Journal of Geovisualization and Spatial Analysis (2025) 9:1
https://doi.org/10.1007/s41651-024-00202-3

HFNet: Semantic and Differential Heterogenous Fusion Network for Remote Sensing Image Change Detection

Yang Han[1] · Jiayi Li[2] · Yang Qu[1] · Leiguang Wang[2,3] · Xiaofeng Pan[4] · Xin Huang[1]

Accepted: 6 November 2024 / Published online: 25 November 2024

Abstract

Existing methods of change detection that rely on individual image modalities, such as concatenation and differencing, risk overlooking the distinct change patterns, and detailed features inherent in different modalities. This oversight can result in missed detections and suboptimal delineation of change boundaries. To address these limitations, this study presents a novel semantic and differential heterogeneous fusion change detection method, termed HFNet (Heterogeneous Fusion Network). HFNet integrates two different bitemporal image input modes to analyze changes and detailed information within image pairs. To effectively integrate bitemporal-difference features, we have introduced a module called CFM (Correlation Fusion Module) in HFNet, designed to handle the bitemporal features of the concatenation branch and the differencing features of the Siamese branch. Furthermore, to exploit multiscale change information in images, we designed the GSFM (Global Scale Frequency Fusion) in the decoder stage to establish inter-scale dependencies and adaptively compute multiscale weights. In addition, to fully explore the spatiotemporal features of bitemporal images, we constructed a change detection feature extraction module with a hybrid spatiotemporal Convolution-Transformer structure, named STConvTrans. This feature extraction module is designed to efficiently extract and restore fine features from the temporally stacked branches of bitemporal images. Extensive experiments on public datasets demonstrate the superior performance of HFNet. Compared to the current state-of-the-art model, HFNet achieves improvements of 3.7%, 2.9%, and 0.4% in IoU on the CDD, WHU_CD, and SYSU_CD datasets, respectively. Moreover, branch line comparative experiments, ablation experiments, and feature extraction module comparison experiments further validate the effectiveness of HFNet's components.

Keywords Change detection(CD) · Hetrogeneous fusion · Convolution neural network(CNN) · Transformer

✉ Yang Qu
quyang@whu.edu.cn

Yang Han
han_rs@whu.edu.cn

Xiaofeng Pan
xfpan@meeb.sz.gov.cn

Xin Huang
xhuang@whu.edu.cn

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 129 Luoyu Road, 430079, Hubei, China

[2] Institute of Big Data and Artificial Intelligence, Southwest Forestry University, Kunming 650224, Yunnan, China

[3] Key Laboratory of State Forestry and Grassland Administration on Forestry and Ecological Big Data, Southwest Forestry University, Kunming 650224, Yunnan, China

[4] School of Environmental and Mapping Engineering, Suzhou University, Suzhou 234000, Anhui, China

Introduction

Remote sensing image change detection is a technique that employs multi-temporal remote sensing data to identify alterations in land cover. This method offers invaluable insights for post-disaster assessment (Eftekhari et al. 2023), urban development (Huang et al. 2013; Wen et al. 2021), land use monitoring (Li et al. 2017; Tu et al. 2021), and ecosystem monitoring (Jiang et al. 2022).

In recent years, deep learning-based change detection methods have gained significant prominence in the field of remote sensing, demonstrating commendable performance (Fang et al. 2019; Zhang and Lu 2019). CNNs, due to their merits of parameter sharing capabilities and multi-scale feature extraction, have been extensively applied in remote sensing image change detection (Daudt et al. 2018; JST 2015; Zhan et al. 2017; Hou et al. 2023). For instance, the Siamese CNN is utilized to extract features pixel by pixel

from images and encode them into feature vectors (Zhan et al. 2017). Changes in optical aerial images are identified by comparing the similarity between these two vectors. The integration of CNN in this endeavor enables the model to autonomously acquire knowledge of landcover semantic patterns and structures within the data, consequently enhancing the accuracy and robustness of change detection. However, these 2D convolution-based methods insufficiently preserve spatiotemporal information between bitemporal images.

To address this issue, Re3FCN (Song et al. 2018) employed a 3D fully convolutional network for change detection, enhancing the utilization of spatiotemporal information from bitemporal images compared to 2D convolutions. Nevertheless, traditional 3D convolution often leads to mutual interference among diverse spatial features, including spatial and channel features, as well as temporal features in images, resulting in the loss of temporal information. In addition, with the increase of image spatial resolution, despite the advantages of CNN-based methods in spatiotemporal feature representation, the locality of convolutional operations limits their receptive field (Zhou et al. 2014), resulting in insufficiency context information integration.

Recently, Transformers, which excel at extracting global contextual information, have demonstrated performance comparable to or even surpassing that of CNNs in the field of remote sensing change detection (Chen et al. 2021; Zhang et al. 2022). However, Transformers tend to prioritize global context over local features, which may pose challenges for accurately identifying edge pixels and could lead to insufficient supervision for smaller objects.

To tackle this challenge, researchers have endeavored to establish a mechanism that unifies local semantics from CNNs with the global contextual features from Transformers (Chen et al. 2021; Li et al. 2022; Yuan et al. 2022; Zhang et al. 2022). For instance, BIT_CD (Chen et al. 2021) employs ResNet as a feature extractor, followed by Transformer-based feature enhancement, and ultimately outputs change detection results through prediction heads. This methodology demonstrates promising results on the LEVIR-CD dataset. Similarly, TransUNetCD (Li et al. 2022) employs transformers for encoding image patches, which are then fused with high-resolution CNN feature maps to improve the extraction of local–global semantic features. While the aforementioned methods integrating CNNs and Transformers for change detection yield promising results. Nevertheless, contemporary methodologies primarily concentrate on straightforward connections of features produced by CNN and Transformer, placing a significant emphasis on spatial features. These approaches fall short in fully leveraging the temporal features inherent in bitemporal images and fail to unlock the complete potential of synergizing the strengths of both CNN and Transformer. Therefore, it is essential to propose a change detection method that deeply integrates the advantages of both CNNs and Transformers.

Furthermore, The existing deep learning change detection frameworks often concentrate solely on a specific type of image feature, such as differential (e.g., FC-Diff (Daudt et al. 2018), TransUNet (Li et al. 2022)) or bitemporal feature (e.g., FC-Siam-Conc (Daudt et al. 2018), FC-Siam-EF (Daudt et al. 2018), and SNUNet (Fang et al. 2021)). However, relying solely on a particular type of image feature may constrain network performance. To be specific, while bitemporal features can effectively capture extensive land cover information, they may obscure critical details related to changes, such as the reconstruction or renovation of buildings. And differential features emphasize areas of change by comparing images captured at different times. However, they are susceptible to interference from false changes due to the loss of semantic information related to land cover (Zhao et al. 2019). From this perspective, the establishment of a unified mechanism that integrates differential and bitemporal information holds the potential to enhance the accuracy and robustness of change detection. In summary, it is urgent to explore a comprehensive and effective change detection method that can deeply combine semantic features and differential features, while fully leveraging the advantages of both CNN and transformer.

Taking into consideration the issues and challenges, this study proposes a Semantic and differential Heterogeneous Fusion Network (HFNet) for change detection. In order to effectively extract temporal global features containing change information, our design incorporates two branches within the Encoder: the primary branch for temporal concatenated features and the auxiliary branch for differencing features. To achieve seamless integration of information from these branches, we merge data from each scale of the auxiliary branch into the primary branch. This process guarantees that the primary branch can adequately consider significant change information in differencing features at various scales. In contrast to single-branch networks exclusively relying on specific input types (land cover or difference), our proposed framework concurrently capitalizes on the advantages of both land cover features and differencing features. This approach mitigates issues such as false positives and false negatives resulting from one-sided information extraction in single-branch networks (Daudt et al. 2018; Fang et al. 2021; Zhang et al. 2018), providing heightened robustness and extracting the most discriminative features.

To effectively integrate the rich image details, present in temporal cascade features with the variation information found in difference features, we propose the Correlation Fusion Module (CFM). This module leverages correlation and interaction mechanisms between separate branches and bitemporal image pairs to fully exploit the advantages of temporal and differential features in terms of semantic detail

and change information. Specifically, the features processed by each CFM are subsequently fed into the next stage, progressively emphasizing the interplay between bitemporal and differential information. Subsequently, the introduction of a Global Scale Fusion Module (GSFM) aims to enhance the fusion and integration of information from various scales during the decoding phase. Furthermore, a novel hybrid spatio-temporal 3D convolution-Transformer network module, called STConvTrans, was designed for the change detection network. This module segregates 3D convolution operations in the temporal and spatial domains and combining them with the Transformer to explore global and local information. Finally, features from each period are aggregated through temporal convolution. This approach alleviates the impact of temporal perturbations during spatial information acquisition, enhancing temporal feature representation and effectively unleashing the potential of the feature extraction module in extracting features from bitemporal images.

The specific technical contributions include the following aspects:

1. A heterogeneous fusion change detection network framework, named HFNet, is proposed. At each stage of the encoder, this network concurrently introduces branches for land cover features and sequential differencing features, achieving interactive fusion. Additionally, GSFM is devised in the decoder to enhance information interaction, interaction, and feature utilization efficiency.
2. A novel change detection module, named STConvTrans, is introduced. This innovative hybrid feature extraction module combines 3D CNN and Transformer. This module simultaneously leverages the spatiotemporal features, translational invariance of 3D CNN, and the global receptive field of Transformer to better capture subtle changes and details of objects in images, widespread changes, large-scale terrain changes, or the evolution of overall scenes. This approach addresses the inadequacy of regionally correlated information when using CNN alone and the insensitivity to local changes when using Transformer alone, thereby enhancing the performance of change detection.

The remainder of this paper is organized as follows: In the "Related Work" section, an overview of the current state of change detection research based on deep learning is provided. The "HFNet" section presents the overarching framework of HFNet and comprehensively describes its components. In the "Result and Discussions" section, we conduct comparative experiments between HFNet and other state-of-the-art algorithms. The effectiveness of the various structures within HFNet and the significance of STConvTrans are verified through ablation experiments and feature extraction module comparison. Additionally, the network's parameters and computational time are discussed. The "Conclusion" section concludes the article and presents prospects for future work.

Related Work

Change Detection Methods Based on the Joint Mechanism of CNN and Transformer

Recently, the collaborative integration of CNNs and Transformers has gained prominence due to its potential to amalgamate the merits of both models, promising enhanced performance. However, common fusion methodologies often involve the direct incorporation of land cover semantic features, extracted by CNNs, into Transformers (Li et al. 2022; Liu et al. 2022a; Song et al. 2022; Wang et al. 2022). While this strategy initially integrates local and global information, the direct input of features may lead to information loss. For example, TransUNetCD (Li et al. 2022) and CTD-Former (Zhang et al. 2023a, b) utilize a weight-shared twin convolutional encoder to extract temporally invariant features (mainly pertaining to land cover) for each period individually, which are then concatenated and fed into Transformer for subsequent exploration of global features. However, this temporal concatenation approach still presents disorder in channel and time information when further input into the Transformer. In contrast, ACABFNet (Song et al. 2022) introduces CNN and Transformer branches to concurrently explore differential features. Despite the fusion of features in an inter-scale manner for fusion, there is still a lack of interaction within the same scale, resulting in incomplete information fusion.

In summary, contemporary methodologies address the issue of missing local details and inadequate contextual associations by leveraging the interaction between CNN and Transformer. However, they do not yet fully address the problem of mutual interference between temporal, spatial, and channel features in change detection. This suggests that there is still ample room for exploration in the collaborative application of CNN and Transformer. Hence, the proposition of a deeply integrated CNN and Transformer change detection feature extraction module becomes imperative.

Framework of Change Detection

Most existing deep learning-based change detection methods adopt an Encoder-Decoder structure, which, based on the form of input data, can be roughly categorized into two types. One approach is to fuse the two temporal images and then feed the fused image to the change detection network as a single input, known as early fusion (EF) (Daudt et al. 2018; Lebedev et al. 2018; Peng et al. 2019), as shown in

Fig. 1a. This approach can fully leverage global information, contributing to a better overall understanding of the scene. For example, registered pairs of remote sensing images were connected as input to an improved UNet + + network (Peng et al. 2019).

However, early fusion is executed without adequate feature extraction in the shallow layers of the network, which may result in inadequate depth and complexity, and consequently failing to capture intricate local relationships. To address this issue, researchers have proposed an alternative feature extraction method that utilizes a neural network to extract deep features from bitemporal images. The extracted bitemporal features are then fused and change detection is performed based on the fused features, named late fusion (LF) (Bao et al. 2020; Chen et al. 2020b; Daudt et al. 2018; Fang et al. 2021; Liu et al. 2020; Zhang et al. 2018; Zhang et al. 2023a, b), as shown in Fig. 1b. For instance, in PPCNet (Bao et al. 2020), bitemporal images are separately input to a pair of Siamese branches and fused at the end of the branches. However, information extracted through deep networks may suffer from some loss of details. Therefore, it is necessary to propose a change detection framework that combines the advantages of EF and LF.

Additionally, researchers have predominantly utilized feature fusion methods such as direct concatenation (Bao et al. 2020; Daudt et al. 2018; De Bem et al. 2020; Liu et al. 2020; Peng et al. 2019; Zhang et al. 2018) or differencing to extract semantic and differential features. However, these change detection frameworks generally rely on a single input modality (e.g., concatenation or differencing), which may lead to potential loss of information and interference in the extracted features, consequently impacting the precision and efficiency of change detection. For instance, Bi-SRNet (Ding et al. 2022) and SCanNet (Ding et al. 2024) concatenate the outputs of a pair of weight-shared convolutional encoders along the channel dimension and input them into another separate encoder to extract change features. This concatenation input approach effectively extracts information about the images themselves, but it may also lead to a situation where extraneous data obscures critical change information. Conversely, FC-Siam-Diff (Daudt et al. 2018) identifies changes by utilizing features derived from the differencing of bitemporal images. This approach effectively emphasizes change information, but it may also be affected by factors such as shadows and seasonal variations. It is clear that these two methods serve to complement one another.

Therefore, DMINet (Feng et al. 2023) introduced a new module to guide attention distribution between features and designed pixel level differencing and channel-level concatenation to capture potential multi-level differences. However, DMINet did not fully exploit the extraction of temporal concatenation and differencing features, nor did it adequately leverage their advantages for effective complementarity. Thus, the development of a change detection method that combines differencing and temporal concatenation features to preserve both detail and change information is highly necessary.

At the same time, in the aspect of multiscale information fusion, most existing methods predominantly employ feature pyramid structures in Encoder and Decoder to extract multiscale information. For example, SwinSUNet (Zhang et al. 2022) constructs a feature pyramid by downsampling in the Encoder and upsampling in the Decoder to comprehend and utilize features at different scales. However, the approach described focuses solely on achieving fusion between adjacent scales, neglecting fusion among multiple scales. This limitation results in incomplete utilization of multiscale features, leading to challenges like unclear edge detection and missed identification of small objects. Therefore, it is necessary to propose a suitable multi-scale fusion module to improve the aforementioned problems. ECAM (Fang et al. 2021) achieves weight redistribution between channels and scales through channel attention, effectively realizing multiscale information fusion. However, considering the differences in scene space and semantic information implied by features at different scales, features at larger scales often cover changes in overall structure and layout aspects, including but not limited to the overall outline of buildings and road layouts. Meanwhile, features at smaller scales are more sensitive, capturing subtle changes and details, such as variations in building textures and trees. Hence, it is crucial to

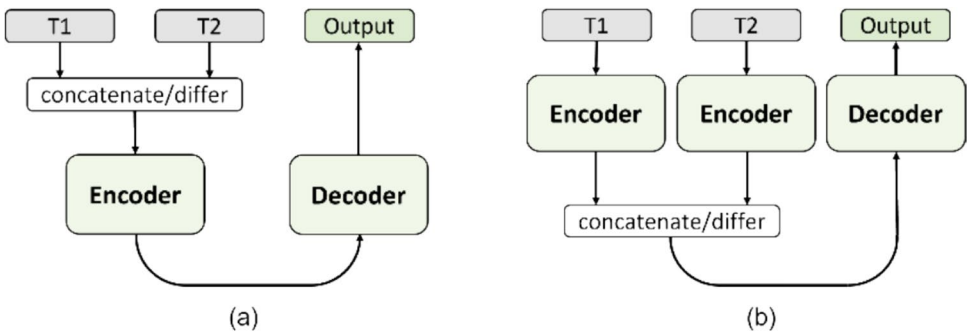

Fig. 1 Current change detection architectures. **A** Early fusion and (**b**) late fusion

adopt differentiated feature extraction and computation for information at different scales.

Hence, in this study, we propose a change detection network, HFNet, which comprises a bitemporal concatenation branch and a Siamese differencing branch. Additionally, we present a novel feature extraction module, STConvTrans, which combines spatial–temporal 3D convolution and Transformer for the bitemporal concatenation branch of the network. Furthermore, we design two modules: CFM for bitemporal-differencing branch fusion and GSFM for multi-scale fusion.

HFNet

Overall Structure

The overall architecture of the network is shown in Fig. 2. Unlike previous studies, we have designed two branches in the Encoder, namely the bitemporal concatenation branch and the Siamese differencing branch. The concatenated branch uses our proposed STConvTrans to capture land-cover changes (e.g., redevelopment of old residential areas) and global imaging condition differences, while the weight-shared Siamese differencing branch utilizes CNN to generate multi-scale difference features that are independent of time, aiming to eliminate the interference of local changes related to phenology, as each land cover has different phenological features. Then, a CFM is leveraged to fuse time-difference feature information. The fused features are input into the bitemporal concatenation branch and Decoder. In addition to the stage-wise up-sampling as current decoders (Zhang et al. 2022), a GSFM is designed to ensure the comprehensive extraction and fusion of diverse information inherent in features at different scales, achieving the effective utilization of multiscale spatiotemporal differential information.

In specific terms, our network takes bitemporal RGB images as input. To begin, we employ patch partitioning and linear embedding to predivide the original bitemporal images into patches and map the channel number to C, where the feature map size is $\mathbf{R}$. Subsequently, within the Encoder, the features are separately propagated through these two heterogeneous branches. To enhance multi-scale feature extraction, both branches comprise four stages, each achieving a twofold spatial downsampling, with the temporal and channel dimensions remaining unchanged. The output of each stage in the bitemporal concatenation branch is fused with the information from the differencing branch through the CFM block, which is transmitted to the subsequent stage of the concatenation branch. Each CFM output is not only utilized within the bitemporal concatenation branch but also input to the corresponding stage of the Decoder to recover change information.

Then, similar to the Encoder, the Decoder consists of three stages, with feature dimensions as $f_i^{De} \in \mathbb{R}^{(H/2^{i+1})\times(H/2^{i+1})\times(2^{i-1}C)}$, $i \in \{1,2,3\}$, i is the stage index. Each stage encompasses several STConvTrans units and a UM block (upsampling and merging block) (Zhang et al. 2022). Following the three stages, the multi-scale features from the Decoder are fed into the GSFM, to establish a bottom-up approach for multi-scale feature propagation and fusion. Ultimately, the multiscale features are restored to $Output \in \mathbb{R}^{H\times W\times C}$ through UM blocks and MLP to generate the change map.

STConvTrans

We notice that different land features have different semantic representations (such as buildings and roads), and their scales may vary. Additionally, the rates and directions of their temporal changes may differ (for example, artificial structures and vegetation exhibit distinct local phenological characteristics). Moreover, global imaging condition differences (such as cloudy and sunny conditions) in bitemporal images can also confound changes in land features. Therefore, in land change detection, attention should be paid not only to exploring the global and local semantic information of land features in the multiscale representation of the encoder but also to organically integrate local differential information and global differential information in the multiscale representation process. Therefore, in the hierarchical representation process, we introduce a novel hybrid spatio-temporal convolution-Transformer feature extraction module, named STConvTrans (Spatio-Temporal Convolution and Transformer). It first uses self-attention to highlight the spatial semantic of land features and employs temporal convolution to emphasize the temporal semantic differences. As illustrated in Fig. 3, STConvTrans is composed of two operations:

1. Temporal independent spatial feature mapping: Given the input of temporally independent features concatenated at two moments (i.e., mainly referring to land cover semantic features) f_1^C, f_2^C, we use a pair of weight-shared 3×3 spatial convolutions ($SConv$) to map the local detailed features f_1^L, f_2^L for each temporal instance. Subsequently, utilizing weight self-attention operations further captures the spatiotemporal information of f_1^L, f_2^L at a global scale, yielding the global spatiotemporal feature f_1^G, f_2^G. It is noted that the operation of weight sharing allows the convolution/self-attention to focus on extracting spatial features while alleviating interference from temporal information. Thus, f_1^G, f_2^G is sensitive to the local detailed land-cover changes and is also robust to the false alarms (e.g., shadow) caused by the viewing differences between different VHR images.

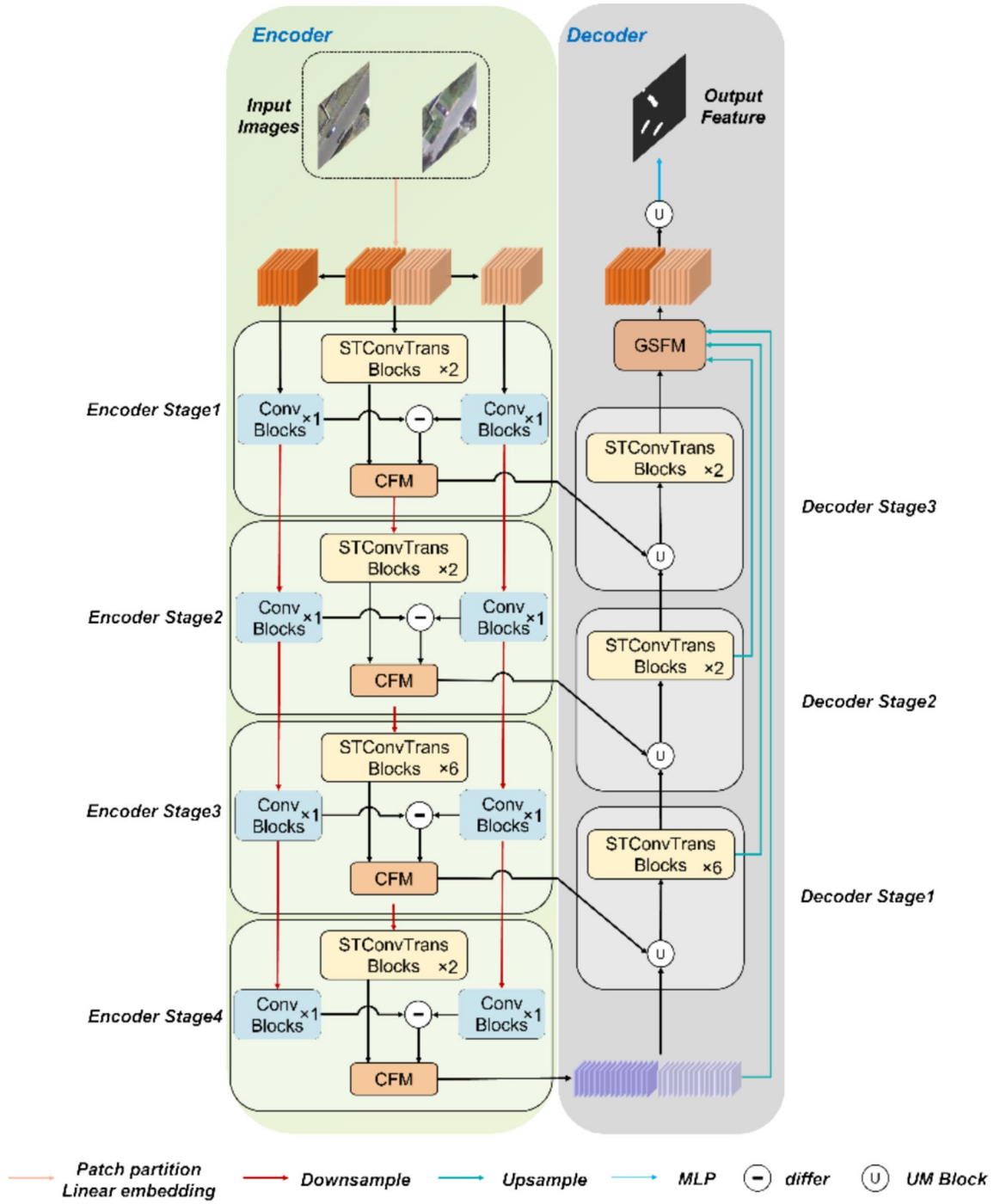

Fig. 2 Framework of the proposed HFNet

2. Bitemporal relevant feature mining: To achieve temporal interaction and extract the global imaging differences of temporally sensitive land features, the f_1^G, f_2^G is further input into a 1×1 temporal convolution (*TConv*) to enhance feature representation capabilities. Unlike conventional 1×1 convolutions acting on the channel dimension, *TConv* achieves information interaction between different times by operating on the temporal dimension.

STConvTrans can be employed as a universal module for change detection networks, with the feature size remaining unchanged before and after this module. The primary operation of STConvTrans is expressed by the following equation:

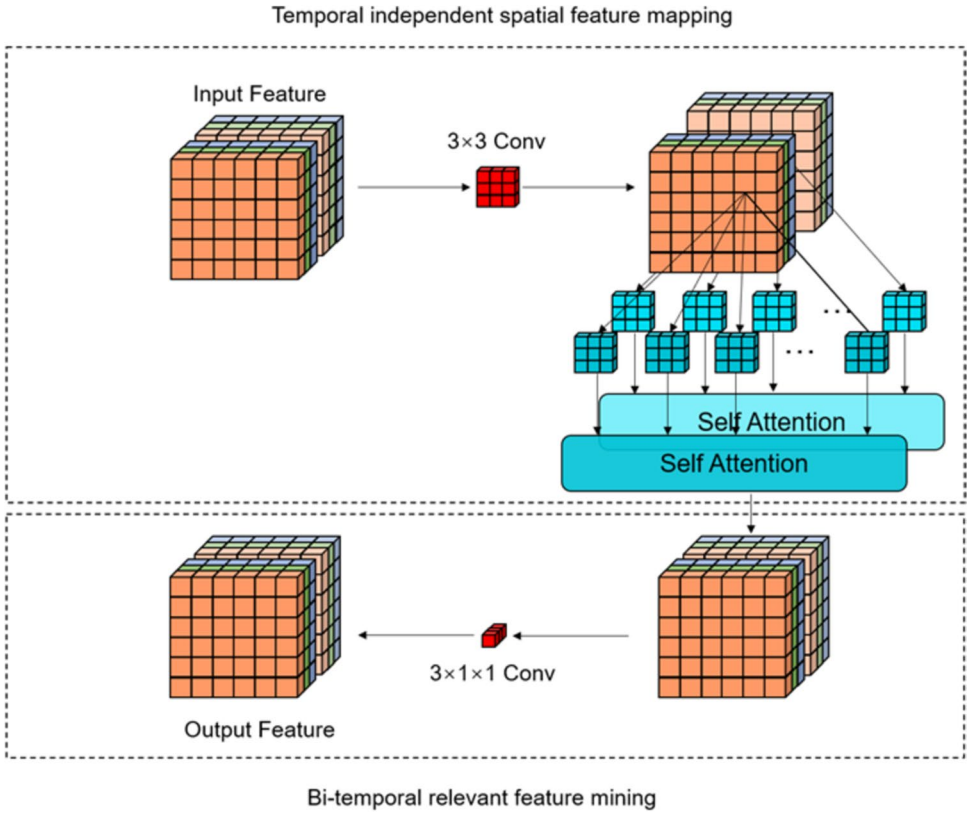

Fig. 3 STConvTrans

$$f_{st} = TConv(Att(SConv(f_1^C), (SConv(f_2^C))) \quad (1)$$

where f_1^C, f_2^C is the input feature, f_{st} is the output feature, *Att* is a self-attention, *SConv* is a space convolution with shared weights, *TConv* is a temporal convolution.

Correlation Fusion Module

CFM consists of two main components: correlation interaction and temporal fusion, as illustrated in Fig. 4. In this process, the region for correlation interaction effectively couples the detailed information of bitemporal features and the change information of.

differential features by mapping and calculating the similarity between them. Then, the temporal fusion region achieves temporal information interaction of the fused bitemporal features through 3D convolution.

1. Correlation interaction: In response to the global bitemporal land cover semantic features, f_i^P, from the primary branch, and the local features, f_i^{L1} and f_i^{L2}, from the auxiliary branch (where i represents the i stage output of the Encoder), CFM receives land cover semantic features f_i^P and difference feature $f_i^{Di} = f_i^{L1} - f_i^{L2}$. This is done to separately calculate the similarity information between the two temporal features and the difference feature; we first separate f_i^P into bitemporal features f_i^{T1}, f_i^{T2}. Subsequently, the bitemporal features and differential features are mapped through an MLP layer to reshape them into a new smoothed representation. Finally, pixel-wise multiplication is used to compute the similarity between each temporal feature and the differential feature. The correlation features are correspondingly multiplied with the original corresponding temporal features and the residual sum is obtained.

Due to each branch extracting different features, and there is some overlap in the information contained in different features, we need to use correlation calculation to fully extract the independent information between different features and filter out semantically irrelevant information. This process can be seen as utilizing prominent change information from the differential features to guide an increased focus on the change regions within the temporal features, thereby reducing redundant information. This approach facilitates a more effective fusion of detailed information from the bitemporal concatenation features and change information from the differential features, thus leveraging the advantages of both strategies. It is worth noting that all MLP operations mentioned above for single-temporal features involve weight sharing. This is due to the Siamese operation's focus on spatial information while being insensitive to imaging condition differences and temporal features. The aforementioned steps can be expressed by the following formula:

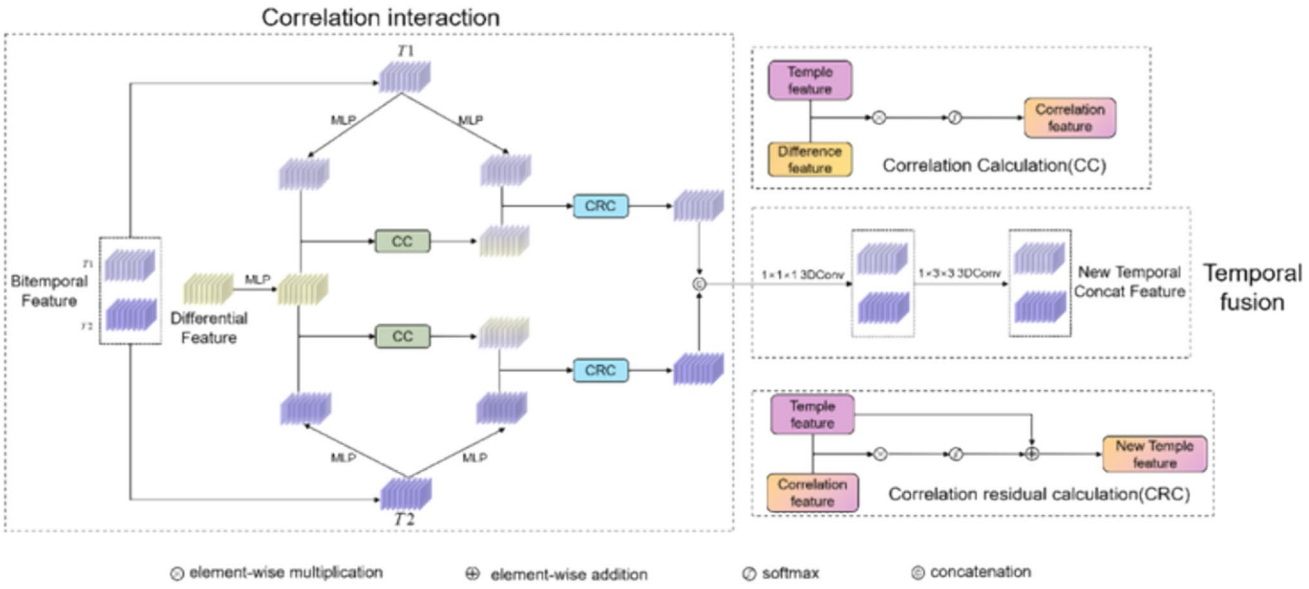

Fig. 4 Correlation fusion module. It consists of the correlation interaction region and the temporal fusion region. The correlation interaction region utilizes MLP to map the separated bitemporal features and difference features into a new smoothed representation. Subsequently, pixel-wise multiplication calculates the similarity between each temporal feature and the difference feature. Finally, the calculated correlation features are correspondingly multiplied with the original corresponding temporal features and subjected to residual summation. The temporal fusion region achieves temporal information interaction for the fused bitemporal features through the use of 3D convolution

$$f_i^{S1} = \sigma(MLP(f_i^{T1}) \otimes MLP(f_i^{Di})) \tag{2}$$

$$f_i^{S2} = \sigma(MLP(f_i^{T2}) \otimes MLP(f_i^{Di})) \tag{3}$$

$$f_i^{a1} = \sigma(f_i^{S1} \otimes MLP(f_i^{T1}) + MLP(f_i^{T1})) \tag{4}$$

$$f_i^{a2} = \sigma(f_i^{S2} \otimes MLP(f_i^{T2}) + MLP(f_i^{T2})) \tag{5}$$

where f_i^{S1}, f_i^{S2} is the correlation of bitemporal images, f_i^{a1}, f_i^{a2} is the bitemporal feature with correlation, σ is the SoftMax function, and $\otimes$ is element-wise multiplication.

2. Temporal fusion: The features from the two temporal phases are assumed to be mutually independent, which neglects the spatial and temporal interactions within the bitemporal features. Therefore, to attain a feature representation with enriched temporal change semantics, we aggregate the outputs of the bitemporal features in the channel dimension. This operation ensures that the dimensions of the input and output features of the module are consistent. Subsequently, these concatenated features are sequentially fed into spatial and temporal convolutions, facilitating the flow of information between the fused features of the bitemporal phases. In summary, CFM establishes correlations across different branches through an effective interaction mechanism for the bitemporal images, which can thereby maximize the utilization of multi-source information, suppress irrelevant information interference, and enriching image feature representation.

Global Scale Fusion Module

As HFNet extracts features at different scales in each stage, and these features internally contain various information, in order to purposefully utilize the internal information at different scales and achieve efficient multi-scale feature fusion, we have proposed the GSFM, which fuses the outputs of different scales from each stage of the Decoder, as illustrated in Fig. 5.

Specifically, to achieve preliminary feature interaction, we combine the input features from Decoder at three different scales, denoted as f_i^{De} (where i represents the i layer output of the Decoder), with the outputs from the Encoder, denoted as f_4^{En}, resulting in the fused feature f^S. Subsequently, we employ global average pooling (GAP) to compress the global spatial information into channel descriptors, to obtain features that encapsulate global channel information.

Furthermore, to consider characteristics at different scales, we selectively generate multi-scale global channel weights. We map the global channel features of various scales through MLP layers and Sigmoid function, aiming to learn the feature weights for each channel, adjust them, and reshape them into corresponding scale-specific global channel weights f_i^{GAP}. This process can be expressed as:

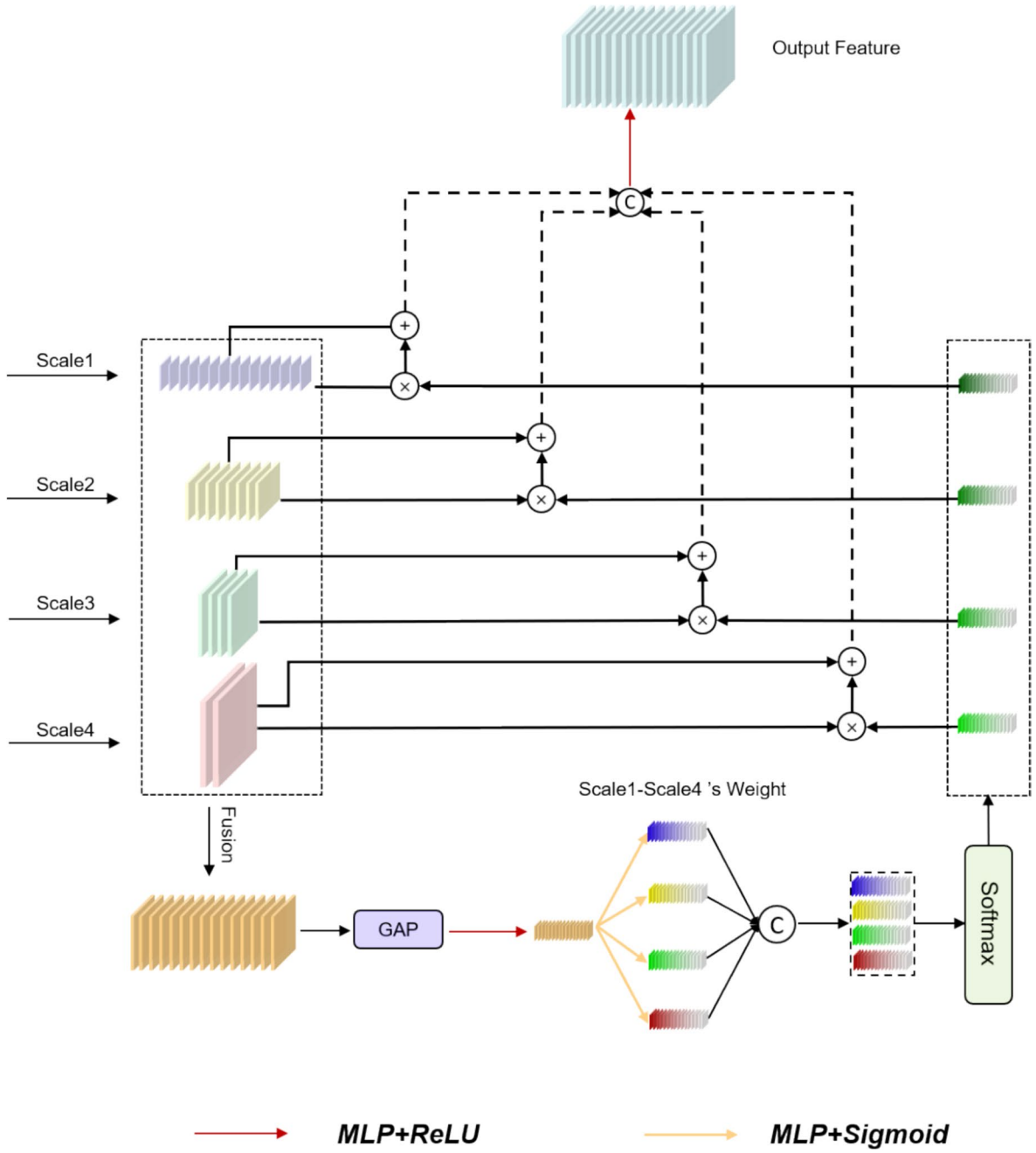

Fig. 5 Global scale fusion module. It utilizes channel-wise summation with MLP for the fusion of input multiscale information and employs GAP to extract multi-scale global channel information. Subsequently, through combinations of multiple MLPs and Sigmoid functions, the multi-scale global channel information is recalibrated at each scale to adjust channel weights accordingly. Furthermore, GSFM redistributes channel weights at different scales using SoftMax for inter-scale weight reallocation, obtaining channel weight information with scale weights. This information is then multiplied and residual connected with original features at different scales. Finally, GSFM utilizes channel-wise summation with MLP for fusion, yielding output multi-scale features

$$f_i^{GAP} = \sigma_{sig}(MLP(GAP(concat(f_4^{En}, f_1^{De}, f_2^{De}, f_3^{De})))) \quad (6)$$

where σ_{sig} represents Sigmoid, f_i^{GAP} represents channels of different scales, and i indicates four scales, $i \in \{0, 1, 2, 3\}$.

Subsequently, to achieve attention between scales and allocate reasonable inter-scale weights, the interaction and recalibration of weights f_i^{GAP} for different scales are conducted through the SoftMax function σ, resulting in new channel weights $f_i^W \in \mathbb{R}^{1\times C}$ for each scale. Following this step, the channel weights f_i^W of each scale are element-wise multiplied with their corresponding input features f_i^{De}, followed by a summation of residuals, to enhance attention

to key scales, suppress the influence of uninterested scales, thereby achieving the purpose of fully utilizing highly aggregated information. Ultimately, the four resultant features f_i are aggregated across the scale dimension and subsequently processed through an MLP to adjust the dimensionality to C, thereby obtaining the final output feature $f_{gs} \in \mathbb{R}^{H\times W\times C}$. This process can be represented as:

$$f_i = (\sigma(f_i^{GAP}) \otimes f_i^D) \oplus f_i^D \tag{7}$$

$$f_{gs} = MLP(concat(f_1, f_2, f_3, f_4)) \tag{8}$$

GSFM calculates attention within and between scales through global information embedding and adaptive recalibration. This establishes a bottom-up information exchange between layers, achieving efficient utilization of multi-scale information and full integration of multi-scale features.

Result and Discussions

To evaluate the performance of the proposed HFNet, we conducted a comparative analysis against six state-of-the-art (SOTA) models on the CDD (Lebedev et al. 2018), WHU_CD (Ji et al. 2018), and SYSU_CD (Shi et al. 2021). Additionally, we conducted ablation experiments to assess the effectiveness of our proposed STConvTrans, CFM, and GSFM modules. Furthermore, we carried out comparative experiments focusing on the feature extraction module to evaluate the advantages of STConvTrans in comparison to other well-established feature extraction modules. Finally, we discussed the network's parameter volume and computational time.

Datasets

The CDD dataset (Lebedev et al. 2018) was created using Google Earth imagery for change detection. It comprises 7 pairs of satellite images with dimensions of 4725×2700 and 4 pairs of images sized 1900×1900, featuring seasonal variations. The spatial resolution ranges from 3 to 100 cm/px. CDD encompasses diverse change features, including variations in natural objects and artificial changes. Following Lebedev et al. (2018), we partitioned all 11 pairs of images into 16,000 pairs of bitemporal images, each sized 256×256, using rotations and cropping. Among these, 10,000 pairs were assigned to the training set, while 3000 pairs each were allocated to the validation and test sets.

WHU_CD (Ji et al. 2018) is a building change detection dataset. It consists of a pair of aerial images with 32507×15354 pixels. WHU_CD documents the post-earthquake reconstruction of the Christchurch region in New Zealand, capturing aerial images in 2012 and 2016. It encompasses 12,796 buildings (increased to 16,077 in 2016) within a 20.5 square kilometer area. In line with Zhang et al. (2022), we cropped the images into non-overlapping pairs of size 224×224, subsequently dividing these bitemporal image pairs randomly into three sets: 8060 pairs for training, and 1,007 pairs each for validation and testing.

SYSU_CD (Shi et al. 2021) consists of 20,000 pairs of high-resolution aerial images captured in Hong Kong between 2007 and 2014. Each image pair has dimensions of 256×256 pixels with a spatial resolution of 0.5 m. Notably, the dataset includes various significant change types: (a) new urban buildings, (b) suburban expansion, (c) construction zones, (d) vegetation changes, (e) road expansion, and (f) marine construction. Among these, 12,000 pairs of images serve as the training set, while 4000 pairs each are allocated to the validation and test sets (Shi et al. 2021).

Evaluation Criteria and Experimental Setup

Evaluation Criteria

We assess the performance of CD results using four evaluation criteria: Precision, Recall, F1-score, and Intersection over Union (IoU).

To compute these criteria, we first calculate TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). TP represents the count of correctly detected change pixels; TN denotes the count of correctly detected unchanged pixels. Additionally, FP indicates the count of unchanged pixels erroneously detected as change, while FN represents the count of change pixels erroneously detected as unchanged. After calculating TP, TN, FP, and FN, various accuracy metrics can be computed using them, defined as follows:

Precision: The accuracy of correctly detected change pixels.
Recall: The ratio of correctly detected change pixels among all actual change pixels.
F1-score: A balanced measure that combines precision and recall for binary classification models.
IoU: The ratio of the intersection to the union of true positive change pixels and detected change pixels.

Clearly, higher precision indicates fewer false positives, while higher recall indicates fewer false negatives. However, precision and recall can yield misleading estimates in the case of class imbalance. F1-score and IoU help overcome this issue, offering better insight into overall performance.

Experimental Setup

PyTorch was used as the framework for all experiments, conducted on an NVIDIA GeForce RTX 3090 with 24 GB memory. In our experiments, end-to-end training was achieved through random initialization of all parameters. In the training of the model, Adam optimizer is used for fast convergence during training. The initial learning rate is set to 0.00005, the loss function employed is cross-entropy, the batch size is 8, and the global weight decay parameter is set to 1e-8. We also utilized a cosine annealing strategy for learning rate adjustment to facilitate rapid convergence during training. For data augmentation, only flip operations were applied to the training data.

Comparison to State-of-The-Art Change Detection Methods

In this section, we compare HFNet with six SOTA change detection methods on the CDD, WHU_CD, and SYSU_CD. To demonstrate the advanced and comprehensive nature of our approach, we selected change detection methods employing various feature extraction techniques, including three pure convolutional approaches (FC-EF (Daudt et al. 2018), FC-Siam-conc (Daudt et al. 2018), FC-Siam-diff (Daudt et al. 2018)), an attention-based method (STANet (Chen and Shi 2020a)), a Transformer-based method (SwinSUNet (Zhang et al. 2022)), and a hybrid CNN-Transformer approach (BIT_CD (Chen et al. 2021)), a method for setting up CNN Transformer dual branch lines (ACABFNet (Song et al. 2022)), a method with three encoders based on CNN (Bi-SRNet(Ding et al. 2022)), another method with three encoders based on CNN and Transformer (SCanNet(Ding et al. 2024)), and a method based on deep supervision (IFN(Zhang et al. 2020)). These methods comprehensively cover the research status, demonstrating the improvements of our approach beyond these foundations.

Introduction of Comparative Methods

FC-EF (Daudt et al. 2018) (fully convolutional early fusion) is a UNet-based fully convolutional early fusion change detection model. It consists of 4 layers of Encoder and 4 layers of Decoder, with corresponding skip connections. EF (early fusion) indicates that the bitemporal images are fused in the channel dimension before entering the network. FC-EF is a representative change detection method that combines pure CNN and EF.

FC-Siam-Conc (Daudt et al. 2018) (Fully Convolutional Siamese-Concatenation) is a UNet-based fully convolutional Siamese-concatenation change detection model. Unlike FC-EF, FC-Siam-Conc employs a pair of Siamese networks to simultaneously extract features from the bitemporal images. In the Decoder, skip connections are used to concatenate the two feature maps of equal size from the Siamese convolutional networks along the channel dimension. FC-Siam-Conc is a representative change detection method that combines pure CNN with image concatenation input.

FC-Siam-Diff (Daudt et al. 2018) (Fully Convolutional Siamese-Difference) is similar to FC-Siam-conc (Daudt et al. 2018), being a UNet-based fully convolutional Siamese-difference change detection model. It focuses on detecting differences between the bitemporal images. FC-Siam-Diff employs absolute difference values of the bitemporal images at each size for skip connections. FC-Siam-Diff is a representative change detection method that combines pure CNN with image differencing input.

STANet (Chen and Shi 2020a) utilizes ResNet18 for feature extraction and employs pyramid spatiotemporal attention modules to acquire detailed information, which can effectively alleviate false positives and false negatives caused by registration errors.

BIT_CD (Chen et al. 2021) is a change detection method that combines Transformer with CNN. BIT employs ResNet for feature extraction, and then BIT_CD employs spatial attention to transform feature maps into a compact set of tokens. The subsequent step involves using Transformer to model context for the two tokens and reproject them into pixel space. Finally, feature difference maps are computed from the two enhanced feature maps and input to CNN for pixel-level change prediction.

SwinSUNet (Zhang et al. 2022) is a pure Swin-Transformer change detection approach with a Siamese UNet structure. It utilizes Swin-Transformer as the basic unit, including Encoder, Fusion, and Decoder. SwinSUNet is the first pure Transformer network for CD tasks, enabling better extraction of global spatiotemporal information and addressing the computational complexity issue in ViT (Dosovitskiy et al. 2020).

ACABFNet (Song et al. 2022) operates in parallel through CNN and Transformer branches, employing a bidirectional fusion method to integrate local and global features in the change detection process. It can effectively utilize both fine-grained and global features of the images.

Bi-SRNet (Ding et al. 2022) uses a three-encoder-decoder architecture for change detection, where two CNN-based encoders extract semantic features and merge them through a temporal stacking method and a CNN-based deep CD unit encoder.

SCanNet (Ding et al. 2024) extracts temporal features and change representations through a three-encoder-decoder network, and then introduces Transformer to jointly learn spatio-temporal dependencies in token space.

IFN (Zhang et al. 2020) employs weight-shared parallel streams to separately extract features from bitemporal images, followed by convergence of the two streams into a

change detection stream. Outputs at each stage of this stream are output for deep supervision, culminating in the generation of change maps at the end of the stream.

Results on the CDD Dataset

In Table 1, we present quantitative comparison results of our method on the CDD dataset against currently popular methods. The results in Table 1 demonstrate that our method achieves the highest scores across all evaluation criteria.

Table 1 Quantitative evaluation of different approaches applied to the CDD dataset. The best values are indicated in italic (column wise)

Methods	F1-score	Precision	Recall	IoU
FC-EF	0.581	0.767	0.467	0.409
FC-Siam-Conc	0.648	0.815	0.538	0.480
FC-Siam-Diff	0.640	0.870	0.506	0.471
STANet	0.855	0.793	0.927	0.746
BIT_CD	0.932	0.943	0.921	0.872
SwinSUNet	0.915	0.938	0.894	0.844
ACABFNet	0.927	0.942	0.912	0.864
Bi-SRNet	0.866	0.801	0.942	0.763
SCanNet	0.907	0.892	0.921	0.829
IFN	0.898	0.874	0.922	0.814
HFNet	*0.953*	*0.956*	*0.949*	*0.909*

Compared to classical CNN methods (namely, FC-EF, FC-Siam-Conc, and FC-Siam-Diff), our method shows an improvement of 30.5% to 37.2% in F1 score and 42.9% to 50% in IoU. When compared to advanced Transformer models like BIT_CD and SwinSUNet, our method achieves a respective improvement of 2.1% and 3.8% in F1-Score and 3.7% and 6.05% in IoU.

For a more intuitive understanding of the differences between our method and classical approaches, we provide visual change detection results for each method on the CDD dataset in Fig. 6. It is worth noting that the phenomena of WHU_CD and SYSU_CD are similar, so we focus on visualizations of CDD. From the figure, it is evident that HFNet achieves superior results compared to the current SOTA methods. For instance, with the multi-scale information fusion capability of GSFM, HFNet effectively segments the complete outline of changed buildings, as shown in Fig. 6e. When it comes to small-scale changes, from f and g in Fig. 6, it can be observed that when facing smaller targets, HFNet can better distinguish change areas compared to other popular methods. Additionally, due to the Siamese difference branch of HFNet, which has the ability to eliminate phenological interference, HFNet exhibits strong robustness as shown in a of Fig. 6. As depicted in Fig. 6c and f, shadows on buildings lead to missed detections and false alarms, which are significantly mitigated by HFNet, benefiting from the Siamese characteristics in STConvTrans.

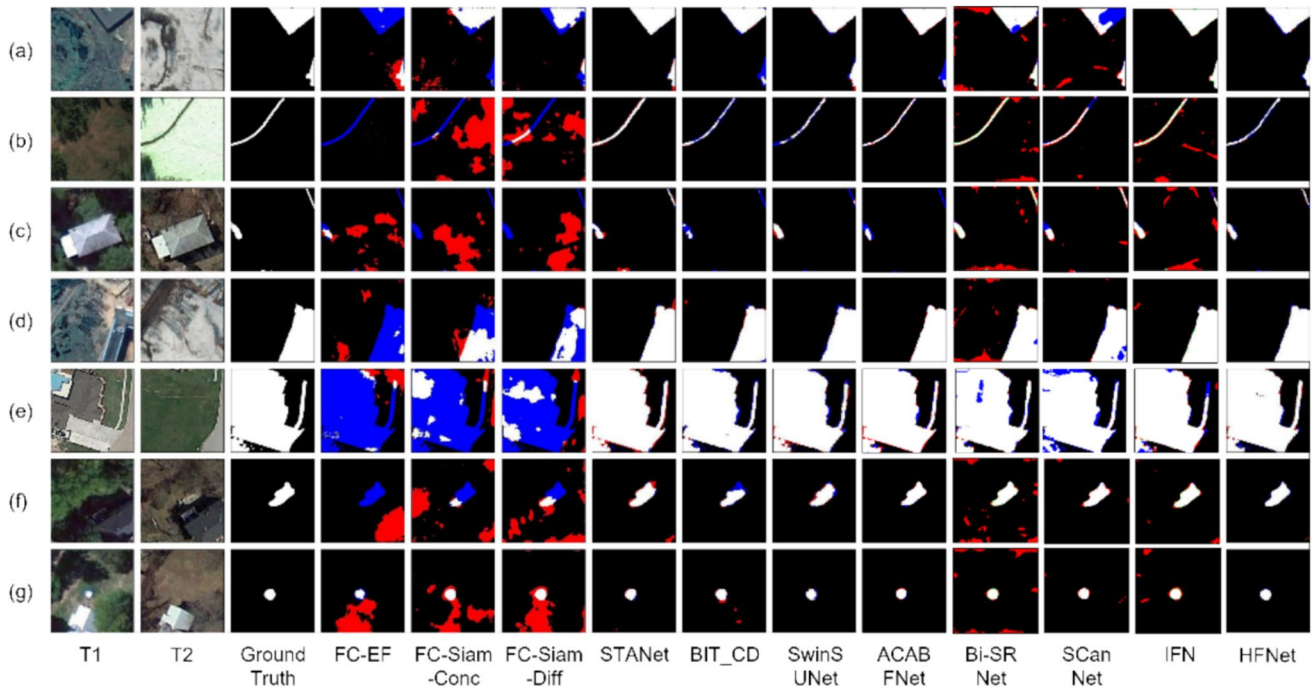

Fig. 6 Visualization results of comparison algorithms and our algorithm on the CDD dataset, where white represents true positives, black represents true negatives, red represents false positives, and blue represents false negatives. **a**–**g** are seven sets of images selected from the CDD test set

Results on the WHU_CD Dataset

Table 2 displays the quantitative comparison results of our method against current popular methods on the WHU_CD dataset. From the table, it can be seen that our method achieves the best results across the majority of categories. Furthermore, in terms of average criteria, our method outperforms the second-best model SwinSUNet by 1.5% in F1-score and 2.9% in IoU. Although FC-EF and STANet achieve high recalls of 99.8%–99.9%, their F1-scores and IoU are significantly lower than those of our proposed method. This could be attributed to the presence of a substantial number of unchanged regions in the WHU-CD dataset, leading to a wider range of false positives.

Table 2 Quantitative evaluation of different approaches applied to the WHU_CD dataset. The best values are indicated in italic (column wise)

Methods	F1-score	Precision	Recall	IoU
FC-EF	0.895	0.817	0.998	0.810
FC-Siam-Conc	0.807	0.854	0.765	0.676
FC-Siam-Diff	0.888	0.864	0.913	0.799
STANet	0.854	0.745	*0.999*	0.744
BIT_CD	0.907	0.917	0.997	0.829
SwinSUNet	0.929	0.900	0.960	0.868
ACABFNet	0.869	0.901	0.839	0.769
Bi-SRNet	0.910	0.837	0.996	0.834
SCanNet	0.928	0.876	0.978	0.860
IFN	0.921	0.883	0.962	0.854
HFNet	*0.946*	*0.926*	0.967	*0.897*

Figure 7 illustrates the change detection visual results for each method on the WHU_CD dataset. It is evident that WHU_CD focuses solely on detecting changes in buildings. In comparison to the current popular methods, HFNet accurately identifies changed regions and their boundaries for both small clusters and large newly constructed buildings, as demonstrated in Fig. 7b, c, and g. This is attributed to the powerful multi-scale feature extraction capability of STConvTrans and the multi-scale fusion capability of GSFM. In g of Fig. 7, it can be observed that the shadows of the buildings have a minimal interference with HFNet's ability to extract building edges. This demonstrates the effectiveness of our method for datasets concerning building change.

Results on the SYSU_CD Dataset

Table 3 presents the results of various methods on the SYSU-CD dataset. It is observed that our method achieves the best results across most of the categories. Compared to the second-best model ACABFNet, our method outperforms with a 0.4% higher F1-score and a 0.4% higher IoU on average. However, it can be observed that the change detection

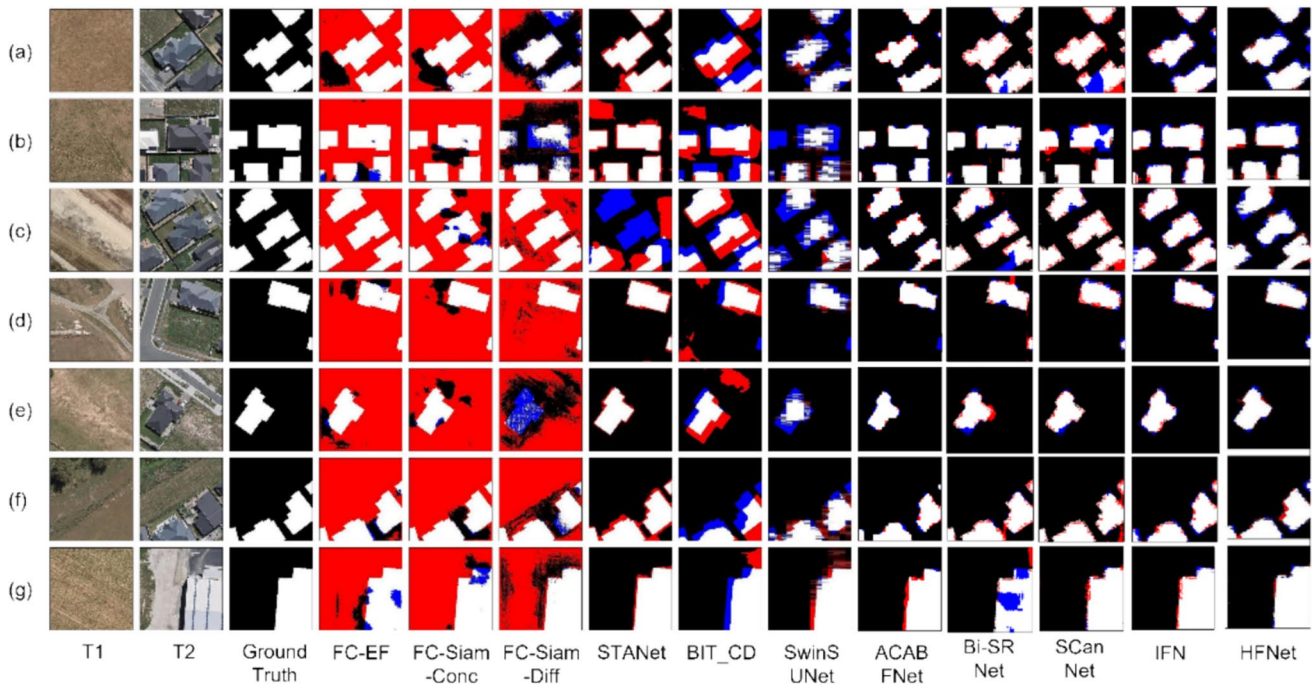

Fig. 7 Visualization results of comparison algorithms and our algorithm on the WHU_CD dataset, where white represents true positives, black represents true negatives, red represents false positives, and blue represents false negatives. **a**–**g** are seven sets of images selected from the WHU_CD test set

Table 3 Quantitative evaluation of different approaches applied to the SYSU_CD dataset. The best values are indicated in italic (column wise)

Methods	F1-score	Precision	Recall	IoU
FC-EF	0.751	0.743	0.758	0.601
FC-Siam-Conc	0.764	0.825	0.710	0.618
FC-Siam-Diff	0.726	*0.891*	0.612	0.570
STANet	0.754	0.706	0.810	0.606
BIT_CD	0.765	0.916	0.716	0.619
SwinSUNet	0.758	0.802	0.719	0.611
ACABFNet	0.784	0.815	0.756	0.646
Bi-SRNet	0.738	0.634	0.884	0.585
SCanNet	0.777	0.707	0.863	0.635
IFN	0.755	0.712	0.805	0.607
HFNet	*0.788*	0.828	0.752	*0.650*

capability of HFNet in SYSU_CD is evidently lower than that in CDD and WHU_CD. This might be attributed to the complexity of SYSU_CD dataset, which includes 6 types of complex change, and the Hong Kong region characterized by high-density urban structures and intricate backgrounds. This complexity introduces challenges such as building shadows, reflections, and occlusions, thus increasing the difficulty of change detection.

Figure 8 showcases the change detection visual results for each method on the SYSU_CD dataset. With a diverse range of target types in this dataset, these visualizations confirm the consistency of HFNet's performance with the results in Table 3. HFNet generates change maps with finer details, which are closer to the reference. Due to the ability of STConvTrans to extract and process global features, although all image pairs presented in Fig. 8 exhibit noticeable imaging differences, and HFNet demonstrates strong robustness in handling these variations. Figure 8a involves changes in roads and vegetation, Fig. 8b and c illustrate changes caused by boats, and Fig. 8f and g are related to building changes. From these images, we can see that compared to other methods, HFNet generates fewer detection errors, exhibits its strong change detection capabilities for multiscale features, and predicts more accurate edges. In particular, in Fig. 8e, it can be seen that HFNet can effectively mitigate the impact of shadows in the bitemporal images and accurately detect change areas.

In summary, both quantitative and qualitative results affirm the applicability and superiority of HFNet, which has higher accuracy and lower false positives than other methods in the vast majority of cases.

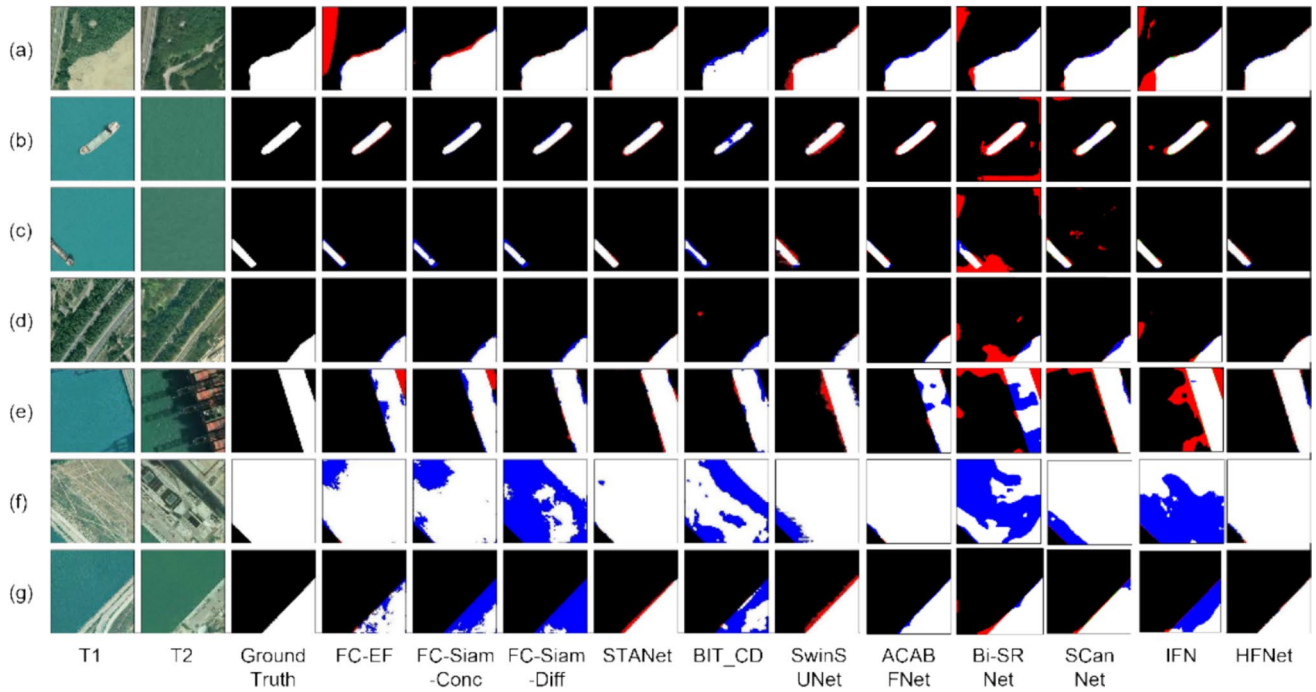

Fig. 8 Visualization results of comparison algorithms and our algorithm on the SYSU_CD dataset, where white represents true positives, black represents true negatives, red represents false positives, and blue represents false negatives. **a**–**g** are seven sets of images selected from the SYSU_CD test set

Mutual Enhancement of Semantic and Differential Features

To verify the effectiveness of the dual-branch structure in HFNet, experiments were conducted to analyze the mutual enhancement between the bitemporal concatenation branch and Siamese differencing branch. The tested schemes included utilizing only the Siamese differencing branch, only the temporal concatenation branch, and integrating both branches. In all experiments, multi-scale feature fusion methods involved stacking features across channels and using MLP to restore to the desired number of channels. This fusion method is defined in this paper as Cat^{scale}. The experimental results are shown in Table 4.

As shown in Table 4, the performance of the Siamese difference branch scheme alone is significantly inferior to that of the other schemes. One reason for this is that, although the difference scheme effectively highlights changes, it also retains a large amount of pseudo-change information, which prevents the model from accurately identifying true changes. When comparing Scheme 1 with Scheme 2, it is evident that utilizing solely the bitemporal concatenation branch yields significantly superior results compared to relying exclusively on the Siamese difference branch. However, it should be noted that although the Siamese difference branch has its limitations, it can still generate valuable insights by emphasizing change areas. Most notably, Scheme 3 demonstrated superior performance, indicating that the dual-branch fusion strategy proposed in this study effectively harnesses the semantic features extracted from the bitemporal concatenation branch and the difference features obtained from the Siamese difference branch. This approach effectively addresses the issue of insufficient change information extraction and the difficulty in discerning pseudo-changes.

Ablation Study

To validate the rationality and effectiveness of the modules in HFNet, we conducted ablation experiments on the CDD dataset. We utilized the four aforementioned metrics to evaluate the impact of different components on change detection performance, and the ablation results are summarized in Table 5. In these experiments, we explored the effects of the three key components: STConvTrans, CFM, and GSFM.

Experiment 6 represents the complete HFNet change detection approach. In Experiment 1, the network feature extraction module employs the advanced universal module Swin Transformer. For the fusion method between branches, this experiment stacks features from different branches across the channel dimension and uses MLP to restore them to be consistent with the bitemporal features. This fusion method is defined as $Cat^{feature}$. Consistent with the branch comparison experiment, this experiment employs Cat^{scale} for multi-scale feature fusion. Subsequently, Experiment 2 modified the network feature extraction module to our proposed STConvTrans. In Experiments 3, 4, and 5, the change detection networks were based on Experiment 6 with GSFM, STConvTrans, and CFM replaced by Cat^{scale}, Swin-Transformer, and $Cat^{feature}$, respectively, to assess the influence of GSFM, STConvTrans, and CFM on the change detection capability of HFNet. The results of the ablation experiments are summarized as follows.

GSFM

Comparing the results of Experiment 3 and Experiment 6, using GSFM instead of Cat^{scale} leads to improvements in all evaluation criteria. Compared to Cat^{scale}, GSFM employs differential computation on information from different scales, enabling the network to have a more comprehensive

Table 4 Quantitative evaluation of the different input schemes using the CDD dataset. The best values are indicated in italic (column wise)

ID	bitemporal concatenation branch	Siamese difference branch	F1-score	Precision	Recall	IoU
1	/	√	0.766	0.871	0.684	0.621
2	√	/	0.914	0.940	0.890	0.842
3	√	√	*0.929*	*0.944*	*0.915*	*0.868*

Table 5 Quantitative evaluation of the ablation study results using the CDD dataset. The best values are indicated in italic (column wise)

ID	STConvTrans	CFM	GSFM	F1-score	Precision	Recall	IoU
1	/	/	/	0.929	0.944	0.915	0.868
2	√	/	/	0.932	0.944	0.921	0.873
3	√	√	/	0.944	0.954	0.935	0.895
4	/	√	√	0.941	0.949	0.932	0.888
5	√	/	√	0.945	0.945	0.945	0.896
6	√	√	√	*0.953*	*0.956*	*0.949*	*0.909*

capability in multi-scale feature extraction. Specifically, F1-Score increases by 0.9% and IoU increases by 1.4%. This indicates that simple summation fusion of features from different stages is not the optimal fusion strategy due to the varying content of features at different stages. In contrast, our GSFM can effectively extract differentiated information inherent in different scales and establish an efficient joint mechanism.

CFM

The comparison between Experiment 5 and Experiment 6 reveals a significant enhancement in change detection performance with the inclusion of CFM. On the CDD dataset, the increase in F1 and IoU amounts to 0.8% and 1.3%, respectively. In HFNet, different branches extract different features, but they overlap to some extent. However, $Cat^{feature}$, which simply concatenates branch features in the channel dimension, might overlook the distinctive characteristics of different branch features, potentially resulting in suboptimal feature combination. In contrast, CFM calculates correlations among different features to extract temporally uncorrelated features and filter out irrelevant semantic changes. This underscores the significant role of CFM in more effectively fusing features from different branches.

STConvTrans

Compared to Experiment 4, Experiment 6 shows a 1.2% improvement in F1-Score and a 2.1% improvement in IoU. This indicates that, compared to the Swin Transformer, the spatiotemporal convolution in our proposed STConvTrans contributes to preserving the global differential information extraction capability while achieving temporal interaction and enhancing the local feature extraction ability. Notably, the performance improvement from STConvTrans in Experiment 6 is significantly higher than that in Experiment 2. This indicates that effective inter-branch fusion and multi-scale fusion can assist STConvTrans in more efficient feature representation.

Comparison to State-of-The-Art Feature Extraction Module

To validate the advancement of STConvTrans, this section compares STConvTrans with several state-of-the-art feature extraction modules, including Vision Transformer (ViT) (Dosovitskiy et al. 2020), Convolutional Vision Transformer (CvT) (Wu et al. 2021), Swin-Transformer (SwinT) (Liu et al. 2021), and Video-Swin-Transformer (3DSwin-T) (Liu et al. 2022b). These feature extraction modules have been widely used in the field of change detection (Chen et al. 2021; Huang et al. 2022; Zhang et al. 2022). To ensure fairness, consistent hyperparameters and evaluation criteria were employed, and the experiments were conducted on the CDD dataset using the HFNet framework. The comparative results are presented in Table 6.

From Table 6, it is evident that our proposed STConvTrans outperforms the other feature extraction modules for all evaluation criteria. Specifically, compared to the state-of-the-art 2D Transformer-based approach (Swin-T), our STConvTrans achieves improvements of 1.2% in F1-Score and 2.1% in IoU. This highlights STConvTrans's ability to enhance change detection accuracy by combining the locally perceptive capabilities of CNN with the global attention of Transformers. In comparison with the approach that combines 2D CNN with Transformers (CvT), our STConvTrans shows increments of 3.1% in F1-Score and 5.4% in IoU. This demonstrates that STConvTrans surpasses the simple concatenation of CNN and Transformer methods. Furthermore, when compared to the state-of-the-art 3D Transformer-based approach incorporating spatiotemporal features (3DSwin-T), our STConvTrans achieves improvements of 1.5% in F1-Score and 2.5% in IoU. This confirms that STConvTrans can not only effectively leverage the advantages of both CNN's local receptive fields and Transformer's global attention but also fully exploit the spatiotemporal features of bitemporal imagery.

Complexities of Feature Extraction Module

Furthermore, in this section, we validate the efficiency of our proposed STConvTrans by comparing its parameter count and training time per epoch. Considering both Tables 6 and 7, we observe that STConvTrans achieves a 4.1% and 3.1% improvement in F1-Score over ViT and CvT, respectively, while requiring significantly less training time. When compared to 3DSwinT, STConvTrans maintains a 1.5% F1-Score lead while remaining considerably lighter in terms of parameters and training time. This implies that STConvTrans can achieve accuracy improvements while keeping the network lightweight.

Although the parameter counts and training time of STConvTrans are slightly higher than those of Swin-T, this is

Table 6 Quantitative comparison of STConvTrans change detection results on CDD with other SOTA feature extraction modules. The best values are indicated in italic (column wise)

Feature extraction modules	F1-score	Precision	Recall	IoU
ViT	0.912	0.924	0.901	0.838
CvT	0.922	0.933	0.911	0.855
Swin-T	0.941	0.949	0.932	0.888
3DSwin-T	0.938	0.945	0.931	0.884
STConvTrans	*0.953*	*0.956*	*0.949*	*0.909*

Table 7 Quantitative comparison of parameter quantity and calculation time of STConvTrans with other SOTA feature extraction modules. The best values are indicated in italic (column wise)

Feature extraction modules	Pram.(M)	Time(s/epoch)
ViT	303.51	1890
CvT	81.46	1102
Swin-T	*81.23*	*517*
3DSwin-T	131.46	901
STConvTrans	107.67	638

attributed to the integration of CNN in STConvTrans to efficiently capture local features. CNN modules typically come with a certain parameter count and computational complexity. However, considering that STConvTrans introduces an effective approach for spatiotemporal feature extraction, surpassing Swin-T's performance by dynamically attending to neighboring elements and enhancing local feature extraction, the computational cost of STConvTrans remains reasonable.

Conclusion

In this study, we proposed a novel bitemporal-difference fusion method, HFNet, for change detection. Leveraging the complementary nature of bitemporal concatenated features and difference features, we established a dual-branch network for bitemporal-difference fusion, enabling the simultaneous extraction of detailed bitemporal concatenated features and change information. To ensure effective interaction between the bitemporal and difference branches, fusion between them was performed at each stage, with the fused features fed into the bitemporal concatenated branch for further computation. Subsequently, we designed CFM to effectively couple bitemporal concatenated features and difference features, constructing a more robust inter-branch connection mechanism. In the decoder, we introduced a multiscale semantic information fusion module, GSFM, which allowed features of different scales to establish mutual dependencies between channels and scales. GSFM adaptively calculated weights and bridged the semantic gap across scales. In addition, recognizing the importance of both local and global contextual features in remote sensing image change detection, we proposed a hybrid CNN-Transformer feature extraction module, STConvTrans, to capture and integrate the local detailed features and long-distance relationships between bitemporal images.

Nonetheless, this study has limitations. As our experiments involve multiple change detection branches, HFNet's parameter count is relatively higher than other change detection networks. However, considering HFNet's substantial performance advantages over SOTA methods through the effectively integrating bitemporal concatenated and difference features, its computational cost is acceptable. In future work, we plan to further optimize our network structure to reduce the parameter count and computational requirements.

Acknowledgements The authors would also like to thank the editors and anonymous reviewers for the insightful suggestions, which significantly improved the quality of this article.

Funding The research leading to these results received funding from the major scientific and technological projects of Yunnan Province under grant 202202AD080010, and the National Natural Science Foundation of China under Grants 42471391 and 42271328.

Data Availability Data available within the article or its supplementary materials.

Declarations

Competing Interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Bao T, Fu C, Fang T, Huo H (2020) PPCNET: a combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection. IEEE Geosci Remote Sens Lett 17:1797–1801

Chen H, Shi Z (2020) A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing 12:1662

Chen J, Yuan Z, Peng J, Chen L, Huang H, Zhu J, Liu Y, Li H (2020) DASNet: dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. IEEE J Sel Top Appl Earth Obs Remote Sensi 14:1194–1206

Chen H, Qi Z, Shi Z (2021) Remote sensing image change detection with transformers. IEEE Trans Geosci Remote Sens 60:1–14

Daudt RC, Le Saux B, Boulch A (2018) Fully convolutional siamese networks for change detection. In: 2018 25th IEEE international conference on image processing (ICIP), pp 4063–4067. https://doi.org/10.1109/ICIP.2018.8451652

De Bem PP, de Carvalho JOA, Fontes GR, Trancoso Gomes RA (2020) Change detection of deforestation in the Brazilian Amazon using landsat data and convolutional neural networks. Remote Sens 12:901

Ding L, Guo H, Liu S, Mou L, Zhang J, Bruzzone L (2022) Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/tgrs.2022.3154390

Ding L, Zhang J, Guo H, Zhang K, Liu B, Bruzzone L (2024) Joint spatio-temporal modeling for semantic change detection in remote sensing images. IEEE Trans Geosci Remote Sens 62:1–14. https://doi.org/10.1109/tgrs.2024.3362795

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Eftekhari A, Samadzadegan F, Javan FD (2023) Building change detection using the parallel spatial-channel attention block and edge-guided deep network. Int J Appl Earth Obs Geoinf 117:103180

Fang B, Pan L, Kou R (2019) Dual learning-based siamese framework for change detection using bi-temporal VHR optical remote sensing images. Remote Sensing 11:1292

Fang S, Li K, Shao J, Li Z (2021) SNUNet-CD: a densely connected Siamese network for change detection of VHR images. IEEE Geosci Remote Sens Lett 19:1–5

Feng Y, Jiang J, Xu H, Zheng J (2023) Change detection on remote sensing images using dual-branch multilevel intertemporal network. IEEE Trans Geosci Remote Sens 61:1–15

Hou S, Zhang G, Cui H, Li X, Chen Y, Li H, Ma X (2023) Stable prototype guided single-temporal supervised learning for change detection and extraction of building. IEEE Trans Geosci Remote Sens 61:1–22

Huang X, Zhang L, Zhu T (2013) Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7:105–115

Huang X, Dong M, Li J, Guo X (2022) A 3-D-Swin transformer-based hierarchical contrastive learning method for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–15

Ji S, Wei S, Lu M (2018) Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans Geosci Remote Sens 57:574–586

Jiang J, Xiang J, Yan E, Song Y, Mo D (2022) Forest-CD: forest change detection network based on VHR images. IEEE Geosci Remote Sens Lett 19:1–5

Lebedev M, Vizilter YV, Vygolov O, Knyaz VA, Rubis AY (2018) Change detection in remote sensing images using conditional adversarial networks. Int Arch Photogramm Remote Sens Spat Inf Sci 42:565–571

Li Q, Huang X, Wen D, Liu H (2017) Integrating multiple textural features for remote sensing image change detection. Photogramm Eng Remote Sens 83:109–121

Li Q, Zhong R, Du X, Du Y (2022) TransUNetCD: a hybrid transformer network for change detection in optical remote-sensing images. IEEE Trans Geosci Remote Sens 60:1–19

Liu Y, Pang C, Zhan Z, Zhang X, Yang X (2020) Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geosci Remote Sens Lett 18:811–815

Liu M, Chai Z, Deng H, Liu R (2022) A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15:4297–4306

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022b) Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3202–3211. https://doi.org/10.1109/CVPR52688.2022.00320

JST C (2015) Change detection from a street image pair using cnn features and superpixel segmentation. In Procedings of the British Machine Vision Conference, pp 61.

Peng D, Zhang Y, Guan H (2019) End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sensing 11:1382

Shi Q, Liu M, Li S, Liu X, Wang F, Zhang L (2021) A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE Trans Geosci Remote Sens 60:1–16

Song A, Choi J, Han Y, Kim Y (2018) Change detection in hyperspectral images using recurrent 3D fully convolutional networks. Remote Sensing 10:1827

Song L, Xia M, Weng L, Lin H, Qian M, Chen B (2022) Axial cross attention meets CNN: bibranch fusion network for change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16:21–32

Tu L, Li J, Huang X (2021) High-resolution land cover change detection using low-resolution labels via a semi-supervised deep learning approach-2021 IEEE data fusion contest track MSD. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp 2058–2061. https://doi.org/10.1109/IGARSS47720.2021.9555033

Wang W, Tan X, Zhang P, Wang X (2022) A CBAM based multiscale transformer fusion approach for remote sensing image change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15:6817–6825

Wen D, Huang X, Bovolo F, Li J, Ke X, Zhang A, Benediktsson JA (2021) Change detection from very-high-spatial-resolution optical remote sensing images: methods, applications, and future directions. IEEE Geoscience and Remote Sensing Magazine 9:68–101

Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L (2021) Cvt: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 22–31. https://doi.org/10.1109/ICCV48922.2021.00009

Yuan J, Wang L, Cheng S (2022) STransUNet: a siamese TransUNet-based remote sensing image change detection network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15:9241–9253

Zhan Y, Fu K, Yan M, Sun X, Wang H, Qiu X (2017) Change detection based on deep siamese convolutional network for optical aerial images. IEEE Geosci Remote Sens Lett 14:1845–1849

Zhang W, Lu X (2019) The spectral-spatial joint learning for change detection in multispectral imagery. Remote Sensing 11:240

Zhang M, Xu G, Chen K, Yan M, Sun X (2018) Triplet-based semantic relation learning for aerial remote sensing image change detection. IEEE Geosci Remote Sens Lett 16:266–270

Zhang C, Yue P, Tapete D, Jiang L, Shangguan B, Huang L, Liu G (2020) A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J Photogramm Remote Sens 166:183–200. https://doi.org/10.1016/j.isprsjprs.2020.06.003

Zhang C, Wang L, Cheng S, Li Y (2022) SwinSUNet: pure transformer network for remote sensing image change detection. IEEE Trans Geosci Remote Sens 60:1–13

Zhang K, Zhao X, Zhang F, Ding L, Sun J, Bruzzone L (2023) Relation changes matter: cross-temporal difference transformer for change detection in remote sensing images. IEEE Trans Geosci Remote Sens 61:1–15. https://doi.org/10.1109/tgrs.2023.3281711

Zhang X, Cheng S, Wang L, Li H (2023) Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection. IEEE Trans Geosci Remote Sens 61:1–15

Zhao W, Mou L, Chen J, Bo Y, Emery WJ (2019) Incorporating metric learning and adversarial network for seasonal invariant change detection. IEEE Trans Geosci Remote Sens 58:2720–2731

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2014) Object detectors emerge in deep scene CNNs. arXiv preprint arXiv:1412.6856