International Journal of
Geo-Information

MDPI

Article

Hybrid-TransCD: A Hybrid Transformer Remote Sensing Image Change Detection Network via Token Aggregation

Qingtian Ke and Peng Zhang *

School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China; keqt3@mail2.sysu.edu.cn
* Correspondence: zhangpeng5@mail.sysu.edu.cn

Abstract: Existing optical remote sensing image change detection (CD) methods aim to learn an appropriate discriminate decision by analyzing the feature information of bitemporal images obtained at the same place. However, the complex scenes in high-resolution (HR) remote images cause unsatisfied results, especially for some irregular and occluded objects. Although recent self-attention-driven change detection models with CNN achieve promising effects, the computational and consumed parameters costs emerge as an impassable gap for HR images. In this paper, we utilize a transformer structure replacing self-attention to learn stronger feature representations per image. In addition, concurrent vision transformer models only consider tokenizing single-dimensional image tokens, thus failing to build multi-scale long-range interactions among features. Here, we propose a hybrid multi-scale transformer module for HR remote images change detection, which fully models representation attentions at hybrid scales of each image via a fine-grained self-attention mechanism. The key idea of the hybrid transformer structure is to establish heterogeneous semantic tokens containing multiple receptive fields, thus simultaneously preserving large object and fine-grained features. For building relationships between features without embedding with token sequences from the Siamese tokenizer, we also introduced a hybrid difference transformer decoder (HDTD) layer to further strengthen multi-scale global dependencies of high-level features. Compared to capturing single-stream tokens, our HDTD layer directly focuses representing differential features without increasing exponential computational cost. Finally, we propose a cascade feature decoder (CFD) for aggregating different-dimensional upsampling features by establishing difference skip-connections. To evaluate the effectiveness of the proposed method, experiments on two HR remote sensing CD datasets are conducted. Compared to state-of-the-art methods, our Hybrid-TransCD achieved superior performance on both datasets (i.e., LEVIR-CD, SYSU-CD) with improvements of 0.75% and 1.98%, respectively.

Keywords: change detection; deep learning; transformer; self-attention

check for updates

Citation: 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. https://doi.org/10.3390/ijgi11040263

Academic Editor: Wolfgang Kainz

Received: 23 February 2022
Accepted: 15 April 2022
Published: 17 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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

Change detection (CD) belongs to the important field of intelligent interpretation of remote sensing images, which is aimed to identify the difference of the objects or scenes between multi-temporal sequence images, playing an important role in land cover monitoring, urban/mining land resource management, natural disaster assessment, and so on [1]. The purpose of CD is to obtain a pixel-level change map by analyzing the registered bitemporal remote sensing image, where each pixel is assigned a probability representing changed and unchanged.

Concurrent change detection methods are mainly divided into pixel-level, semantic-level, and feature-level CD. Pixel-level CD infers the classification prediction pixel by pixel, in which the unchanged pixels are represented as 0, and the changed pixels are represented as 1 [2,3]. Although these methods achieved competitive results by adopting some models to solve the semantic segmentation task, this inevitably caused the complete loss of spatiotemporal relationship of among bitemporal images. For semantic-level methods, the difference of object entities in the scene are compared to obtain the discriminative information between bitemporal images, thus the salient regions are ignored including both large-scale and small-scale objects. The feature-level CD methods usually introduce learning-based models for representing image features, which were subsequently used as yardsticks for classification [4,5], but most network structures severely rely on the feature representing ability of different layers of CNN, thus completely ignoring the effect by modeling context relationship in the feature extraction stage. Most modern CD works applied deep convolutional neural networks (DCNN) to obtain a pixel-level change map, but many of them only utilized effective CNN backbone (i.e., VGG [6], ResNet [7]) to represent bitemporal features pairs, which are limited to the network structure. Although abundant methods adopted attention mechanisms to enhance the global context of features, the computational efficiency faces a significant drop. Therefore, our work introduces a supervised model for binary change detection by introducing a feature-level network integrating transformer.

Although many learning-based methods show a good performance on public HR datasets, there remain two limitations: (1) Under complex scenes, the object's appearance in bitemporal images is not consistent; (2) the measure of difference is difficult to learn by facial spectral behaviors. As in Figure 1, the objects inside the red dashed box have an intrinsic offset in color and visual angle, leading to inconsistent distinguish features. In the meantime, due to the differences caused by seasonal changes, extensive irrelevant local changes representations emerge. Under these conditions, most CNN-based methods find it difficult to perfectly locate the change regions of interest. Although many methods tried to address the problem by introducing different kinds of attention mechanism, almost all consume heavy computational and memory costs. The visual transformer inherited from natural language translation improved the computational efficiency of modeling global context in many visual tasks. Motivated by the vision transformer (ViT [8]), we introduce a transformer-based layer replacing self-attention to build long-distance dependencies of each temporal feature set. Attaching with the transformer encoder, multi-head attention is engaged to construct patch-to-patch interactions. However, prior transformer modules only adopted fixed embedded image sequences, ignoring the multi-scale image token representations, which means that the receptive field within the transformer layer is limited to regular scene objects, leading to a weak field for complex scenes that shows objects varying with diverse sizes. Further, such incomplete feature granularity among each temporal image inevitably causes irrelevant changes at different scales. In this work, we propose a hybrid transformer structure via token aggregation strategy for HR remote-image change detection, which creatively represents feature correlations in the manner of coarse-grained and fine-grained within one transformer layer. Specifically, the multi-head attentions in our hybrid transformer are split into several groups, each of which contains attention information with different specific granularity. For the fine-grained groups, a small amount of tokens containing more detailed local information are aggregated. For most of the remaining coarse-grained groups, the corresponding attention heads show the ability to selectively capture large objects by aggregating large-sized tokens.

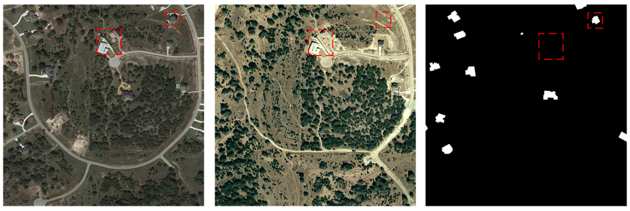

Figure 1. Changes of unrelated attributes in objects between bitemporal high-resolution remote sensing images under complex scenes.

The early late-fusion-based CD methods [9–11] applied either channel-wise concatenation or spatial-level difference operations into feature pair fusion, thus producing high-level semantic feature maps. In our proposed baseline, two manners, named Late-Diff (LD) and Early-Diff (ED), are proposed for producing distinguished features augmented by self-attention: (1) Dual-pair sequential tokens from a hybrid-transformer encoder (H-TE) are followed, re-projected by a Siamese hybrid-transformer decoder (H-TD), and the generated token pair is performed with an absolute difference operation; (2) instead of separately reprojected per temporally encoded tokens from previous H-TE with H-TD, the ED method differentiates between the enhanced token pair in the earlier stage for directly obtaining discriminative difference features to be reprojected. Compared with concurrent attention-based change detection methods, our model fully captures global long-range dependencies in a hybrid manner, making the high-level semantic features from CNN contain richer contextual representations.

To fully restore feature resolution in the decoder, prior methods [12] focused on adding skip-connections from shallow-layer features to high-level layers, thus obtaining features containing detailed texture and highly representative semantic information. U-Net++ [13] adopted a multiple side fusion strategy for generating dense multi-level semantic change maps, but nevertheless occupied a high computation complexity. Ref [14] proposed a scale-selection module to adaptively aggregate the final maps from different levels of features. However, extensive change details are neglected along with the progressive upsampling stages in the decoder, so a cascade feature decoder was proposed to mitigate the absence of various scale representations.

On the whole, in this work, we introduce an improved bitemporal image transformer network to model long-range context within the bitemporal image in a multi-scale manner. The key is that high-level representations of related changes could be represented into serialized visual words. Our contributions can be summarized as follows:

(1) We proposed a transformer-based change detection network (Hybrid-TransCD), which fully alleviates the extremely complex computations caused by early self-attention. The long-range interactions captured by transformer module promoted strengthened feature representations.

(2) A hybrid-transformer encoder (H-TE) and a hybrid-transformer decoder (H-TD) were designed to produce stronger difference tokens (features), both of which captured multi-scale tokens' context within one self-attention block via token aggregation. The proposed transformer structure merges hierarchical tokens among large-scale regions and small-scale objects while maintaining lightweight computational and memory costs.

(3) By building relations between encoded token sequences and the original residual token, we proposed two manners for representing discriminative features between bitemporal images, both of which captured promising difference context containing multi-granularity information. Compared with the traditional CNN that fuses features in the final stage to generate differential features, our designed structures improve the transformer layer so that the discriminative features can be directly obtained when modeling spatiotemporal context.

(4) To utilize richer coarse-grained and fine-grained features among high and shallow layers, a cascade feature decoder was introduced to achieve dense change prediction.

(5) The abundant experiments demonstrate that the proposed approach performs better than other attention-based and learning-based change detection methods concerning F_1 and parameters cost. In particular, we achieve an improvement of 0.75 and 1.98 points on the LEVIR-CD and SYSU-CD datasets, respectively, compared to the state-of-the-art methods.

2. Related Work

2.1. Attention-Based Methods

Recently, global context modeling and long-range dependencies grasping have attracted increasing attention in remote sensing image change detection, and many attention-based mechanisms, including channel-attention, spatial-attention, and self-attention, are being gradually applied to multi-temporal spatiotemporal correlation modeling [15–17]. However, these methods only establish the long-range context of each temporal image separately, or directly update the initially fused image by reweighting in spatial and channel spaces. Some works [18,19] achieved promising performance by dense no-local operations to construct pixel-to-pixel semantic difference correlation between bitemporal images, but the majority have loaded computational/memory cost, causing an inefficient learning process for HR remote sensing images. Zhang [20] pointed out that the current deep-learning-based change detection methods have certain limitations in deep feature fusion and supervision, so they proposed a deep supervised image fusion network with a dual-branch architecture, improving the ability to distinguish differences by inserting a spatial attention module and a channel attention module into multi-level feature layers. Raza [21] proposed ultimate fusion strategies based on spatial/channel attention by repeating multiple times, thus acting to refine the multi-scale features.

2.2. Vision Transformer

Motivated by ViT, BiT [22] firstly proposed a bitemporal image transformer network for effectively modeling spatial–temporal contexts, which innovatively proved the enhancing ability by combining a CNN and a transformer. TransCD [23] considered the limitation of local receptive field of traditional CNN networks, so they incorporated a Siamese vision transformer (SViT) in a feature difference SCD framework to solve the scene change detection task. However, these mentioned transformer-based CD frameworks are merely capable of capturing global interdependencies of single-scale objects within each transformer layer, which tend to lose robustness in rich spatial scenes of remote sensing images.

Recent versatile vision transformer models such as PVT [24] and swin transformer [25] proposed effective solutions to heavy computational cost existing in a pure transformer-based network: the former considered representing high-resolution features by replacing coarse-grained image patches with fine-grained ones, and the token lengths were subsequently reduced, adopting a progressive pyramid strategy. The latter introduced a hierarchical structure to reduce the heavy computational cost existing in token-level self-attention, and the shifted windows method facilitates interaction between adjacent patch groups. Although both of them alleviated the computational and memory costs caused by large-resolution feature maps, the transformer encoder blocks either merely modeled local context within the narrow region or captured mixed fine-grained information between objects and irrelevant backgrounds. To solve the above limitations, we introduce a hybrid transformer structure for maintaining multi-granularity global dependencies among projected token pairs, thus acquiring difference representations between bitemporal images for both large and small objects. The effect comparisons are shown in Figure 2.

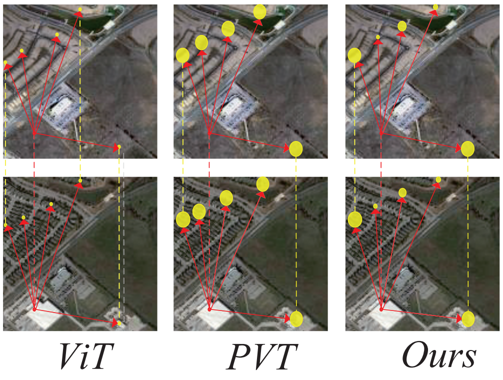

Figure 2. Comparison of recent attention mechanisms in change detection. The areas connected by the yellow dotted line represent the difference in token information between the multi-temporal image regions, where the number of yellow circles reveals the number of computations required for self-attention while the size of the circle represents the receptive field where the token is located.

Other works [26] built multi-scale feature interactions by constructing progressive cross-scale self-attention within one layer, or combined the patch-wise attention to constructing hierarchical reception field (RF) [27]. They represent multi-granularity features at the cost of increasing redundant branches for large-scale patches, causing yet imperfect computational efficiency. In this work, we design a hybrid vision change detection transformer module for enhancing not only large change regions (e.g., building) but also elaborate small objects (e.g., car). Significantly, we propose two transformer-based modules (transformer encoder and difference transformer decoder) to model the semantic context of bitemporal tokens and pixel-level difference tokens respectively, both of which are composed with the various number (N and M) of hybrid transformer blocks.

3. Materials and Methods

3.1. Network Overview

Similar to most late-fusion methods, the proposed method constructs a discriminative difference feature map in the highest layer, which means the extracted features pair from CNN backbone are enhanced by the hybrid transformer first. Then, the produced difference feature maps are forwarded into the decoder to restore context change representation with initial size. Significantly different from the general CD pipeline, which treats fused features of bitemporal images from the highest layer as change semantic representations, here, we introduced transformer structure into the feature fusion stage to obtain pixel-level discriminative features with the compact tokens pair.

The overall network flow is shown in Figure 3. A hybrid transformer module is incorporated into the general CNN-based pipeline to leverage an elaborate bitemporal feature pair extracted by the Siamese backbone expressed as $f^i \in R^{C\times H\times W}, i = 1, 2$, and global context enhanced by the transformer, thus generating an encoded token pair denoted as $T^i \in R^{(P\times C)\times D}, i = 1, 2$, where P represents the embed patch number, while the D represents the predefined parameter of token hidden dimensions to be projected.

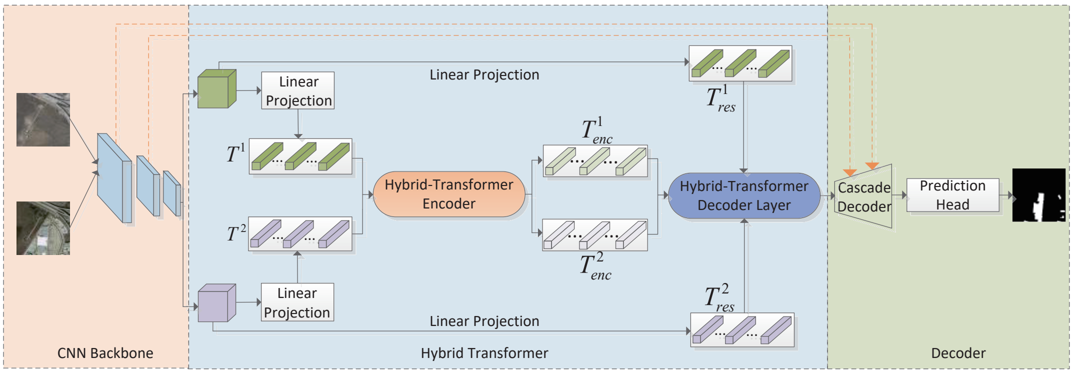

Figure 3. Overall architecture of our Hybrid-TransCD.

Specifically, regarding each temporal feature, a hybrid-transformer encoder (H-TE) is employed for building coarse-grained and fine-grained patch embeddings. The generated semantic tokens pair $T^i_{enc} \in R^{P\times D_L}, i = 1, 2$ together with corresponding residual patch embeddings $T^i_{res} \in R^{P\times D_L}; i = 1, 2$ are forwarded into the hybrid-transformer decoder (H-DE) layer to leverage dependencies between encoded semantic tokens and original pixel-level features, the generated difference tokens $T_{diff} \in R^{P\times D}$ are subsequently restored into a feature-dimensional tensor represented as $f_{diff} \in H \times W \times D$ by performing permutation and reshaping operations, where D is the predefined number of hidden channels. As the absolute difference is first taken from the encoded token pair and then decoded (early difference), or the token pair is decoded first and then made (late difference), the produced features contain abundant semantic change information. Subsequently, the fused features accompanied by skip-connections from CNN backbone are upsampled to restore original resolution by the proposed cascade feature decoder. The prediction head composed of a 1 × 1 convolution is employed for generating a predicted change probability map $P \in R^{2\times H\times W}$.

Significantly, the hybrid ResNet [7], rather than a pure transformer extractor, is employed for leveraging CNN-transformer strength. The Siamese H-TE is constructed by a multiple hybrid encoder transformer block (M) while the H-TD is composed of N hybrid decoder transformer blocks, thus stabilizing appropriate computation efficiency.

3.2. Hybrid-Transformer Encoder

As the major component of ViT, the transformer encoder module is used to extract image features. Specifically, the original two-dimensional image is converted into a one-dimensional embedding sequence, that is, the input $X \in H \times W \times C$ is divided into block sequences of size $P \times P$, and the sequence length is HW/P^2. At the same time, position embedding is added to encode the position information of tokens, avoiding the model, to learn absolute position information with the semantics of image patches. As can be seen from Figure 4, the transformer encoder contains a multi-head attention (MHA), two normalization layers (Norm) and a multi-layer perceptron layer (MLP), performing with the scaled dot product attention, as shown in Figure 5. The query, key, and value are produced by 1×1 convolution, where key and value are paired. According to self-attention, the inner product is calculated by matching k key vectors ($K \in R^{k \times d}$) with query vector ($Q \in R^d$), which is then normalized by $Softmax$. For multi-head attention MHA (Figure 6), h attention heads act on the input sequence, respectively, and in practice, the image sequence patches are divided into h subsequences with the size of $N \times d$, and the outputs from h different attention heads are concatenated together. Finally, a linear transformation is performed to obtain the ultimate output, which is expressed as

$$MSA(X) = Concat(head_1, \ldots, head_h)W^0) \tag{1}$$

where,

$$head_i = Attention(XW_i^Q, XW_i^K, XW_i^V) \tag{2}$$

For each MSA, a feedforward network (FFN) is followed for nonlinear mapping.

Given the high-level feature $f^i \in R^{C \times H \times W}, i = 1,2$ represented by CNN, the tokenization operation is firstly performed to obtain two-dimensional patch sequences denoted as $x_i^j \in R^{P \times C} | i = 1,2; j = 1, \ldots, N$, where the patch size is $P \times P$, and $N = \frac{HW}{P}$ means the length of patches.

Then, the serialized patches are embedded into latent high-dimensional space (D) using learnable linear projections. Specifically, we take a convolution with a $P \times P$ kernel size and the stride of P. To learn patch position information, we added trainable position embedding as follows:

$$T^i = [x^1E; x^2E; \ldots; x^NE] + E^{pos}, i = 1,2 \tag{3}$$

where $E \in R^{(P^2 \cdot C) \times D}$ is the projection weight of patch embeddings, and $E^{pos} \in R^{N \times D}$ encodes the space–time dimensional absolute positions of tokens.

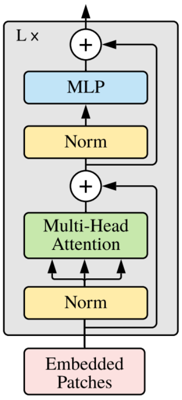

Figure 4. Overall architecture of the ViT encoder.

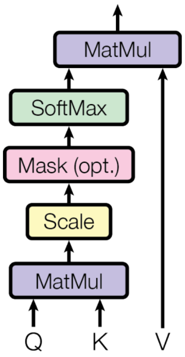

Figure 5. The illustration of scaled dot product attention.

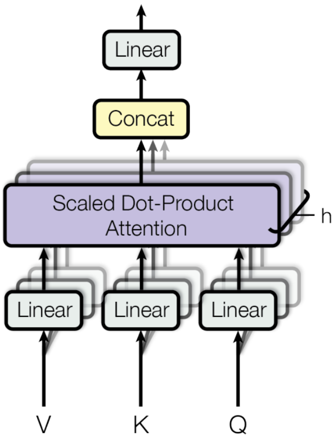

Figure 6. The structure of multi-head attention.

The linearly projected tokens pair is fed into the Siamese encoder containing multiple hybrid transformer encoder blocks for producing richer semantic context among per-temporal image. For each hybrid transformer encoder block, the forwarded input is normalized by the pre-norm residual unit (PreNorm) [8] at first. Then, the output sequences are projected into query (Q), key (K), and value (V). Next, an improved hybrid multi-head self-attention (H-MSA) operation is adopted to parallelly compute hybrid self-attention (Figure 7). Different from the swin [25] that splits broader Q, K, V into multiple small regions, we follow the SRA layer in PVT [24] to adopt a hybrid spatial–reduction attention strategy, to mitigate the computational cost and capture multi-granularity semantic information in the meantime. The comparisons between different self-attention manners are depicted in Figure 8. Instead of applying self-attention globally on final downsampled feature maps or local self-attention on large-scale feature maps within divided small regions, our hybrid self-attention employs token aggregation among multiple key–value pairs, where each key–value pair is generated by downsampling to different sizes. Specifically, at each block b, the K and V from different heads are represented as

$$\begin{aligned} Q_i &= T^{(b-1)}W_i^Q, \\ K_i,V_i &= MTA(T^{(b-1)},r_i)W_i^K, MTA(T^{(b-1)},r_i)W_i^V, \\ V_i &= V_i + LA(V_i) \end{aligned} \tag{4}$$

where W_i^Q, W_i^K, and W_i^V are learnable linear projection weights for the previous output $T^{(b-1)}$ in the i-th head. The $MTA(.;r_i)$ perform a multi-scale token aggregation operation with the downsampling rate of r_i in the i-th head. Here, a convolution layer with kernel size 1×1 and stride of r_i is implemented. In actuality, various r_i within one layer among multiple heads brings multi-scale self-attention computed by K and V. $LA(\cdot)$ is the local augment stage of MTA, implemented by a depthwise convolution for V. Compared with the SRA module in PVT, the transformer encoder learns complementary fine-grained and coarse spatiotemporal spectrum information. The *i*-th attention head is calculated by

$$head_i = \sigma(\frac{Q_i K_i^T}{\sqrt[2]{D_h}})V_i \tag{5}$$

where D_h represents projected channel dimension and σ is the softmax function. The multi-head self-attention (MSA) then perform a concatenation operation to fuse representation information from different dimension spaces. Specifically,

$$\begin{aligned} T_{enc}^{(b-1)} &= MSA(LN(Concat(head_1, \ldots, head_h)W)) + T_{enc}^{(b-1)}, \\ T_{enc}^{(b-1)} &= MLP(LN(T_{enc}^{(b-1)})) + T_{enc}^{(b-1)} \end{aligned} \tag{6}$$

where $W \in R^{hD_h \times C}$ is the linear projection weight matrix while h denotes the head number. The MLP is the feedforward layer to reproject tokens normalized by layer norm (LN); here, we give an improved detail-enhancement (DE) feedforward layer to complement local representations specified for details. As in Figure 9, compared with traditional ViT and PVT, we add a DE layer between two fully connected layers, thus preserving fine-grained local details, where $DWConv$ and $GELU$ are depthwise separable convolution and nonlinear activation functions. The formula is

$$MLP(T^{(b-1)}) = \sigma(T^{(b-1)}W_1 + DE(T^{(b-1)}W_1))W_2 \tag{7}$$

where W_1 and W_2 are learnable weight parameters of FC.

From above, our hybrid transformer block is capable of capturing objects of different scales. By controlling the downsampling rate r, we can achieve the available performance at the cost of efficient computational costs. Specifically, the larger r is, the more short tokens (K, V) are merged, thus producing richer semantic tokens for large regions in a lightweight manner. On the contrary, the smaller r preserves more local details for small objects. The integration of multiple r within one attention block learns multi-granularity features. In our work, we construct Siamese H-TE with different numbers of hybrid transformer blocks, generating encoded semantic tokens $T_{enc}^i, i = 1, 2$ among bitemporal images.

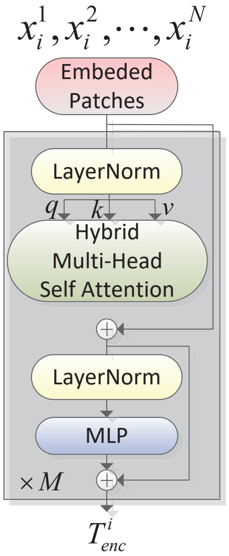
$x_i^1, x_i^2, \cdots, x_i^N$
Embeded Patches
LayerNorm
q
k
v
Hybrid Multi-Head Self Attention
LayerNorm
MLP
$\times M$
T_{enc}^i

Figure 7. The architecture of the hybrid-transformer encoder block. The embed patches of each temporal feature are encoded by our improved hybrid multi-head self attention and MLP.

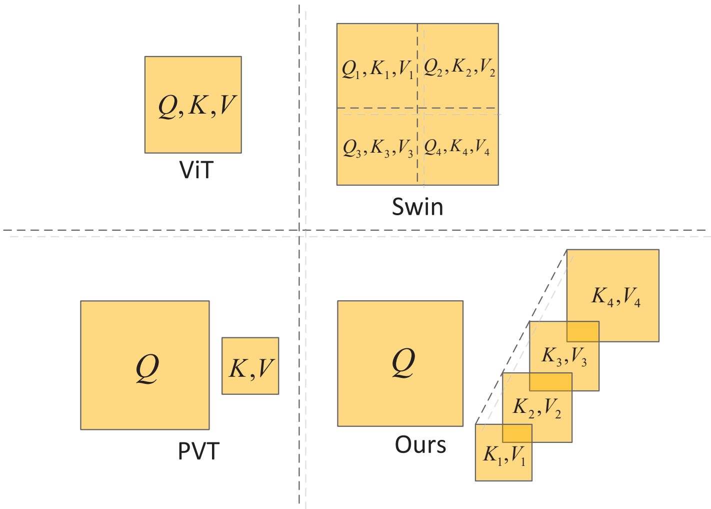

Figure 8. Comparisons of self-attention between different transformer-based models. The Q, K, V indicate query, key, and value, respectively.

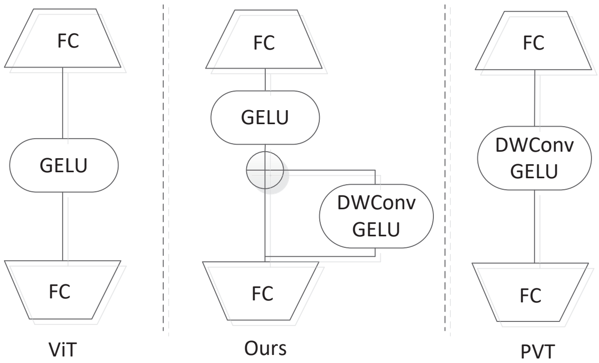

Figure 9. Comparisons of the feedforward layer between different transformer-based models. The *GELU* and *DWConv* represent Gaussian error linear units function and depthwise convolution, respectively.

3.3. Hybrid-Transformer Decoder

In this section, we introduce the improved hybrid-transformer decoder. The MLP, LayerNorm, and self-attention operation here are identical to the hybrid-transformer encoder, except the MHA is replaced by MA.

In order to capture strong discriminate semantic information, the hybrid-transformer decoder (H-TD) layer containing two H-TD structures is proposed for projecting encoded tokens back to pixel-space, thus producing refined change features. Specifically, the projected tokens pair $T^i_{res}, i = 1,2$ from original features pair $f^i, i = 1,2$ together with context-encoded tokens pair $T^i, i = 1,2$ are forwarded into the H-TD layer to exploit separate relations between each pixel per features and corresponding encoded token T^i_{enc} (Figure 10), or change relations between each pixel of difference features and encoded difference token (Figure 11).

Given feature tokens pair $T^i_{res} \in R^{B\times n\_patch\times D}, i = 1,2$ and rich context tokens pair $T^i_{enc} \in R^{B\times n\_patch\times D}, i = 1,2$, the first decoder structure adopts the Siamese hybrid-transformer decoder to obtain decoded representations $T^i_{dec} \in R^{B\times n\_patch\times D}, i = 1,2$ for each temporal image, which then perform reshape and permute operations to restore into the final pixel-level features pair $f_i \in R^{B\times D\times H\times W}, i = 1,2$. Finally, the change discrimination feature maps are generated by performing absolute difference between f_1 and f_2. Different from the first late-difference (LD) manner, the second early-difference (ED) structure performs the difference operations in a earlier stage. In practice, the residual tokens pair T^1_{res}, T^2_{res} and encoded tokens pair T^1_{enc}, T^2_{enc} subtract, respectively, the outputs that are efficiently exploited with the H-TD. By difference relations modeling directly, the produced token sequence represents a pixel-level semantic change discrimination. The permutation and reshape operations are also accomplished to obtain high-level change features.

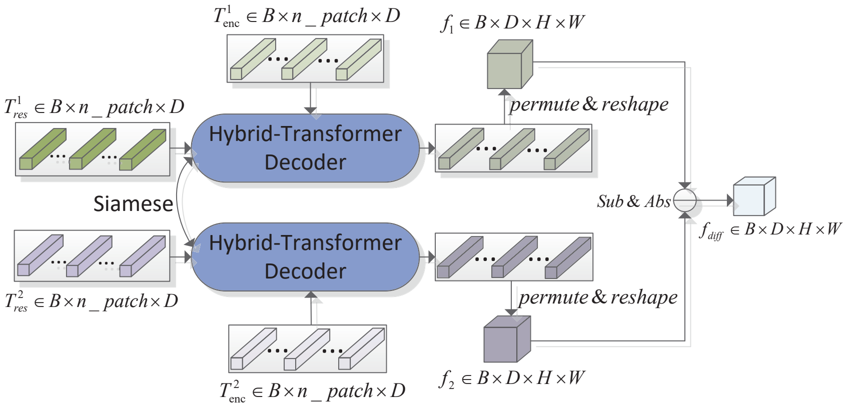
$T_{enc}^1 \in B \times n\_patch \times D$
$f_1 \in B \times D \times H \times W$
$T_{res}^1 \in B \times n\_patch \times D$
Hybrid-Transformer Decoder
permute & reshape
Siamese
Sub & Abs
$f_{diff} \in B \times D \times H \times W$
Hybrid-Transformer Decoder
$T_{res}^2 \in B \times n\_patch \times D$
permute & reshape
$T_{enc}^2 \in B \times n\_patch \times D$
$f_2 \in B \times D \times H \times W$

Figure 10. The late-difference structure of the hybrid-transformer decoder.

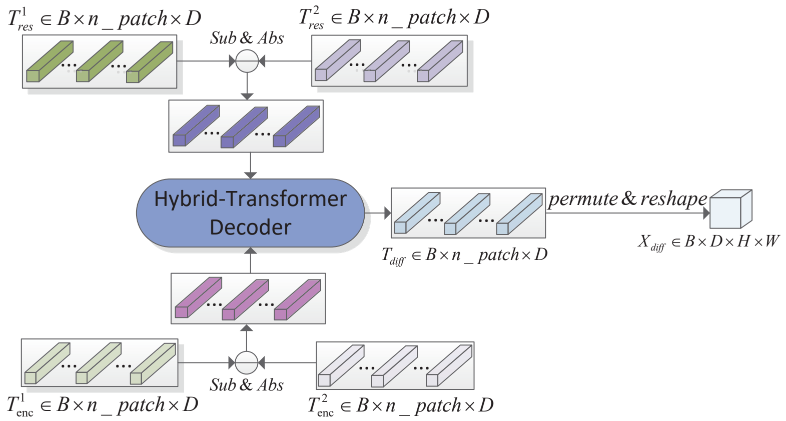
$T_{res}^1 \in B \times n\_patch \times D$
$T_{res}^2 \in B \times n\_patch \times D$
Sub & Abs
Hybrid-Transformer Decoder
permute & reshape
$T_{diff} \in B \times n\_patch \times D$
$X_{diff} \in B \times D \times H \times W$
Sub & Abs
$T_{enc}^1 \in B \times n\_patch \times D$
$T_{enc}^2 \in B \times n\_patch \times D$

Figure 11. The early-difference structure of the hybrid-transformer decoder.

Significantly, our H-TD consists of N blocks of the hybrid transformer decoder block, each of which constructs improved hybrid multi-head attention (H-MA) and DE feedforward layers. Rather than build self-attention within encoded tokens, the MA strongly builds mutual attention between encoded tokens and the original unprocessed ones. In addition, both ED and LD capture multi-scale representations thanks to the multiple values of r_i, thus capturing small-change objects surrounded by large background regions. The specific structure of the hybrid transformer decoder block is illustrated in Figure 12; the only difference from the H-TD block is that the queries in MA are derived from $\{T_{res}^1, T_{res}^2\}$ or $|T_{res}^1 - T_{res}^2|$ rather than the pure tokens $T_{enc}^i, i = 1, 2$. The formulations are defined as

$$\begin{aligned} Q_i &= T_{res}^{(b-1)} W_i^Q, \\ K_i, V_i &= MTA(T_{enc}^{(b-1)}, r_i) W_i^K, MTA(T_{enc}^{(b-1)}, r_i) W_i^V, \\ V_i &= V_i + LA(V_i) \end{aligned} \tag{8}$$

to obtain decoded tokens, formulated as

$$\begin{aligned} T_{dec}^{(b-1)} &= MSA(LN(Concat(head_1, \ldots, head_h)W)) + T_{dec}^{b-1}, \\ T_{dec}^{(b-1)} &= MLP(LN(T_{dec}^{b-1})) + T_{dec}^{b-1}, \\ where\ head_i &= \sigma(\frac{Q_i K_i^T}{\sqrt[2]{D_h}}) V_i \end{aligned} \tag{9}$$

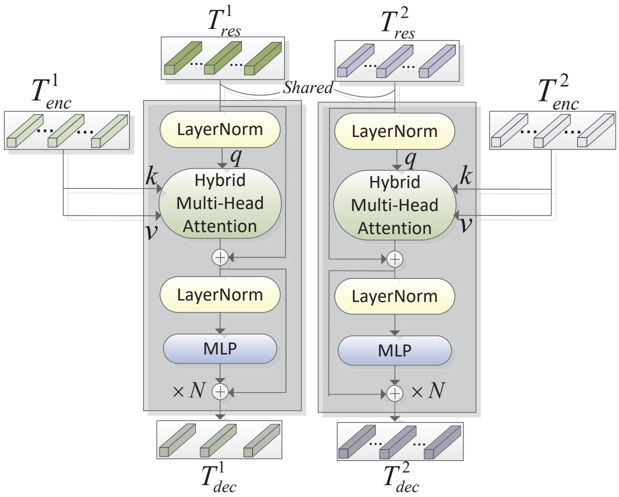

Figure 12. The architecture of the hybrid-transformer decoder block. The encoded tokens pair is decoded by the Siamese transformer decoder, and the self-attention is replaced by hybrid multi-head attention.

The decoded difference tokens set $T_{diff} \in R^{n\_patch \times D}$ is finally unfolded into three-dimensional features $f_{diff} \in D \times H \times W$.

3.4. Cascaded Feature Decoder

Concurrent vision works demonstrate the efficiency of multi-scale feature fusion in encoded low-level and high-level layers; skip-connections in decoder stages powerfully mitigate the missing details caused by global upsampling processes. Here, we propose a cascaded feature decoder (CFD) to aggregate semantic features varying multiple scales in a dense manner. As in Figure 13, the feature maps enhanced by our hybrid transformer are upsampled to the common scale from the highest layer of CNN backbone, and the output together with previous skip-connection acts as the input of the next decoder block. The features in the n-th decoder stage can be formulated as

$$F_n = Concat(g_n(F_{n-1}), u_n(F_{n-1})) \tag{10}$$

where $g(\cdot)$ and $u(\cdot)$ are 3×3 convolution and upsampling operations, respectively. In our work, four decoder blocks are employed for generating decoded features, where each block contains upsampling with bilinear interpolation, concatenation, and two convolutions with kernel size of 3×3. For multiple decoder stages, the decoded channel numbers are $[256, 128, 64, 16]$.

Until now, we have obtained upsampled feature maps $f \in R^{C \times H \times W}$, where the spatial size is identical to the input image. To obtain change probability maps $P \in 2 \times H \times W$, a prediction head composed of a light convolution and a softmax function are utilized to map dense prediction result, where the convolution kernel size is 3×3 and padding is 1. The pixel-wise probability map among each channel of P represents the changed and unchanged probability corresponding to this pixel, where the higher value will be determined. In the inference stage, a pixel-wise *Argmax* operation is adopted for producing a visual prediction map.

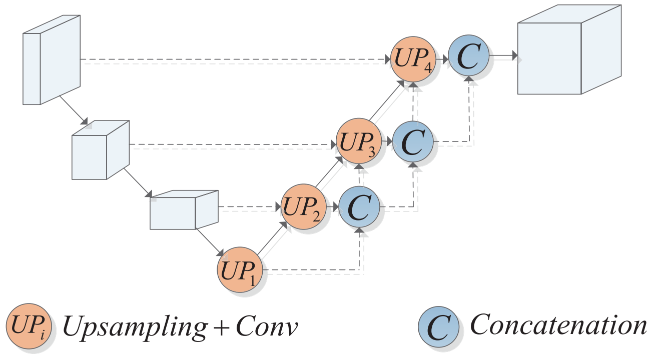

Figure 13. The architecture of the cascaded feature decoder. The upsampling is achieved by bilinear interpolation.

4. Experiments

4.1. Datasets and Implementation Details

In this work, we conduct experiments on two HR remote images change detection datasets. The first is LEVIR-CD [16], which was publicly collected from Google Earth (GE) covering multiple regions at different times (2002 to 2018). It contains 637 bitemporal image patch pairs of size 1024 × 1024, where the great majority of land-cover changes focus on manmade building changes. Following the default split ratio, 445:64:128 images are obtained as training/validation/test set. Considering the GPU memory consumption, all the images are cut into small patches of size 256 × 256. Therefore, 7120/1024/2048 pairs of patches are generated for training/validation/test. Another is SYSU-CD [28], which contains 20,000 pairs of 0.5/pixel aerial images of size 256 × 256 within the time period from 2007 to 2014 in Hong Kong. Different from the former, SYSU-CD constructs fine-grained change types including the new construction and destruction of buildings, the replacement of urban ground, seasonal changes in vegetation and oceans, and the road expansion. The default training/validation/test split is 12,000/4000/4000.

To demonstrate the effectiveness of our Hybrid-TransCD, some basic models are set for ablation comparison.

- **Baseline**: A light CNN backbone (ResNet18) with a single-level decoder sub-network. The decoder sub-network comprises four upsampling blocks for progressively restoring the image scale, and the fusion among multiple outputs is used to predict the final change map.
- **H-Res-E4-D4-ED-CFD**: The CNN backbone with the proposed hybrid transformer layer including four H-TE and four H-TD blocks, the ED decoder structure performs as the H-TD layer. In the feature decoder stages, the cascaded feature decoder is utilized.
- **H-Res-E4-D4-LD-CFD**: The same as H-Res-E4-D4-ED-CFD, except that the ED decoder is replaced by LD.
- **H-Res-E1-D1-ED-CFD**: The numbers of H-TE block (M) and H-TD block (N) are both reduced to 1.
- **H-Res-E1-D1-LD-CFD**: Identical to the previous one except the ED decoder structure is replaced by LD.
- **H-Res-E4-D0-LD-CFD**: The H-TD behind H-TE is removed by setting N to 0 while M is 4.
- **H-Res-E0-D4-LD-CFD**: The H-TE is removed by setting M to 0 while four H-TDs are employed.
- **H-E4-D4-LD-CFD**: Different from the above, which combines CNN-based and transformer-based features, the input here is directly processed by our hybrid transformer network. Specifically, the bitemporal images $I^i \in R^{3 \times H \times W, i=1,2}$ are linearly projected rather than f^i.

- **H-Res-E4-D4-LD-Single**: Compared to H-Res-E4-D4-LD-CFD, the cascade feature decoder is not applied to this structure. Specifically, the feature maps from the last decoder stage are concatenated with skip-connections for producing final features.

Our work is implemented by PyTorch with a single NVIDIA 3090 GPU of 24 GB memory. The generic data augmentation operations, including crop, flip, rotation, and Gaussian blur are adopted to avoid overfitting. The Adam solver [29] is utilized as the model optimizer with $\beta 1 = 0.5$ and $\beta 2 = 0.999$. The initial learning rate is 0.0005 and linearly decays according to the training iterations. For both LEVIR-CD and SYSU-CD datasets, our default training epochs is 100. The backbone ResNet (i.e., ResNet18) or hybrid-ViT are pretrained on ImageNet [30]. In the training stage, we use cross entropy function as the loss function of the model, which is defined as

$$L = \frac{1}{H \times W} \sum_{i=1}^{H \times W} [y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \tag{11}$$

To verify the effectiveness of our method, six metrics are used, as follows:

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

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

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

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

$$IoU = \frac{TP}{FP + FN + TP} \tag{16}$$

$$Kappa = \frac{(OA - P)}{(1 - P)}, \tag{17}$$

$$where \quad P = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + FP + TN + FN)^2} \tag{18}$$

where true positive (TP) indicates the number of pixels predicted correctly as changed, true negative (TN) represents the number of pixels predicted correctly as unchanged, false positive (FP) denotes the number of pixels predicted incorrectly as changed while false negative (FN) means the number of pixels predicted incorrectly as unchanged. F_1 comprehensively considers the precision and recall, thus performing the main index. In addition, most public change detection datasets inherently have class-imbalanced characteristics, making the model partial to a single category. Therefore, $Kappa$ achieves penalizing the "bias" index by replacing OA, which means the more unbalanced the confusion matrix is, the higher the p and the lower the $Kappa$, thereby giving a low score to the model with strong "bias".

4.2. Ablation Study of Existing Methods

Here, we experiment on two datasets to compare the proposed method with recent change detection methods, which include pure CNN-based, attention-based, and transformer-based models. As is shown in Table 1, the first four items only build end-to-end deep convolution neural networks without considering feature global contexts, where the fourth item proposes a more complex multi-level resolution fusion network framework. Although U-Net++ achieves a higher F_1 (4.22%) and IoU (0.98%) compared to FC-Siam-Conc, the computational cost is comparatively huge. Our Hybrid-TransCD outperforms the FC-Siam-Conc by 6.37/1.96/3.58 points of F_1, IoU, and $Kappa$, while the OA

(99.00%) achieves insignificant improvement. The next two items are all attention-based CD methods, where DASNet introduces spatial attention and channel attention modules based on metric learning, but for the dual features pair, rather discriminative change features are utilized, resulting in 3.78 points of F_1 and 5.31 points of IOU reduced compared with FC-Siam-Conc. BiT firstly introduced a transformer into a change detection deep network and achieved the highest F_1, OA, and $Kappa$ compared with earlier CNN-based methods. However, BiT merely adopted single-scale ViT within one transformer layer, causing weak $Precision$ (89.24%) and IoU (80.68%), which are 1.42% and 0.26% lower than U-Net++, respectively. Our Hybrid-TransCD performs with hybrid ViTs to model multi-scale attentions per layer, so multiple objects of different scales in the scene are effectively captured. Compared with U-Net++, our method achieves an improvement by 2.15 points of F_1 and 0.98 points of IoU. Compared with attention-based methods such as STANet, we are slightly worse on $Recall$ by 2.3%, but the IoU (81.92%) and $Kappa$ (89.54%) are 3.28 points and 2.88 points higher, respectively. Due to the high-resolution nature of implicit change regions inherent in this dataset, all the methods achieve high OA. The visual comparison results of LEVIR-CD are shown in Figure 14.

Table 1. Comparison results of the LEVIR-CD dataset.

Method	Precision	Recall	F_1	IoU	OA	Kappa
FC-EF	81.26	80.17	80.71	71.53	98.39	84.10
FC-Siam-Conc	90.99	76.77	83.69	79.96	98.49	85.96
FC-Siam-Diff	89.64	82.68	86.02	78.86	98.65	85.78
U-Net++	90.66	85.32	87.91	80.94	98.24	86.79
DASNet	80.76	79.53	79.91	74.65	94.32	85.14
STANet	83.81	**91.02**	87.27	78.64	98.87	86.66
BiT	89.24	89.37	89.31	80.68	98.92	88.97
Hybrid-TransCD (ours)	**91.45**	88.72	**90.06**	**81.92**	**99.00**	**89.54**

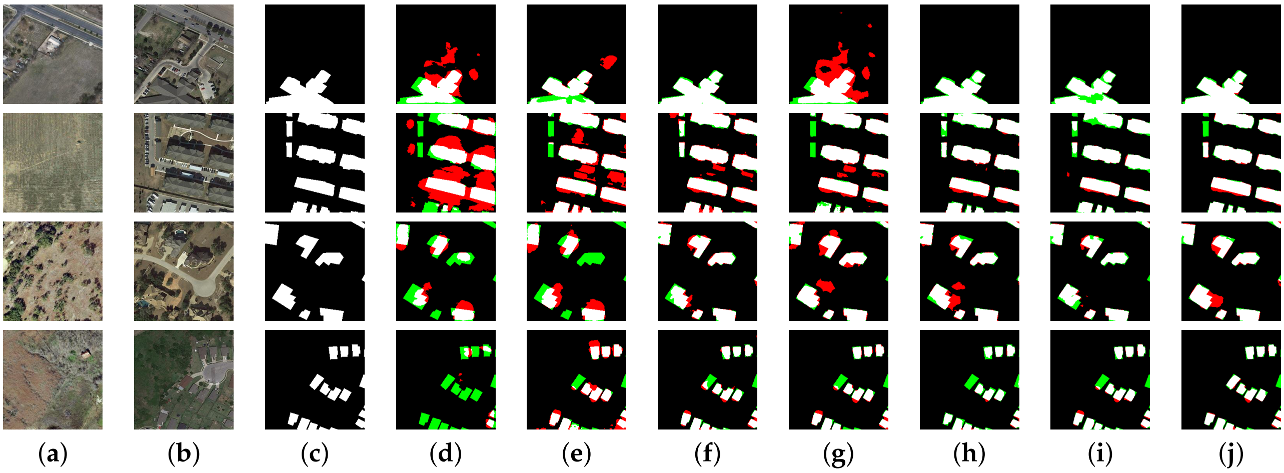

Figure 14. Visualizing comparison results of LEVIR-CD. Red indicates unchanged pixels predicted by error, and green indicates ignored changed pixels. (**a**) Image T1. (**b**) Image T2. (**c**) Ground truth. (**d**) FC-EF. (**e**) FC-Siam-Conc. (**f**) U-Net++. (**g**) DASNet. (**h**) STANet. (**i**) BiT. (**j**) Hybrid-TransCD (ours).

Similarly, the quantitative comparison results of SYSU-CD are shown in Table 2. Our Hybrid-TransCD achieves superior F_1 (80.13%), IoU (66.84%) and $Kappa$ (74.27%) among state-of-the-art learning-based CD methods. Compared to the lightweight networks listed in the first three items, our method acquires significant improvement in all metrics. As for better model FC-Siam-Conc, the 3.78/5.09/1.94 points on F_1, IoU and $Kappa$ are improved. Although U-Net++ achieved slightly inferior $Precision$ (81.36%), which is 7.77 points lower than FC-Siam-Diff (89.13%), the $Recall$ (75.39%) is significantly improved by 14.18 points.

Compared to our model, 1.87 points of F_1 and 4.7 points of IoU are improved. The STANet achieves the best $Recall$ (85.33%), and the $Precision$ is much lower than our 12.29%, causing a 2.76-point drop. The BiT gained an improvement on F_1 and $kappa$, and the improvement of $Recall$ in complex scenarios is still not stable. The visual comparison results on SYSU-CD are shown in Figure 15.

Table 2. Comparison results of the SYSU-CD dataset.

Method	Precision	Recall	F_1	IoU	OA	Kappa
FC-EF	74.32	75.84	75.07	60.09	86.02	72.14
FC-Siam-Conc	82.54	71.03	76.35	61.75	86.17	72.33
FC-Siam-Diff	**89.13**	61.21	72.57	59.96	82.11	71.04
U-Net++	81.36	75.39	78.26	62.14	86.39	72.36
DASNet	68.14	70.01	69.14	60.65	80.14	68.37
STANet	70.76	**85.33**	77.37	63.09	87.96	71.24
BiT	82.18	74.49	78.15	64.13	90.18	73.14
Hybrid-TransCD (ours)	83.05	77.40	**80.13**	**66.84**	**90.95**	**74.27**

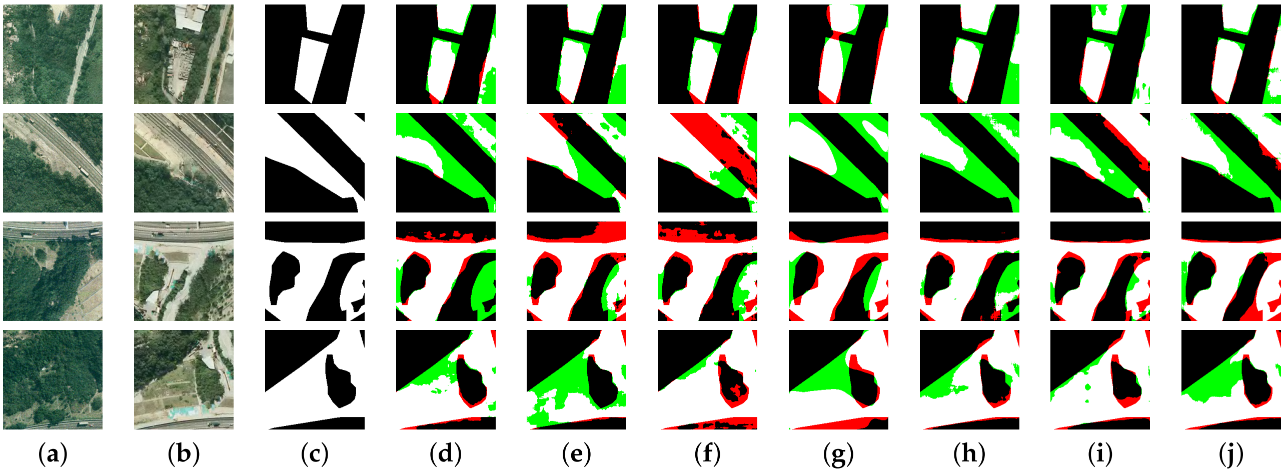

Figure 15. Visualizing comparison results on SYSU-CD. (**a**) Image T1. (**b**) Image T2. (**c**) Ground truth. (**d**) FC-EF. (**e**) FC-Siam-Conc. (**f**) U-Net++. (**g**) DASNet. (**h**) STANet. (**i**) BiT. (**j**) Hybrid-TransCD (ours).

The computational efficiency comparison of different algorithms is shown in Table 3, where our model uses four hybrid encoding blocks and four hybrid decoding blocks. As can be seen, compared with UNet++, our model only increases 35.31 M parameters, but we obtain better results. The attention-based method STANet consumes 116.93 M parameters due to more matrix multiplication operations. Although we only increase parameters of 46.97 M, the effect is significantly improved. BiT adopts the traditional ViT model, but ignores multi-scale representations, so we achieve multi-granularity context capture by improving multi-head attention, and the parameters are only increased by 44.72 M.

Table 3. Comparison of computational efficiency of different methods.

Method	Params (M)	FLOPs (G)
FC-EF	81.35	20.36
FC-Siam-Conc	81.54	21.58
FC-Siam-Diff	81.35	21.42
U-Net++	131.26	47.35
DASNet	108.69	31.33
STANet	116.93	36.58
BiT	121.85	42.99
Hybrid-TransCD (ours)	166.57	51.38

4.3. Ablation Study of Proposed Modules

To assess the effectiveness of the introduced hybrid-transformer, multiple model structures in Section 4.1 were used to experiment on the test data of LEVIR-CD and SYSU-CD datasets. As in Table 4, the necessary metrics including F_1, $Kappa$, and OA are given among multiple methods. In addition, the numbers of model parameters ($Params$) and computational cost ($FLOPs$) indicate the complexity of the corresponding structure. By comparing the first three items, we can observe that both ED and LD transformer decoder manners are effective to our CD task, where ED and LD mean early-difference and late-difference structures of the hybrid-transformer decoder, respectively. The $H - Res$ indicates whether the ResNet backbone is utilized for representing shallow semantic features. The $E(i) - D(j), i, j = 0, \ldots 4$ represents the numbers of H-TE block and H-TD block, where 0 means that the transformer operation is not adopted to the corresponding encoder and decoder. We can observe that H-Res-E4-D4-ED-CFD achieves the highest F_1, $Kappa$, and IoU on both datasets but occupies slightly larger computational costs. Compared with baseline, H-Res-E4-D4-LD-CFD improved by 2.94 and 4.71 points on F_1 and IoU, respectively, while $FLOPs$ (55.18 G) merely about doubled. By comparing H-Res-E4-D4-LD-CFD and H-Res-E1-D1-LD-CFD, more hybrid transformer blocks demonstrate the improvement by 0.83%/0.29% of F_1 and 0.45%/0.41% of IoU on LEVIR/SYSU, and the comparison between H-Res-E1-D1-ED-CFD and H-Res-E4-D4-ED-CFD is similar. By comparing H-Res-E4-D4-LD-CFD and H-Res-E0-D4-LD-CFD, H-Res-E4-D4-LD-CFD, and H-Res-E4-D0-LD-CFD, respectively, it is clearly proved that both H-TE and H-TD can effectively improve the ability to capture global dependencies of our model. In addition, the experiment is also performed with a pure transformer model method (H-E4-D4-LD-CFD) based on the primary bitemporal images to gap the effect of extracted features by CNN. It demonstrates that the designed hybrid-transformer structure strongly represents discriminative semantic features for the CD task loading with lighter computational costs, but there still exist gaps with pure CNN-based structure. Finally, the method (H-Res-E4-D4-LD-Single) of replacing the CFD with a simple single-stream decoder is conducted to verify the efficiency of the CFD. The visualized ablation experiment results are shown in Figure 16, from which we observe that the low-level features from the shallow layers of the CNN are powerfully merged with high-level semantic features enhanced by the hybrid transformer module. In addition, many scenes where objects of different scales are interlaced are accurately distinguished as the changing area, demonstrating the high robustness of the proposed transformer model.

Table 4. Quantitative comparisons of different model structures.

Method	LEVIR-CD			SYSU-CD			Params (M)	FLOPs (G)
	F_1	Kappa	IoU	F_1	Kappa	IoU		
Baseline	86.99	86.35	76.99	75.25	67.92	62.87	16.64	26.86
H-Res-E4-D4-ED-CFD	**90.06**	**89.54**	**81.92**	**80.13**	**74.27**	**66.84**	183.83	67.58
H-Res-E4-D4-LD-CFD	89.93	89.41	81.70	79.53	73.37	66.02	173.73	55.18
H-Res-E1-D1-LD-CFD	89.10	88.36	81.25	79.24	72.72	65.61	27.69	27.72
H-Res-E1-D1-ED-CFD	89.73	89.24	81.44	77.46	70.90	63.22	27.69	27.44
H-Res-E0-D4-LD-CFD	89.21	88.65	81.33	79.05	72.97	65.36	106.08	60.63
H-Res-E4-D0-LD-CFD	88.23	87.63	78.93	71.94	64.00	56.18	106.08	60.63
H-E4-D4-LD-CFD	84.40	83.60	73.01	78.13	71.84	64.12	23.11	8.66
H-Res-E4-D4-LD-Single	88.87	88.65	80.84	78.68	72.68	65.47	166.57	51.38

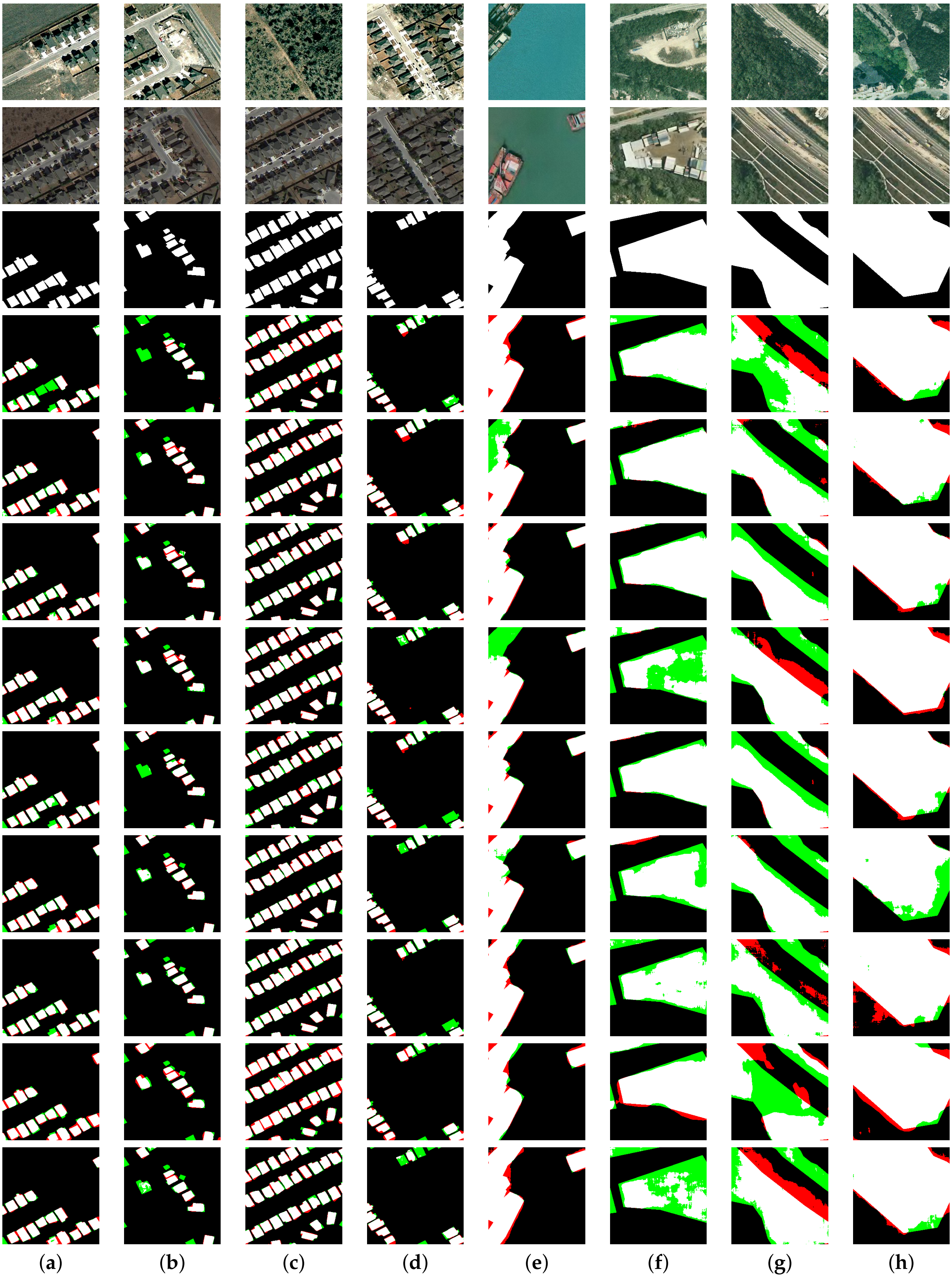

Figure 16. Visual ablation results of different model structures. Panels (**a**)–(**d**) represent the results of LEVIR-CD, and (**e**)–(**h**) represent the results of SYSU-CD. Each column, from top to bottom, represents: baseline, H-Res-E4-D4-ED-CFD, H-Res-E4-D4-LD-CFD, H-Res-E1-D1-ED-CFD, H-Res-E1-D1-LD-CFD, H-Res-E4-D0-LD-CFD, H-Res-E0-D4-LD-CFD, H-E4-D4-LD-CFD, H-Res-E4-D4-LD-Single.

As our hybrid transformer structure proposes a new strategy for aggregating multi-scale tokens, the comparisons of different token aggregation operations are given in Table 5. Compared with linear and convolution aggregation functions, our operation obtains more improvements with similar computation costs. Especially for those complex scale objects, the proposed method adaptively preserves both global and local information. The visual comparison results are given in Figure 17d,e,h, where the first four lines and the last four lines are the results of LEVIR-CD and SYSU-CD.

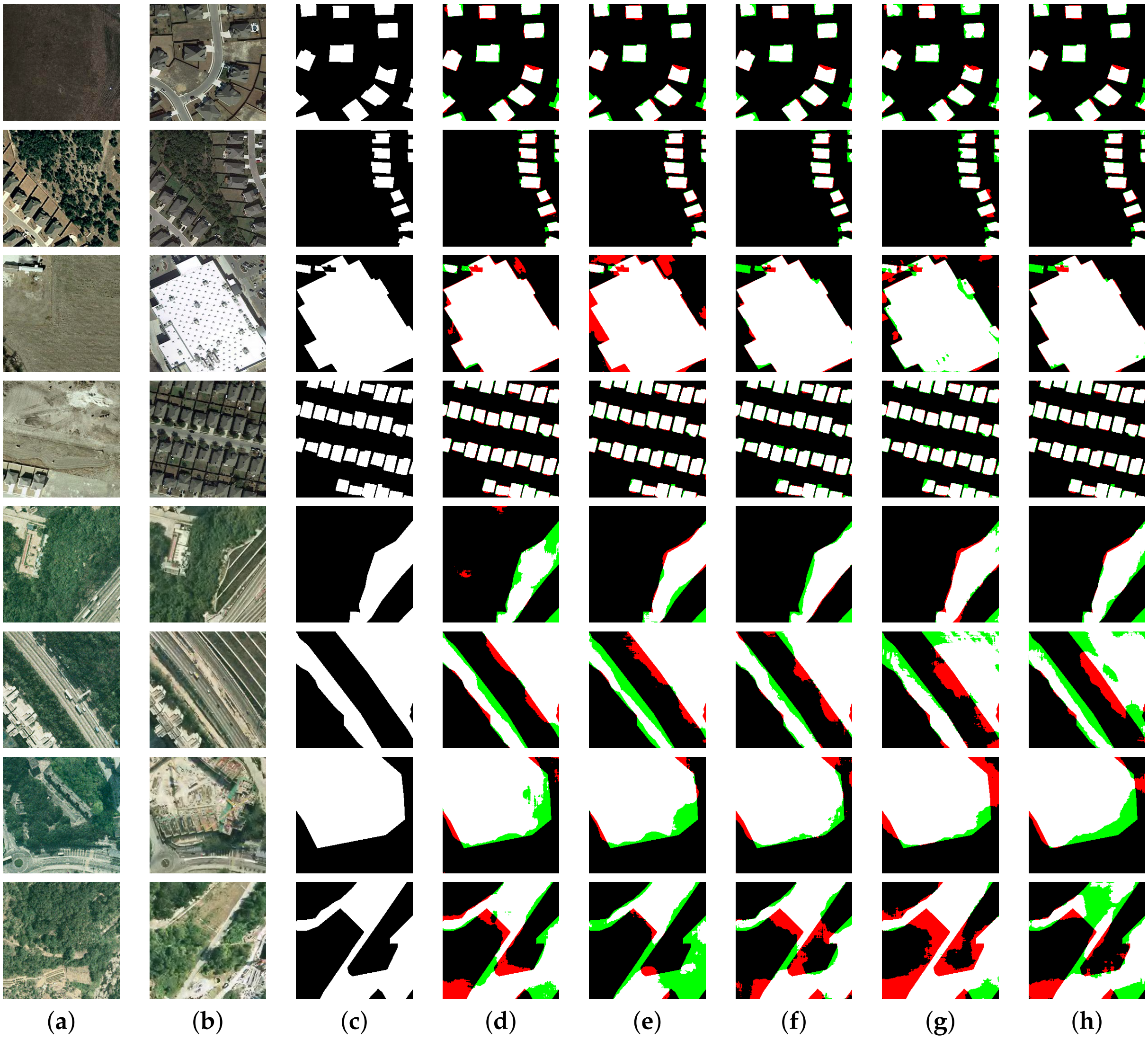

Figure 17. Visual ablations results of different token aggregation operations and different feedforward layer, where the first four lines and the last four lines are the results of LEVIR-CD and SYSU-CD, respectively. (**a**) Image T1. (**b**) Image T2. (**c**) Ground truth. (**d**) Linear token aggregation operation. (**e**) Convolution token aggregation operation. (**f**) Traditional feedforward layer. (**g**) DW-feedforward layer. (**h**) Multi-scale token aggregation operation + DE-feedforward layer.

In ViT [8], the feedforward layer simply conducts MLP operation with two linear functions, causing the negligence of pixel-level information. Subsequent depthwise separable convolutions demonstrate the global information integration capability in the feedforward layer [25], but there is still no interaction of different token information. Therefore, we conduct experiments on our detail-enhanced feedforward layer to demonstrate the ability of global and local token information to complement each other. From Table 6, the traditional feedforward layer, depthwise feedforward layer, and ours are compared. The detail-enhanced feedforward layer achieves the top performance on both datasets, while the model parameters (173.73 M) and FLOPs (55.19 G) are identical to the DW-feedforward layer. The visual comparison results are given in Figure 17f–h.

Table 5. Comparisons of different token aggregation operations.

Aggregation	LEVIR-CD			SYSU-CD			Params (M)	FLOPs (G)
	F_1	Kappa	IoU	F_1	Kappa	IoU		
Linear	89.77	89.24	81.44	76.16	70.04	61.50	147.6	42.18
Convolution	89.82	89.27	81.52	77.17	69.73	62.82	171.71	53.47
Ours	**89.93**	**89.41**	**81.70**	**79.53**	**73.37**	**66.02**	173.73	55.18

Table 6. Comparisons of different feedforward layers.

Layers	LEVIR-CD			SYSU-CD			Params (M)	FLOPs (G)
	F_1	Kappa	IoU	F_1	Kappa	IoU		
FeedForward	87.84	87.24	78.32	76.57	69.80	62.03	173.65	55.14
DW-FeedForward	89.08	88.52	80.30	77.44	70.07	63.18	173.73	55.19
DE-Feedforward	**89.93**	**89.41**	**81.70**	**79.53**	**73.37**	**66.02**	173.73	55.19

5. Discussion

As proposed, the transformer module strongly represents rich semantic features that reveal complex change objects or regions, where tokens containing both global and local concepts effectively model multi-scale attentions. In addition, the cascade feature decoder complements the offset gap between low-level features from CNN and high-level features from the transformer, thus learning elaborate pixel-level semantic information. As in Figure 18, we visualize the bitemporal feature maps generated by our hybrid transformer and the final stage of the decoder on both datasets, where Figure 18b,e represent the attention maps superimposed on the original T^1 and T^2 respectively, and Figure 18h visualizes the heatmap from UP_4. Red represents higher attention factor and blue denotes lower factor. From Figure 18 we can observe that features enhanced by the hybrid transformer actively learn representations related to change regions. Especially for the building changes in the LEVIR-CD dataset, the model even distinguished the delicate object edge. Although the SYSU-CD dataset contains large fuzzy forest and grassland changes, the model still better emphasizes the main changes, except for some unpredictable multi-view changes (e.g., shadows of buildings and trees). To further analyze the semantic information of tokens, Figure 18c,f show the attention maps of tokens for bitemporal images. As can be seen, these tokens fully capture long-range dependencies of per-image patches, thereby discriminating changes in various categories and scales, where strips of different colors represent different token information.

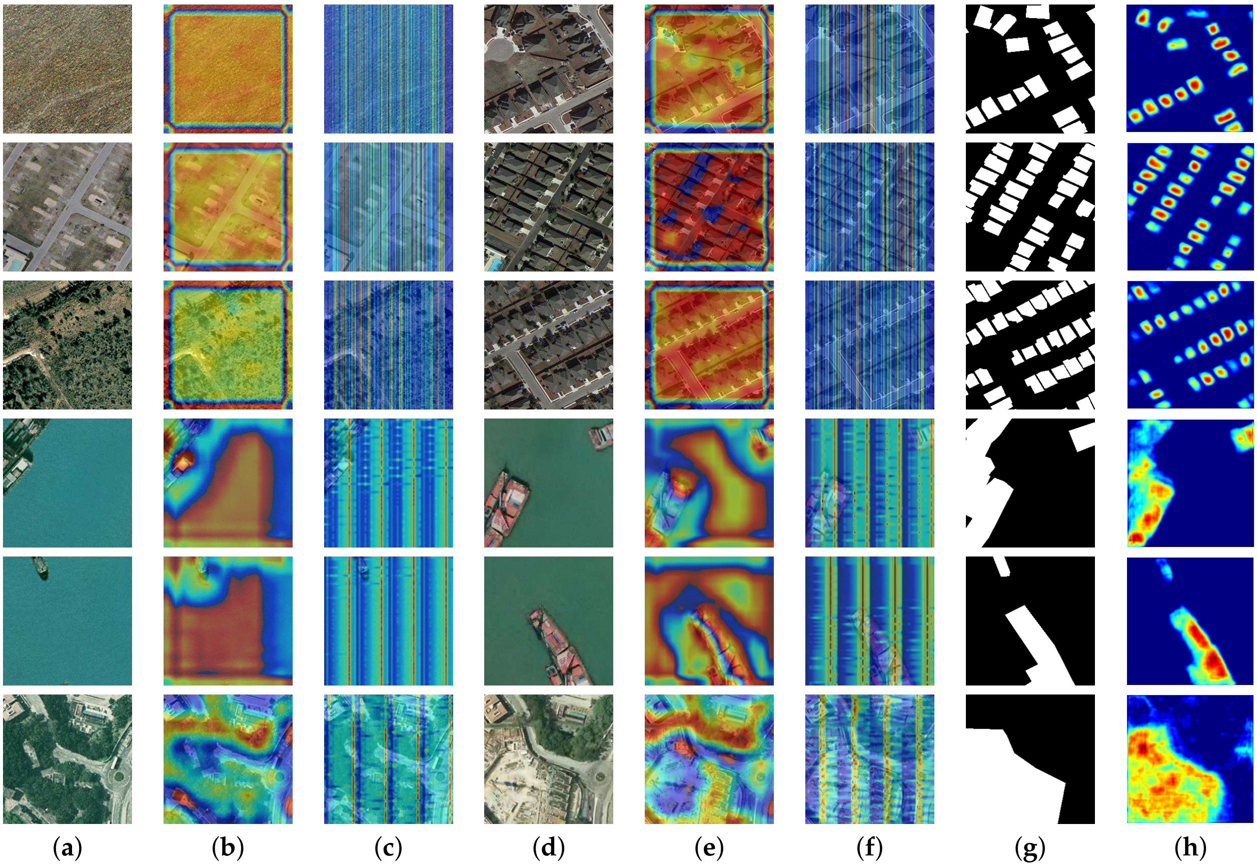

Figure 18. Results of model visualization, where the first three lines and the last three lines are the results of LEVIR-CD and SYSU-CD, respectively. (**a**) Image T1. (**b**) Attention map of Image 1. (**c**) Attention map of Token 1. (**d**) Image T2. (**e**) Attention map of Image 2. (**f**) Attention map of Token 2. (**g**) Ground Truth. (**h**) Heatmap of features from UP_4.

6. Conclusions

In this work, we propose a new transformer-based network Hybrid-TransCD for HR remote sensing image change detection. Compared to early CNN-based and attention-based methods, our model achieved superior performance without increasing heavy computational costs. In the meantime, we introduced a hybrid transformer structure for capturing multiple granularity global context dependencies. Improving on generic ViT and PVT, we designed a new multi-scale self-attention, and the tokens representing fine-grained change detail of small objects and coarse-grained change regions information are aggregated in a hybrid manner, thus effectively preserving spatial–spectral features of complex scenes. Two hybrid-transformer decoder structures are proposed to perform a backprojection on encoded tokens, thus further obtaining context-enhanced difference discriminate features. Both H-TE and H-TD layers are performed multiple times to represent hierarchical difference token information, where self-attention within one block is capable of capturing local and global representations with the cost of lightweight matrix multiplication operations. As low-level features represent rich texture details, we designed a cascade feature decoder for progressive fusion of low-level features and semantically-rich high-level features while restoring features resolution. The experiments on two public HR remote-image change detection datasets demonstrate the efficiency of our method, and the training time is greatly reduced compared with concurrent learning-based methods. In the future, we will be committed to researching more lightweight pure transformer-based change detection models, and further improving the generalization of the model on more datasets.

Author Contributions: Conceptualization, Peng Zhang and Qingtian Ke; Methodology, Qingtian Ke; Software, Qingtian Ke; Validation, Qingtian Ke; Formal Analysis, Peng Zhang and Qingtian Ke; Writing—Original draft preparation, Qingtian Ke; Writing—Review and editing, Peng Zhang and Qingtian Ke; Funding acquisition, Peng Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Shenzhen Science and Technology Program, grant number KQTD20190929172704911.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data presented in this study are available from the author upon reasonable request.

Acknowledgments: The authors sincerely appreciate the helpful comments and constructive suggestions given by the academic editors and reviewers.

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

References

1. 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]
2. 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.* **2020**, *18*, 811–815. [CrossRef]
3. Fang, B.; Pan, L.; Kou, R. Dual learning-based siamese framework for change detection using bitemporal VHR optical remote sensing images. *Remote Sens.* **2019**, *11*, 1292. [CrossRef]
4. Wiratama, W.; Lee, J.; Sim, D. Change detection on multi-spectral images based on feature-level U-Net. *IEEE Access* **2020**, *8*, 12279–12289. [CrossRef]
5. Wu, C.; Zhang, F.; Xia, J.; Xu, Y.; Li, G.; Xie, J.; Du, Z.; Liu, R. Building Damage Detection Using U-Net with Attention Mechanism from Pre-and Post-Disaster Remote Sensing Datasets. *Remote Sens.* **2021**, *13*, 905. [CrossRef]
6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
8. Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
9. Zheng, Z.; Ma, A.; Zhang, L.; Zhong, Y. Change is Everywhere: Single-Temporal Supervised Object Change Detection in Remote Sensing Imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 15193–15202.
10. Liu, R.; Jiang, D.; Zhang, L.; Zhang, Z. Deep depthwise separable convolutional network for change detection in optical aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1109–1118. [CrossRef]
11. Ke, Q.; Zhang, P. CS-HSNet: A Cross-Siamese Change Detection Network Based on Hierarchical-Split Attention. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9987–10002. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
13. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
14. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic segmentation with context encoding and multi-path decoding. *IEEE Trans. Image Process.* **2020**, *29*, 3520–3533. [CrossRef]
15. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Haozhe, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional siamese networks for change detection of high resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *14*, 1194–1206. [CrossRef]
16. 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]
17. Ke, Q.; Zhang, P. MCCRNet: A Multi-Level Change Contextual Refinement Network for Remote Sensing Image Change Detection. *ISPRS Int. J. Geo.-Inf.* **2021**, *10*, 591. [CrossRef]
18. Zhang, Y.; Fu, L.; Li, Y.; Zhang, Y. Hdfnet: Hierarchical dynamic fusion network for change detection in optical aerial images. *Remote Sens.* **2021**, *13*, 1440. [CrossRef]

19. 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*, 1–5. [CrossRef]
20. 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 bitemporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
21. Raza, A.; Liu, Y.; Huo, H.; Fang, T. EUNet-CD: Efficient UNet++ for Change Detection of Very High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
22. Chen, H.; Qi, Z.; Shi, Z. Efficient transformer based method for remote sensing image change detection. *arXiv e-Prints* **2021**, arXiv:2103.00208.
23. Wang, Z.; Zhang, Y.; Luo, L.; Wang, N. TransCD: Scene change detection via transformer-based architecture. *Opt. Express* **2021**, *29*, 41409–41427. [CrossRef]
24. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv* **2021**, arXiv:2102.12122.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030.
26. Wang, W.; Yao, L.; Chen, L.; Lin, B.; Cai, D.; He, X.; Liu, W. CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention. *arXiv* **2021**, arXiv:2108.00154.
27. Lin, H.; Cheng, X.; Wu, X.; Yang, F.; Shen, D.; Wang, Z.; Song, Q.; Yuan, W. CAT: Cross Attention in Vision Transformer. *arXiv* **2021**, arXiv:2106.05786.
28. 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]
29. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.