameters and computational complexity.

Method	#Param (M)	MACs (G)
STANet-BAM(ResNet18) [4]	12.2	49.2
STANet-PAM(ResNet18) [4]	12.2	50.2
DTCDSCN(SE-Res34) [7]	41.1	60.9
L-Unet [31]	8.5	
CDNet [32]	14.3	
MSCANet [33]	16.4	
BiT(ResNet18) [22]	**3.0**	35.0
SNUNet [14]	**3.0**	46.9
ChangeFormer(MiT-b1) [21]	13.9	26.4
IFN(VGG-16) [9]	36.0	316.5
FHD [34]	11.8	
ChangeStar(MiT-b1) [35]	18.4	33.7
Xu et al. [11]	61.4	
ChangerEx(ResNet18) [18]	11.4	**23.9**
ChangeStar(ResNet18) [35]	16.4	32.7
CDNeXt [36]	39.4	31.5
TransUNetCD [23]	95.5	
BAT [27]	6.9	40.3
SWDCA Network	**5.4**	**25.0**

The computational complexity, quantified by multiply–accumulate operations (MACs), was evaluated using bi-temporal image pairs with a resolution of 512 × 512 pixels. The **optimal** value is indicated in red font, whereas the **second-best** value is represented in blue font. # means the number of.

4. BCD Experimental Results

4.1. Experimental Setup

We conducted model training on an NVIDIA RTX 3090 GPU using PyTorch 2.3, with AdamW as the optimizer, a batch size of 8, and the initial learning rate set to 0.0001, applying cosine decay for learning rate adjustment. We utilized cross-entropy as the loss function. The models underwent training for 200 epochs, incorporating a warmup strategy for the initial 20 epochs. Throughout training, we implemented various data augmentation techniques, such as random flipping; rotation; scaling factors of 0.8, 0.9, 1.0, 1.25, and 1.5; random cropping to 768 × 768 pixels; and color adjustments. Following established research practices, we used precision, recall, and F1 score to evaluate the performance of various models. These criteria are formulated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

where TP denotes the number of pixels correctly identified as positive, FP denotes the number of pixels incorrectly identified as positive, and FN denotes the number of pixels incorrectly identified as negative.

Recall evaluates a model's ability to identify regions that have changed. Precision measures a model's effectiveness in excluding irrelevant and unchanged structures from the predictions. The $F1$ score provides an overall assessment of the prediction results.

4.2. LEVIR-CD+ Dataset Experimental Results

The LEVIR-CD+ BCD dataset [24], an extension of the LEVIR-CD dataset [4], contains 985 pairs of near-nadir satellite images. Each image pair has an image size of 1024 × 1024 pixels with 0.5 m per pixel spatial resolution. This dataset covers 20 distinct regions in Texas, spanning a 5-year time frame. The dataset is formally split into a training set comprising 637 pairs of bi-temporal images and a test set containing 348 pairs. Following standard protocols, we utilized the designated training set for network training purposes and reserved the designated test set for evaluation and result reporting.

As illustrated in Table 3, our method demonstrates significantly improved recall rates, affirming the effectiveness of the sliding-window dissimilarity cross-attention mechanism in capturing change features. Despite achieving superior recall rates, our method maintains a precision level comparable to that of top-performing methods, leading to the highest $F1$ score achieved.

Table 3. Evaluation of the SWDCA network against various models on the LEVIR-CD+ dataset.

Method	Precision (%)	Recall (%)	F1 (%)
FC-EF [8]	61.30	72.61	66.48
FC-Siam-Conc [8]	66.24	81.22	72.97
FC-Siam-Diff [8]	74.97	72.04	73.48
DSAMNet [37]	69.76	80.31	74.66
DTCDSCN [7]	80.36	75.03	77.60
L-Unet [31]	78.99	79.18	79.09
STANet-PAM [4]	74.62	84.54	79.31
SNUNet [14]	79.51	81.42	80.45
CDNet [32]	88.96	73.45	80.46
TFI-GR [38]	79.72	83.45	81.54
BiT [22]	82.74	82.85	82.80
A2Net [15]	85.25	81.27	83.21
MSCANet [33]	85.80	81.24	83.46
DCAT [39]	84.72	83.34	84.02
IFN [9]	85.82	83.24	84.51
Hu et al. [40]	88.74	83.63	86.11
AR-CDNet [41]	86.62	86.18	86.40
FHD [34]	89.60	83.83	86.62
CDNeXt [36]	**89.68**	84.73	87.14
BAT [27]	88.29	86.22	87.24
SWDCA Network	88.17	**86.68**	**87.42**

The bold entities denote the best performance.

Figure 5 shows a comparison of building change predictions by BAT and our method. BAT adheres to the mono-temporal task paradigm by assigning greater importance to pixels where the similarity between the query token and key token is higher. In contrast, our SWDCA network assigns greater weight to pixels with bi-temporal similarity tokens that exhibit greater variance. This visualization underscores the superiority of the SWDCA network over BAT in accurately identifying most building changes while successfully filtering out unchanged structures and false changes caused by seasonal shifts, weather changes, lighting variations, or inconsistencies in image sources. We observed that both BAT and SWDCA failed to detect several building changes in the upper-right area of the first image pair, likely due to tree cover obscuring the buildings in the pre-change image.

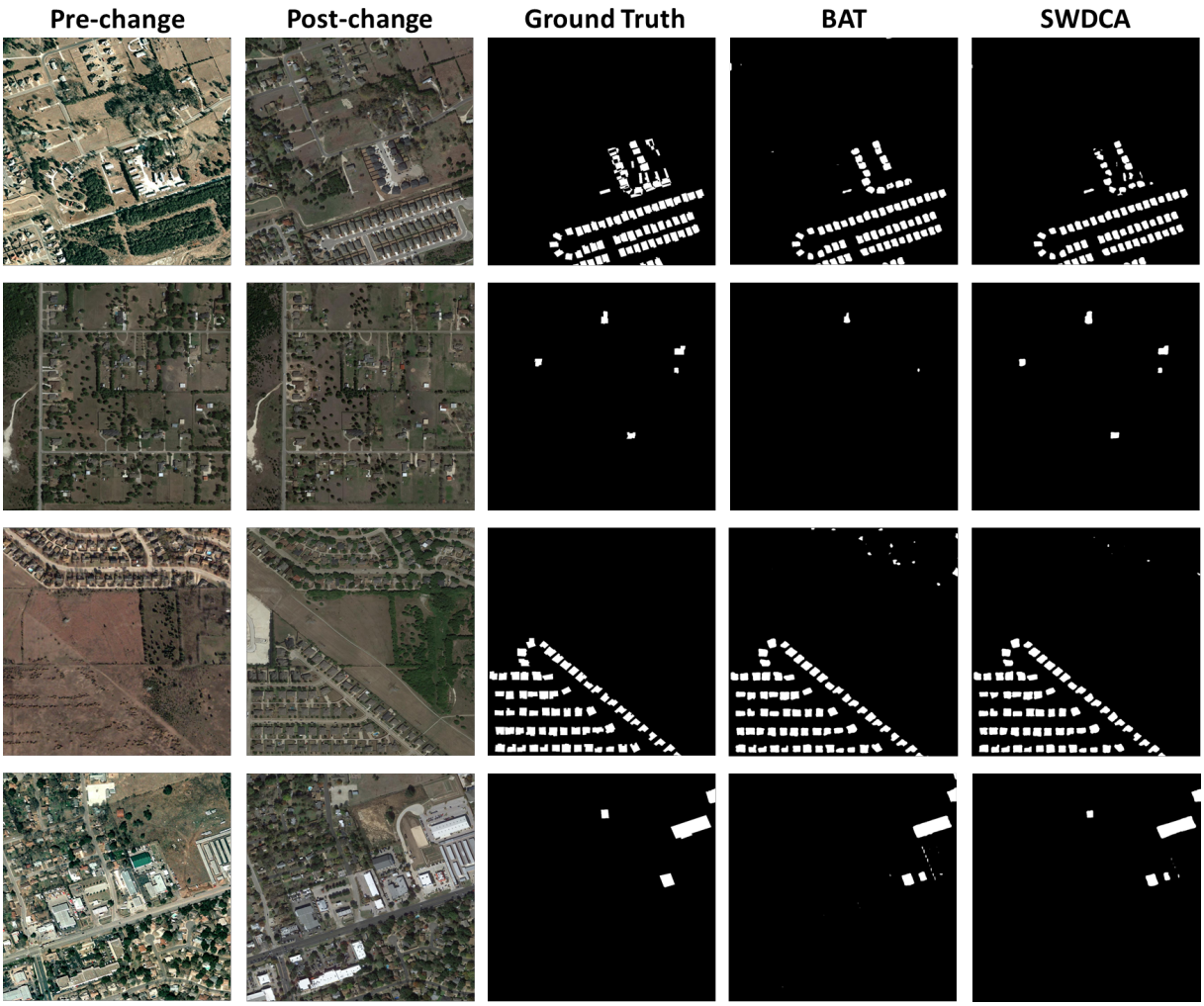

Figure 5. Comparison of building change predictions generated by BAT and the SWDCA network on the LEVIR-CD+ dataset.

4.3. S2Looking Dataset Experimental Results

The S2Looking dataset [24] is both the largest and most challenging BCD dataset available. Unlike the LEVIR-CD+ dataset, which emphasizes urban regions viewed from near-nadir angles, the S2Looking dataset mainly comprises rural regions observed from various large, off-nadir angles. Additionally, it features substantially sparser changes compared to other datasets. The dataset contains 5000 bi-temporal image pairs, divided into 3500 pairs for training, 500 pairs for validation, and 1000 pairs for testing. In accordance with standard practices, we used the training set to train the networks, the validation set to fine tune the networks, and the test set to evaluate and report the results.

As shown in Table 4, the SWDCA network continues to demonstrate a substantial improvement in recall rate on this challenging dataset, resulting in the highest F1 score. While TransUNetCD achieves the highest precision, its recall rates are considerably lower, suggesting that it can only identify evident building alterations. Conversely, our SWDCA network identifies a substantial proportion of building changes, concurrently retaining a precision level that rivals that of leading approaches. Such an advantage arises from the innovative transformer cross-attention mechanism within SWDCA, which detects spatio-temporal semantic discrepancies by explicitly modeling the dissimilarity of bi-temporal tokens.

Table 4. Evaluation of the SWDCA network against various models on the S2Looking dataset.

Method	Precision (%)	Recall (%)	F1 (%)
STANet-BAM(ResNet18) [4]	31.19	52.91	39.24
STANet-PAM(ResNet18) [4]	38.75	56.49	45.97
AMIO-Net [10]	63.94	49.25	53.34
DTCDSCN(SE-Res34) [7]	68.58	49.16	57.27
L-Unet [31]	59.95	58.59	59.26
CDNet [32]	67.48	54.93	60.56
MSCANet [33]	64.63	57.67	60.95
BiT(ResNet18) [22]	72.64	53.85	61.85
Hu et al. [40]	72.53	54.53	62.25
SNUNet [14]	71.94	56.34	63.19
ChangeFormer(MiT-b1) [21]	72.82	56.13	63.39
IFN(VGG-16) [9]	66.46	61.95	64.13
FHD [34]	74.09	56.71	64.25
ChangeStar(MiT-b1) [35]	69.30	59.90	64.30
CGNet [16]	70.18	59.38	64.33
Xu et al. [11]	69.68	61.54	65.36
ChangerEx(ResNet18) [18]	73.59	60.15	66.20
ChangeStar(ResNet18) [35]	70.90	62.20	66.30
CDNeXt [36]	70.78	63.08	66.71
TransUNetCD [23]	**76.41**	59.70	67.03
BAT [27]	70.50	63.99	67.09
SWDCA Network	69.89	**66.01**	**67.90**

The bold entities denote the best performance.

Figure 6 presents a visual comparison of the building change predictions generated by BAT and the SWDCA network. This comparison reveals that the SWDCA network excels in reconstructing the structural details of altered buildings. Furthermore, it consistently outperforms BAT in discerning building alterations from variations in observation angle, illumination, season, and land cover. We observed that both BAT and SWDCA failed to detect two building changes on the right side of the final image pair, likely due to the low contrast in the pre-change image, which poses a big challenge.

Figure 7 shows the building change predictions generated by various methods. The bi-temporal image pair exhibits diverse pseudo-changes arising from lighting, weather, or seasonal vegetation. Among the evaluated methods, BAT and the SWDCA network stand out for their strong adaptability to these varying environmental conditions achieved through the learning of robust representations that are invariant to such variations. Compared to BAT, the SWDCA network shows excellent performance in accurately recovering building boundaries and reducing false detections of changes. However, it failed to detect a building change in the upper half of the image pair due to the building's blurred outline.

Figure 8 presents several failure cases. In the first row, SWDCA incorrectly classifies greenhouses as buildings. This misclassification likely stems from the limited presence of greenhouses in the training set, which reduces the model's ability to distinguish them from buildings. Incorporating height or hyperspectral information may significantly mitigate these errors. In the second and third rows, SWDCA interprets the difference between ground and building foundations and the distinction between an unfinished and a finished building as changes. Building construction and demolition are progressive processes, typically spanning several months. Treating building changes as a binary classification problem and artificially defining buildings as those with completed roofs presents challenges for a model trained with binary cross-entropy, which interprets change as a probability and uses a threshold of 0.5. The greater contextual detail provided by higher spatial resolution imagery, along with more training data, may enable the model to shift its discrimination criterion from structural differences to the presence of completed roofs. In the fourth row,

SWDCA fails to accurately reconstruct the boundaries of small buildings, likely due to the 8× upscaling applied on the final feature map through bilinear interpolation. Reducing the downscaling of the initial feature map, for example, to 1/4, would decrease the extent of upscaling and, thus, preserve more boundary details. However, this approach would also quadratically increase computational complexity. Consequently, a trade-off must be made between prediction detail and efficiency.

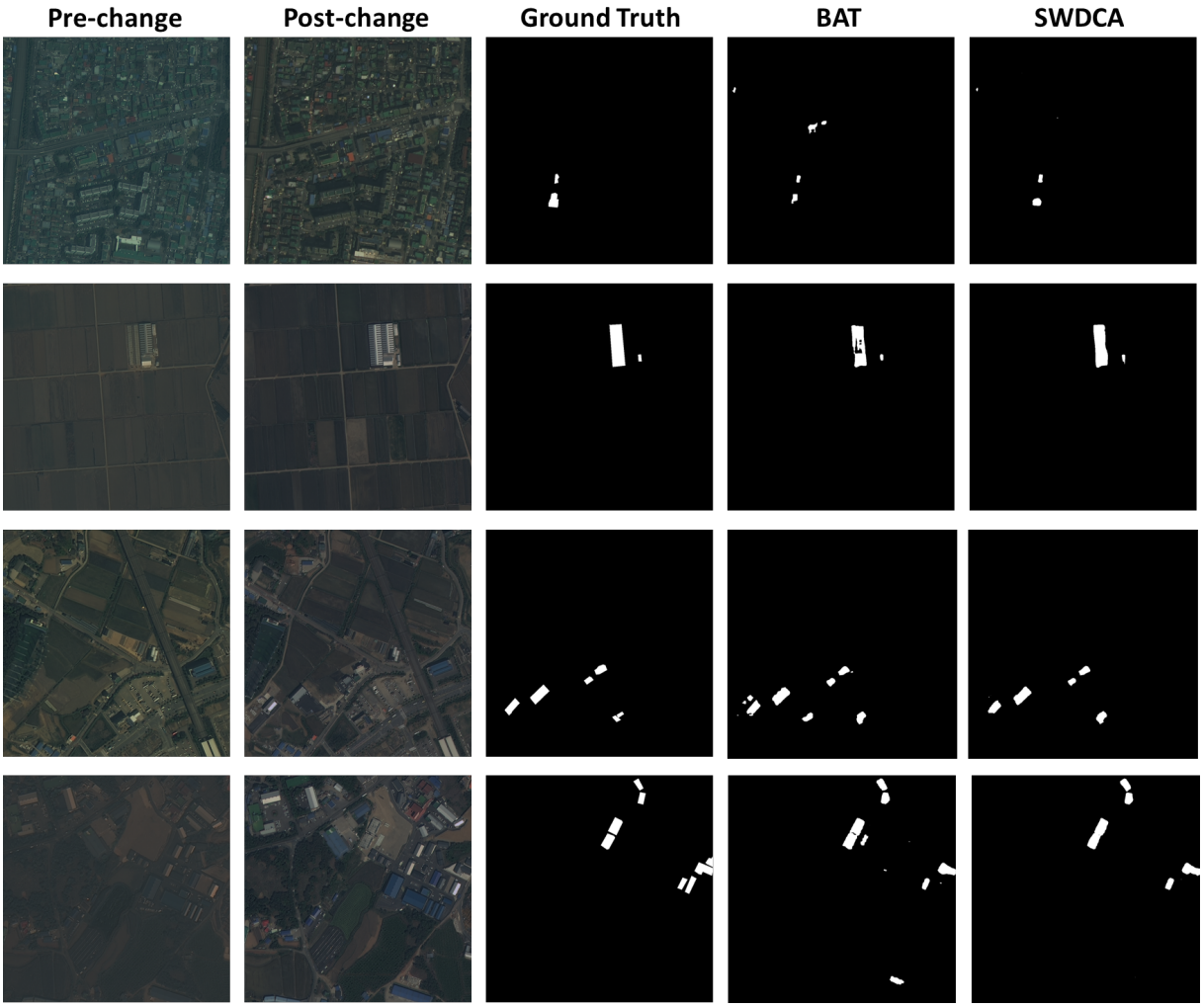

Figure 6. Comparison of building change predictions generated by BAT and the SWDCA network on the S2looking dataset.

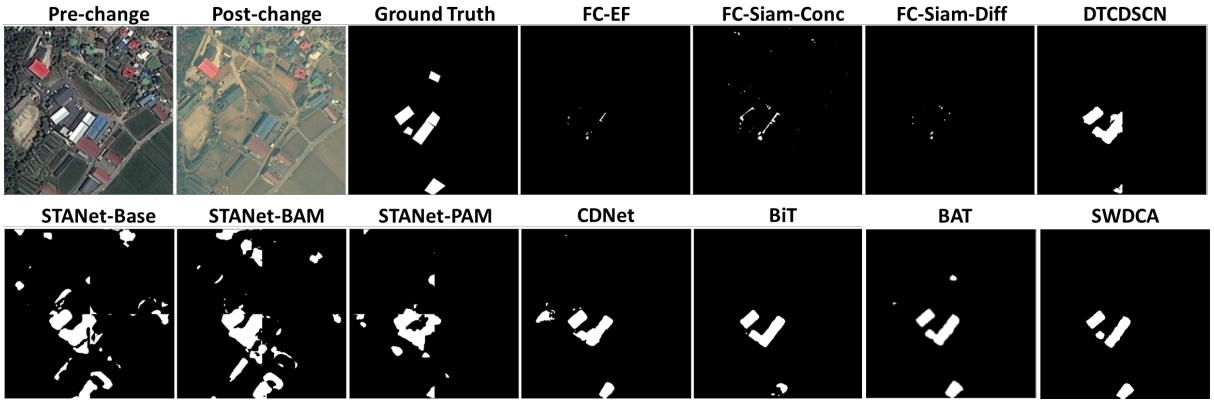

Figure 7. Comparison of building change predictions generated by various methods on the S2looking dataset.

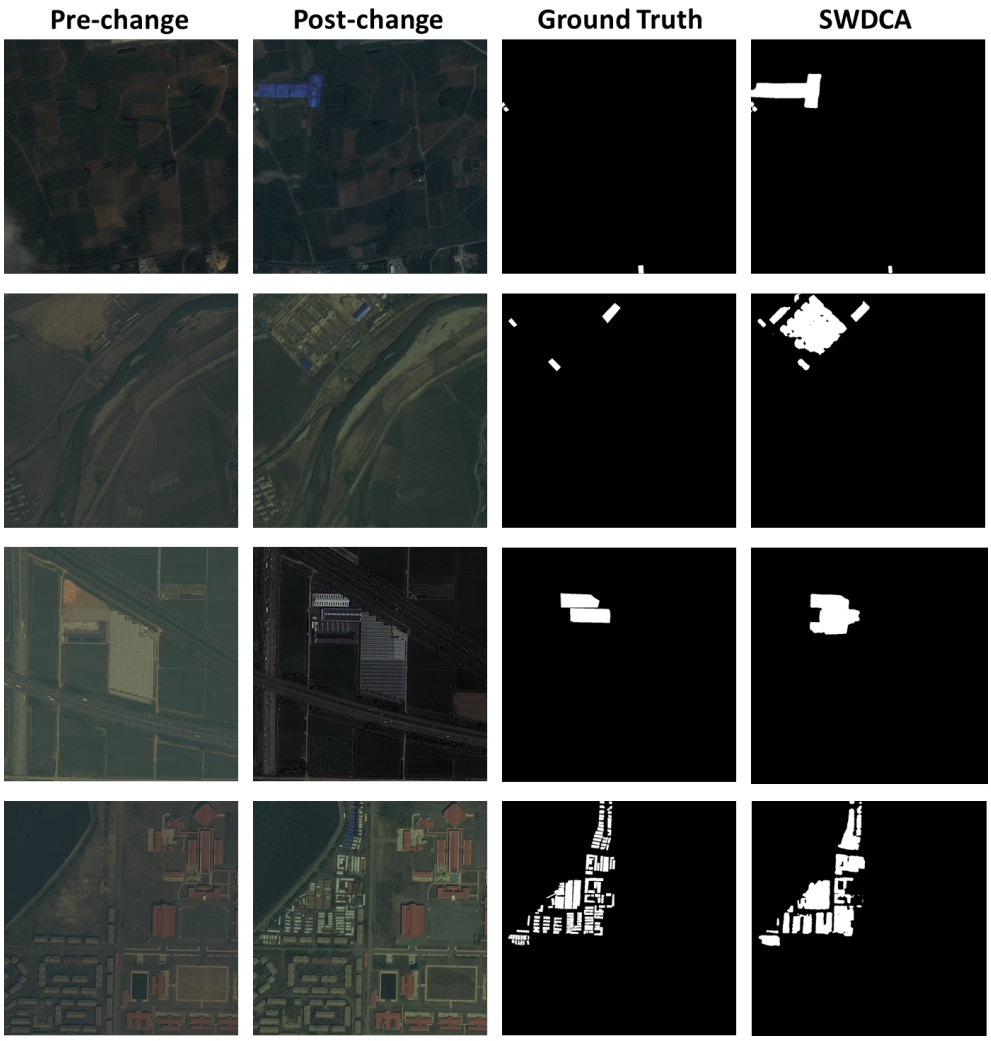

Figure 8. Failure cases on the S2looking dataset.

5. Ablation Analysis

Ablation experiments were conducted on the LEVIR-CD+ dataset to assess the effectiveness of each component within the SWDCA Network.

5.1. Ablation Analysis of Dissimilarity Cross-Attention

The dissimilarity cross-attention within the SWDCA module consolidates the conventional query and key tokens into a single similarity token, which encapsulates feature semantic likeness. It also flips the sign of matrix multiplication before applying the softmax operation, emphasizing dissimilar token pairs and suppressing similar ones, thereby aiming for a more suitable modeling of change detection. To evaluate the efficacy of dissimilarity cross-attention, we constructed a variant network where we replaced it with conventional cross-attention using query, key, and value tokens, without reversing the sign. Therefore, its output at position (i, j), denoted as y_{ij}, can be expressed as follows:

$$CA(q1_{ij}, \widehat{\mathbf{K2}}, \widehat{\mathbf{V2}}) = \sigma(q1_{ij} \times \widehat{\mathbf{K2}}^{\top} / \sqrt{d} + \mathbf{B}) \times \widehat{\mathbf{V2}} \tag{16}$$

$$CA(q2_{ij}, \widehat{\mathbf{K1}}, \widehat{\mathbf{V1}}) = \sigma(q2_{ij} \times \widehat{\mathbf{K1}}^{\top} / \sqrt{d} + \mathbf{B}) \times \widehat{\mathbf{V1}} \tag{17}$$

$$y_{ij} = CA(q1_{ij}, \widehat{\mathbf{K2}}, \widehat{\mathbf{V2}}) \odot CA(q2_{ij}, \widehat{\mathbf{K1}}, \widehat{\mathbf{V1}}) \tag{18}$$

As indicated in Table 5, dissimilarity cross-attention outperforms the conventional cross-attention mechanism by a substantial 2%, resulting in a 1.1% advantage in the overall F1 score. This confirms that dissimilarity cross-attention is more effective than the conventional cross-attention mechanism in detecting changes.

Table 5. Ablation analysis of dissimilarity cross-attention on the LEVIR-CD+ dataset.

Method	Precision (%)	Recall (%)	F1 (%)
Dissimilarity Cross-Attention	**88.17**	**86.68**	**87.42**
Conventional Cross-Attention	88.12	84.61	86.33

The bold entities denote the best performance.

From the perspective of computational complexity, dissimilarity cross-attention generates only two tokens from a feature token, in contrast to conventional cross-attention, which generates three. This results in reduced computational complexity.

5.2. Ablation Analysis of the Dilation Rate

We also investigated the optimal combination of dilation rates within the SWDCA module. As detailed in Table 6, by segmenting the pre-change and post-change feature maps(**F1** and **F2**) into two slices along the channel dimension and employing multi-head dissimilarity cross-attention with dilation rates of 1 and 2, respectively, we achieved higher accuracy in both precision and recall metrics compared to scenarios without segmentation or with one more slice of dilation rate 3. This confirms our hypothesis that a 3 × 3 sliding window with a dilation rate of 2 is adequately sized to encompass the majority of buildings on feature maps downscaled to 1/8 of their original size.

Table 6. Ablation analysis of the dilation rate on the LEVIR-CD+ dataset.

Method	Precision (%)	Recall (%)	F1 (%)
Dilation Rate 1	87.83	85.76	86.78
Dilation Rate 1, 2	**88.17**	**86.68**	**87.42**
Dilation Rate 1, 2, 3	87.57	86.35	86.96

The bold entities denote the best performance.

With respect to computational complexity, the sliding-window size alone determines the complexity of the cross-attention mechanism rather than the dilation rate. Therefore, the variant models presented in Table 6 have the same computational complexity.

It is important to note that the slicing number and dilation rates should be adjusted according to the spatial resolution of the remote sensing images. The optimal slicing and dilation configuration described above is specific to the spatial resolution of the LEVIR-CD+ and S2Looking satellite imagery datasets. When applied to aerial imagery or high-resolution satellite imagery with greater spatial detail, increasing the slicing number and dilation rate would likely yield improved results without added computational complexity.

5.3. Ablation Analysis of the Re-Parameterized 11 × 11 Large Kernel

To minimize computational demands, we utilize only the shallow layers of EfficientNetV2, which inherently possess a restricted receptive field. To overcome this constraint, we integrate three depthwise convolutions with varying kernel sizes (3 × 3, 7 × 7, and 11 × 11) in parallel during training. These convolutions are responsible for capturing small-, medium-, and large-scale features, respectively. The inclusion of larger 7 × 7 and 11 × 11 kernels enhances the network's ability to emulate the receptive field found in deeper networks. During inference, we optimize the smaller 3 × 3 and 7 × 7 kernels by consolidating them into a larger 11 × 11 kernel through batch normalization layer fusion and kernel

parameter combination. As a result, the computational complexity of the original three parallel branches is reduced to that of a single 11 × 11 kernel. Table 7 demonstrates that the re-parameterized 11 × 11 large kernel enhances precision by an impressive 2% at a minimal cost of 2.7 FPS, highlighting its effective feature representation capability.

Table 7. Ablation analysis of the re-parameterized 11 × 11 Large kernel.

Method	Precision (%)	Recall (%)	F1 (%)	Speed (FPS)
With Large Kernel	**88.17**	**86.68**	**87.42**	33.2
Without Large Kernel	86.21	86.59	86.40	**35.9**

The bold entities denote the best performance.

6. Discussion

A near-real-time change detection network allows for the rapid identification of changes in building structures or environments, enabling timely response and intervention where necessary. It also provides continuous monitoring capabilities to detect changes such as construction progress, structural damage, or unauthorized alterations, facilitating informed decision making for urban planning, disaster response, and insurance assessments based on up-to-date information. BCD tasks often encounter diverse environmental conditions, such as variations in lighting, weather, or seasonal vegetation. When dealing with aligned pre-change and post-change images, a fundamental challenge is effectively modeling the spatio-temporal semantic relationships between them. Recognizing the critical importance of discerning disparities between different time points, we propose an innovative transformer cross-attention mechanism. SWDCA detects spatio-temporal semantic discrepancies by explicitly modeling the dissimilarities among bi-temporal tokens, diverging from the mono-temporal similarity attention typically used in conventional transformers. To meet the requirement for near-real-time processing, SWDCA employs a sliding-window scheme to constrain the cross-attention mechanism within a predetermined window or dilated window size. This approach not only excludes distant and irrelevant information but also reduces computational costs. Thorough experiments involving quantitative evaluations and visual analyses confirm that our method can adapt to diverse environmental conditions by learning robust representations that are invariant to these variations. This enhances the model's ability to generalize effectively across different scenarios.

Satellite onboard processing for building change detection offers advantages such as reduced data transfer needs, enhanced privacy and security, autonomous operation, improved responsiveness, and optimized resource allocation. For instance, onboard processing can optimize resource allocation by prioritizing data transmission based on the importance or relevance of detected changes or features. This helps in efficiently utilizing satellite resources and maximizing the utility of satellite missions. To achieve near-real-time inference speed on an embedded system or mobile GPU, we developed a lightweight Siamese backbone for the extraction of building and environmental features. We applied techniques such as batch normalization layer fusion and kernel parameter combination to restructure parallel structures in a series structure, thereby enhancing inference speed.

Although the advantages mentioned earlier are notable, we acknowledge a limitation: our cross-attention mechanism is restricted to analyzing pairs of images captured at two different times and cannot handle the complex relationships within a set of images taken over multiple time periods. Since building construction and demolition typically span several months, treating multi-temporal change detection as a binary or multiclass classification task is insufficient to accurately represent the process. To address this, we

attempt to approach the problem as an ordinal regression task and introduce a novel loss function tailored to a new multi-temporal cross-attention mechanism.

7. Conclusions

In this study, we introduced an innovative transformer cross-attention mechanism aimed at effectively capturing spatial and temporal correlations across aligned pairs of images taken at different times. Additionally, we developed a lightweight Siamese backbone designed for the extraction of features related to buildings and the environment. Integrating the SWDCA module into this backbone enabled the creation of an efficient change detection network that achieved top-tier accuracy on two building change detection datasets. Moreover, when processing bi-temporal image pairs at a resolution of 512 × 512 pixels on a mobile GPU, it achieved a real-time inference speed of 33.2 frames per second. While this study focused on building change detection, our future research will explore applying, adapting, and enhancing the transformer cross-attention mechanism to address change detection tasks in agriculture and climate change domains.

Author Contributions: Conceptualization, W.L. and M.N.; Data curation, W.L.; Formal analysis, W.L.; Funding acquisition, M.N.; Investigation, W.L.; Methodology, W.L.; Project administration, M.N.; Resources, M.N.; Software, W.L.; Supervision, M.N.; Validation, M.N.; Visualization, W.L.; Writing—original draft, W.L.; Writing—review and editing, M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. 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.
2. Weber, E.; Kané, H. Building disaster damage assessment in satellite imagery with multi-temporal fusion. *arXiv* **2020**, arXiv:2004.05525.
3. Gupta, R.; Shah, M. Rescuenet: Joint building segmentation and damage assessment from satellite imagery. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4405–4411.
4. 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]
5. Wang, D.; Chen, X.; Jiang, M.; Du, S.; Xu, B.; Wang, J. ADS-Net: An Attention-Based deeply supervised network for remote sensing image change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *101*, 102348.
6. 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*, 4704817 . [CrossRef]
7. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 811–815. [CrossRef]
8. 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—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2115–2118. [CrossRef]
9. 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]
10. Gao, W.; Sun, Y.; Han, X.; Zhang, Y.; Zhang, L.; Hu, Y. AMIO-Net: An Attention-Based Multiscale Input–Output Network for Building Change Detection in High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2079–2093. [CrossRef]

11. Xu, C.; Ye, Z.; Mei, L.; Yang, W.; Hou, Y.; Shen, S.; Ouyang, W.; Ye, Z. Progressive Context-Aware Aggregation Network Combining Multi-Scale and Multi-Level Dense Reconstruction for Building Change Detection. *Remote Sens.* **2023**, *15*, 1958. [CrossRef]
12. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7296–7307. [CrossRef]
13. 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]
14. 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*, 8007805. [CrossRef]
15. 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*, 5602812. [CrossRef]
16. Han, C.; Wu, C.; Guo, H.; Hu, M.; Li, J.; Chen, H. Change Guiding Network: Incorporating Change Prior to Guide Change Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8395–8407. [CrossRef]
17. Shen, Y.; Zhu, S.; Yang, T.; Chen, C.; Pan, D.; Chen, J.; Xiao, L.; Du, Q. Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5402114. [CrossRef]
18. Fang, S.; Li, K.; Li, Z. Changer: Feature interaction is what you need for change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610111. [CrossRef]
19. Chen, P.; Zhang, B.; Hong, D.; Chen, Z.; Yang, X.; Li, B. FCCDN: Feature constraint network for VHR image change detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 101–119. [CrossRef]
20. Hou, S.; Zhang, G.; Cui, H.; Li, X.; Chen, Y.; Li, H.; Wang, H.; Ma, X. Stable Prototype Guided Single-Temporal Supervised Learning for Change Detection and Extraction of Building. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4406622. [CrossRef]
21. 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; pp. 207–210.
22. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
23. 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*, 5622519. [CrossRef]
24. 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]
25. Zhang, K.; Zhao, X.; Zhang, F.; Ding, L.; Sun, J.; Bruzzone, L. Relation Changes Matter: Cross-Temporal Difference Transformer for Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5611615. [CrossRef]
26. 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*, 5620014. [CrossRef]
27. Lu, W.; Wei, L.; Nguyen, M. Bitemporal Attention Transformer for Building Change Detection and Building Damage Assessment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4917–4935. [CrossRef]
28. Ding, L.; Zhang, J.; Zhang, K.; Guo, H.; Liu, B.; Bruzzone, L. Joint Spatio-Temporal Modeling for Semantic Change Detection in Remote Sensing Images. *arXiv* **2022**, arXiv:2212.05245. [CrossRef]
29. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 July 2022; pp. 12124–12134.
30. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10096–10106.
31. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7651–7668. [CrossRef]
32. Chen, H.; Li, W.; Shi, Z. Adversarial Instance Augmentation for Building Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603216. [CrossRef]
33. 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]
34. Pei, G.; Zhang, L. Feature Hierarchical Differentiation for Remote Sensing Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6514105. [CrossRef]
35. Zheng, Z.; Tian, S.; Ma, A.; Zhang, L.; Zhong, Y. Scalable Multi-Temporal Remote Sensing Change Data Generation via Simulating Stochastic Change Process. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 21818–21827.

36. Wei, J.; Sun, K.; Li, W.; Li, W.; Gao, S.; Miao, S.; Zhou, Q.; Liu, J. Robust change detection for remote sensing images based on temporospatial interactive attention module. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103767. [CrossRef]
37. Liu, M.; Shi, Q. DSAMNET: A deeply supervised attention metric based network for change detection of high-resolution images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 6159–6162.
38. Li, Z.; Tang, C.; Wang, L.; Zomaya, A.Y. Remote sensing change detection via temporal feature interaction and guided refinement. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5628711. [CrossRef]
39. 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]
40. Hu, R.; Pei, G.; Peng, P.; Chen, T.; Yao, Y. Feature Difference Enhancement Fusion for Remote Sensing Image Change Detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: Cham, Switzerland, 2022; pp. 510–523.
41. Li, Z.; Tang, C.; Li, X.; Xie, W.; Sun, K.; Zhu, X. Towards Accurate and Reliable Change Detection of Remote Sensing Images via Knowledge Review and Online Uncertainty Estimation. *arXiv* **2023**, arXiv:2305.19513.