remote sensing

MDPI

Article

Frequency–Spatial–Temporal Domain Fusion Network for Remote Sensing Image Change Captioning

Shiwei Zou [1], Yingmei Wei [1], Yuxiang Xie [1,*] and Xidao Luan [2]

[1] Department of System Engineering, National University of Defense Technology, Changsha 410000, China; zsw0915@nudt.edu.cn (S.Z.); weiyingmei@nudt.edu.cn (Y.W.)
[2] Department of Computer Science, Changsha University, Changsha 410000, China; xidaoluan@ccsu.cn
* Correspondence: yxxie@nudt.edu.cn

Abstract: Remote Sensing Image Change Captioning (RSICC) has emerged as a cross-disciplinary technology that automatically generates sentences describing the changes in bi-temporal remote sensing images. While demonstrating significant potential for urban planning, agricultural surveillance, and disaster management, current RSICC methods exhibit two fundamental limitations: (1) vulnerability to pseudo-changes induced by illumination fluctuations and seasonal transitions and (2) an overemphasis on spatial variations with insufficient modeling of temporal dependencies in multi-temporal contexts. To address these challenges, we present the Frequency–Spatial–Temporal Fusion Network (FST-Net), a novel framework that integrates frequency, spatial, and temporal information for RSICC. Specifically, our Frequency–Spatial Fusion module implements adaptive spectral decomposition to disentangle structural changes from high-frequency noise artifacts, effectively suppressing environmental interference. The Spatia–Temporal Modeling module is further developed to employ state-space guided sequential scanning to capture evolutionary patterns of geospatial changes across temporal dimensions. Additionally, a unified dual-task decoder architecture bridges pixel-level change detection with semantic-level change captioning, achieving joint optimization of localization precision and description accuracy. Experiments on the LEVIR-MCI dataset demonstrate that our FSTNet outperforms previous methods by 3.65% on BLEU-4 and 4.08% on CIDEr-D, establishing new performance standards for RSICC.

Keywords: image change captioning; remote sensing; frequency; state space model; change detection

Academic Editors: Min Huang, Changjiang Xiao, Nengcheng Chen, Runze Li and Orhan Altan

Received: 17 February 2025
Revised: 11 April 2025
Accepted: 16 April 2025
Published: 19 April 2025

Citation: Zou, S.; Wei, Y.; Xie, Y.; Luan, X. Frequency–Spatial–Temporal Domain Fusion Network for Remote Sensing Image Change Captioning. *Remote Sens.* **2025**, *17*, 1463. https://doi.org/10.3390/rs17081463

1. Introduction

Remote Sensing Image Change Captioning (RSICC) provides crucial technical support for multi-temporal earth observation and automated change detection analysis, facilitating a comprehensive understanding of persistent surface changes. This task is dedicated to interpreting the semantic changes between bi-temporal remote sensing (RS) images and expressing them in natural language. As an integrative framework combining remote sensing image captioning with semantic change detection, it establishes the most user-friendly interpretation paradigm, with extensive applications in urban planning, agricultural monitoring, disaster response operations, and military damage assessment [1–5].

Depending on the desired output from the change detector, change detection (CD) tasks can be categorized into binary change detection (BCD) and semantic change detection (SCD). BCD only localizes the changed regions in bi-temporal RS images while ignoring the semantic categories. The semantic change masks produced by SCD are not easily

interpretable for further analysis. In Comparison, RSICC interprets changes between two RS images at a semantic level. It uses natural language to describe the changed regions as well as the attributes and relationships of objects in the images, making it more user-friendly for interpretation. Similar to image captioning tasks, the classic architecture of RSICC follows an encoder–decoder framework. During the feature extraction stage, RSICC identifies and understands the changed regions and their attributes in bi-temporal RS images of the same area. In the generation stage, it transforms change features into textual features and generates descriptive sentences about the changes. Hoxha et al. [6,7] were the first to propose this emerging change interpretation task, designing CNN-RNN and CNN-SVM architectures to represent changes through image-level and feature-level representations, which then guide the decoder in predicting words step-by-step. Liu et al. [8] later designed a dual-branch Siamese Transformer model, further establishing the baseline and benchmark datasets for RSICC, marking a milestone in this field. Recent advancements [4,5,9–11] introduced attention mechanisms to extract multi-scale change features and localize change-related characteristics, further improving the accuracy of RSICC.

Despite the satisfactory performance of current methods [6–11], RSICC still faces several challenges: First, as shown in Figure 1a,c, existing approaches fail to account for the distribution gap between image pairs caused by acquisition conditions such as illumination effects and shadows of trees, lacking the ability to distinguish genuine surface changes of interest from pseudo-changes induced by environmental factors. Second, the temporal attributes of bi-temporal RS image pairs are often overlooked. While bi-temporal image pairs resemble two frames extracted from a video sequence with inherent temporal order, few methods effectively capture temporal characteristics or achieve satisfactory spatial-temporal feature fusion. Third, as illustrated in Figure 1b, the absence of change masks as guidance makes it difficult to distinguish newly emerged buildings due to their color similarity with surrounding areas. The inherent connection between change detection and change captioning tasks remains underutilized, as standalone textual descriptions cannot achieve comprehensive interpretation. Frequency domain features have demonstrated significant applications in both RS change detection and image segmentation, revealing structural and texture information that enhances model accuracy [12–14]. The design of low-pass filters also contributes to suppressing high-frequency noise and enhancing model robustness. Consequently, by processing the frequency components of bi-temporal image pairs, we can precisely locate regions of interest while mitigating interference from pseudo-changes. Moreover, the frequency domain facilitates the capture of dependencies across deep feature channels, enabling effective extraction of bi-temporal depth change features for accurate captioning. The State Space Model (SSM) Mamba [15] introduces a specialized scanning mechanism for sequence data processing, enabling sequential scanning of text or image tokens to capture sequence characteristics. Leveraging this temporal scanning mechanism, SSM can perform Spatial–Temporal Modeling of change features with relatively low complexity, comprehensively identifying spatial changes while understanding their temporal development and evolution.

Based on the aforementioned analysis, this paper proposes the Frequency–Spatial–Temporal domain Fusion Network (FST-Net) to address current challenges. The framework is composed of four essential modules: a multi-scale feature extractor, a Frequency–Spatial Fusion module (FSF), a Spatial–Temporal Modeling module (STM), and a dual-task decoder (DTD). After extracting spatial features from bi-temporal images through the feature extractor, the FSF adaptively extracts frequency domain features from the bi-temporal images and fuses them with spatial features to obtain noise-suppressed fusion features. The STM performs three types of sequential scans on the fused features to conduct Spatial–Temporal Modeling, comprehensively identifying spatial changes while understanding

their temporal development and evolution. Subsequently, the multi-level features incorporating frequency domain information are fed into the change detection decoder to generate change masks. The feature maps containing temporal, frequency, and spatial information are then element-wise multiplied with the change mask to further filter out irrelevant change areas before being input into the change captioning decoder to generate change captions. This approach enables accurate change captions and eliminates the interference of pseudo-changes.

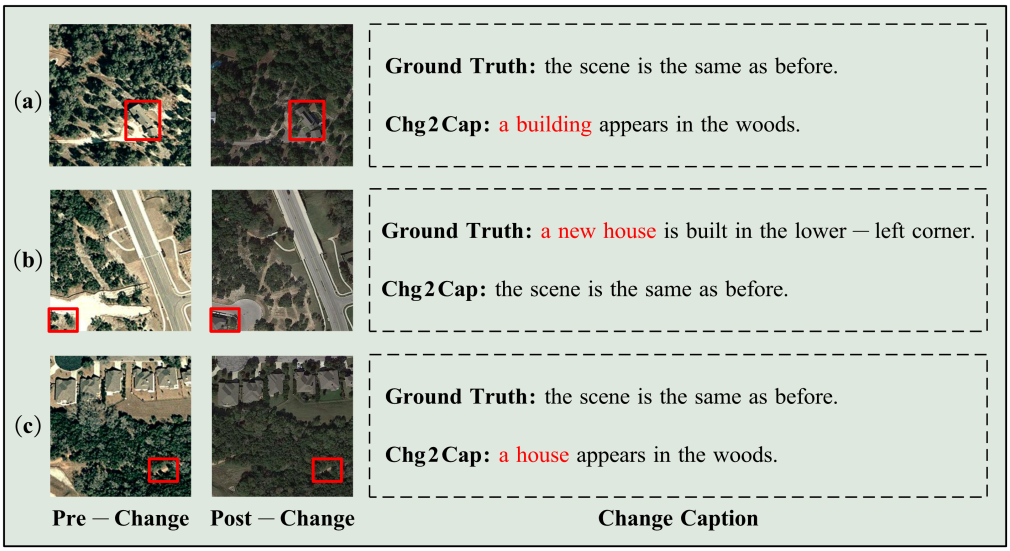

Figure 1. Challenges faced by RSICC in describing semantic changes between bi-temporal images in different scene situations: (**a**) seasonal and illumination changes, (**b**) buildings with land color similarity, and (**c**) shadows produced by trees and illumination, where the red box is marking the change target location. Red words highlight change objects misidentified by Chg2Cap [10].

In summary, our contributions include the following:

1. We propose the Frequency–Spatial–Temporal Domain Fusion Network (FST-Net) to accurately locate the changes of interest and generate accurate and comprehensive change interpretations.
2. We introduce the Frequency–Spatial Fusion module (FSF) and the Spatial–Temporal Modeling module (STM). By adaptively filtering high-frequency features containing noise and pseudo-changes, these modules suppress the interference of pseudo-changes on the model. They also fully capture the temporal features of bi-temporal RS image pairs, enhancing the model's temporal understanding and spatial perception capabilities.
3. We construct a dual-task change interpretation framework that integrates change detection and change captioning tasks. This framework can simultaneously generate pixel-level and semantic-level change results, addressing the limitation of single-output change captioning tasks and further improving the performance of the change captioning model.

Section 2 systematically surveys advancements across three interrelated domains: Remote Sensing Image Change Captioning, remote sensing change detection, and natural change captioning. Section 3 elaborates on the architectural innovations of our proposed FST-Net. Section 4 benchmarks FST-Net against the existing methods on the LEVIR-MCI dataset and the WHU-CDC dataset, validating its superiority through quantitative metrics. Section 5 concludes by summarizing FST-Net's core contributions, analyzing its current limitations, and outlining potential future research directions.

2. Related Works

2.1. Remote Sensing Image Change Captioning

RSICC represents an emerging branch of vision-language understanding and generation tasks in the RS domain, focusing on automatically identifying and describing changes over time from bi-temporal RS image pairs. With the advancement of RS technologies and deep learning, RSICC has evolved into an active research field. From early encoder–decoder architectures based on traditional machine learning to more complex network structures utilizing attention mechanisms and Transformer models, these approaches have demonstrated increasingly superior performance and accuracy in RSICC tasks.

Hoxha et al. [7] pioneered the implementation of RSICC using CNN-RNN and CNN-SVM architectures while establishing two small-scale RSICC datasets, LEVIR CCD and DUBAI CCD, for further research. Liu et al. [8] developed RSICCFormer based on Transformer, leveraging its capability to model change features and filter irrelevant attention results for generating change captions. Subsequently, Liu et al. [9] conducted further explorations by designing PSNet, which employs multiple stacked difference-aware layers and scale-aware enhanced attention modules to capture multi-scale change features. They [16] also decoupled the RSICC task into two sub-tasks, "whether changes occur" and "what changes occur", thereby generating more accurate change captions. Chang et al. [10] designed an attention encoder incorporating self-attention modules and residual blocks to extract visual embeddings and dynamically locate change-related features. Zhou et al. [17] constructed a single-stream extractor network (SEN) pretrained on bi-temporal RS images, achieving lower computational costs. The SFE and CAGD modules of SEN enable better modeling of change features.

Current methods primarily focus on spatial dimension changes, with limited consideration of the temporal attributes of bi-temporal image pairs and a lack of effective methods to suppress the interference of pseudo-changes. While attention mechanisms can improve the detection accuracy of change regions, they may overemphasize certain areas while overlooking other significant change information. Therefore, we propose a Frequency–Spatial–Temporal fusion approach to precisely locate features of interest, suppress pseudo-change interference, and comprehensively understand bi-temporal image changes from both spatial and temporal dimensions. Through a dual-task decoder, we achieve a comprehensive interpretation of changes, further enhancing RSICC performance.

2.2. Remote Sensing Change Detection

Remote sensing change detection(RSCD) primarily focuses on locating and identifying pixel-level changes in bi-temporal RS images, outputting binary change maps indicating changed regions or semantic change maps that represent the type of change for each pixel. Existing methods [18–27] predominantly rely on CNN-based or Transformer-based models to learn image features from bi-temporal images, followed by feature discrimination to determine regions of interest. Li et al. [18] constructed a temporal feature interaction module that captures multi-level change features through interactions between multi-level bi-temporal features in CNNs. Peng et al. [19] introduced a dense hierarchical attention mechanism that employs high-level semantic priors to adaptively guide low-level feature selection through category-aware channel reweighting. In parallel, Liu et al. [20] developed the Prior-Aware Transformer (PA-Former), a multi-stage feature integration framework where low-level spatial features are first processed through Transformer encoder layers to capture global contextual relationships. Zhang et al. [21] introduced SwinSUNet, an end-to-end Transformer architecture for RSCD, which innovatively integrates Swin Transformer blocks into a Siamese U-Net framework. Chen et al. [22] introduced a GAN-based instance-

level change augmentation method for synthesizing new training samples containing building-related changes and corresponding generated bi-temporal images.

Compared with RSCD, RSICC advances further by not only detecting changed regions but also semantically understanding and providing detailed descriptions of the attributes and characteristics of these regions. This approach delivers richer semantic information, requiring an in-depth comprehension of the nature of changes, which is crucial for applications such as environmental monitoring, urban planning, and disaster assessment. RSICC's capability to generate comprehensive change interpretations supports more precise decision-making and strategic planning.

2.3. Natural Image Change Captioning

Compared with traditional image captioning [28–31], change captioning [32–37] primarily focuses on the differences between image pairs, requiring the understanding and articulation of specific changes. Park et al. [32] addressed the practical issue of viewpoint changes in dynamic environments by proposing DUDA, which localizes and describes changes through direct subtraction of bi-temporal image pairs. They also constructed the CLEVER-Change dataset, containing semantic changes and pseudo-changes caused by viewpoint variations. Some researchers have focused on exploring explicit modeling methods for change features that facilitate multi-modal input alignment. VAM [33] designed a viewpoint-independent matching encoder that summarizes and removes common attributes between image pairs for change discrimination. MCCFormers-S and MCCFormers-D [34] employ pure Transformer architectures to implicitly model the interaction between images and caption generation.

Significant distinctions exist between RSICC and NICC. RS images possess higher spatial resolution, providing abundant surface information, but they also face interference from pseudo-changes caused by environmental factors. Bi-temporal RS images are typically captured from the same viewpoint, reducing the impact of viewpoint variations on change localization and captioning. However, they still require handling changes in ground objects due to temporal intervals. Consequently, RSICC must not only accurately identify and localize changes but also conduct in-depth analysis of the environmental context surrounding these changes.

3. Methodology

FST-Net adopts an enhanced encoder–decoder framework based on a Transformer. As illustrated in Figure 2, the overall structure comprises four main components: (1) a multi-scale feature extractor that derives paired image features at four scales from a Transformer-based backbone; (2) a Frequency–Spatial Fusion (FSF) module to adaptively suppress high-frequency features while enhancing low-frequency features across multi-scale representations, mitigating pseudo-change interference; (3) a Spatial–Temporal Modeling (STM) module that employs three scanning mechanisms to process high-level bi-temporal features, strengthening the model's temporal understanding and spatial perception capabilities; and (4) a dual-task decoder that leverages bi-temporal features at different hierarchical levels to generate both change masks and change captions.

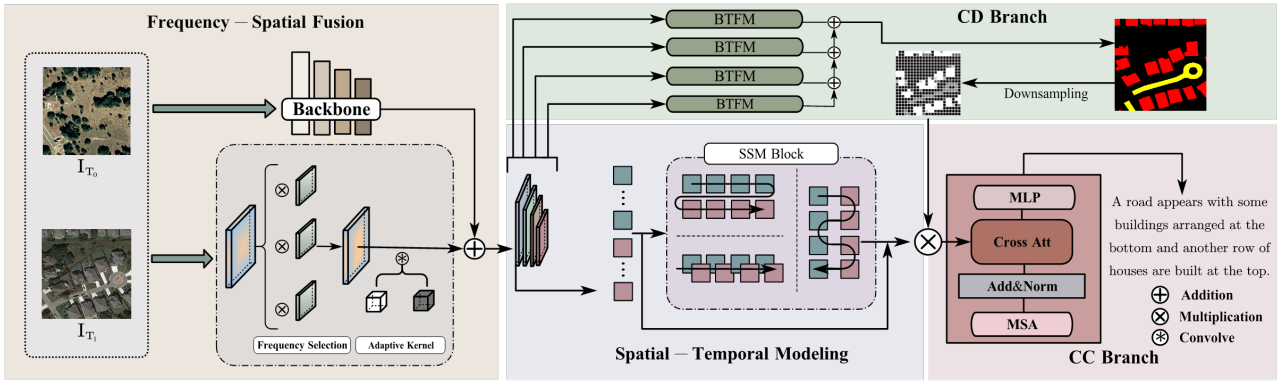

Figure 2. Architecture of the proposed FST-Net. The framework is composed of four essential modules: a multi-scale feature extractor for deriving paired image features at four scales, a Frequency–Spatial Fusion (FSF) module to dynamically attenuate high-frequency components while amplifying low-frequency components, a Spatial-Temporal Modeling (STM) module employing three scanning mechanisms to process high-level bi-temporal features, and a dual-task decoder (DTD) for generating change masks and captions from hierarchical bi-temporal features.

3.1. Multi-Scale Feature Extractor

The feature extractor is primarily designed based on a Siamese SegFormer architecture, leveraging the Mix Vision Transformer (MVT) [38] to extract multi-scale bi-temporal features. Given a pair of bi-temporal images, the overlapping patch merging process in MVT divides each image into multiple overlapping patches, which are subsequently converted into image embeddings through the convolutional layers. Subsequently, four Transformer blocks process these image embeddings, progressively extracting features at different scales. At each stage, the feature resolution gradually decreases while the feature dimension increases. As illustrated in Figure 1, the multi-scale bi-temporal features obtained from the feature extractor are denoted as X^i_{T0} and X^i_{T1}. ($X^0_{T0}, X^0_{T1} \in \mathbb{R}^{8H\times 8W\times C0}$, $X^1_{T0}, X^1_{T1} \in \mathbb{R}^{4H\times 4W\times C1}$, $X^2_{T0}, X^2_{T1} \in \mathbb{R}^{2H\times 2W\times C2}$, $X^3_{T0}, X^3_{T1} \in \mathbb{R}^{H\times W\times C3}$).

3.2. Frequency–Spatial Fusion Module

The Fourier transform facilitates the conversion of a time-domain signal $x(t)$ into its frequency-domain representation by employing the subsequent integral formula, which essentially serves to filter and decompose the signal's frequency components:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t}\, dt \tag{1}$$

Given that the signal is discrete and finite, the integral is converted to a summation, and X becomes a function with a period of 2π. Further, the integral interval is discretized into N sampling points. The interval between sampling points is $\frac{2\pi}{N}$, $\omega = \frac{2\pi}{N}$. Therefore, the Discrete Fourier Transform (DFT) formula is

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi k}{N} n}, k = 0, 1, \ldots, N-1 \tag{2}$$

An image can also be treated as a two-dimensional discrete signal, where x and y represent the pixel coordinates, and $f(x, y)$ denotes the pixel value at the corresponding coordinates. Therefore, by applying DFT to a RS image, it can be converted from the spatial domain to the frequency domain, yielding its frequency-domain representation as follows:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-i2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{3}$$

where $f(x, y)$ represents the pixel values of the image in the spatial domain, $F(u, v)$ denotes the complex representation of the image in the frequency domain, and $e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}$ is the complex exponential function that transforms the image from the spatial domain to the frequency domain.

The physical significance of DFT lies in decomposing an image into a combination of sine and cosine waves of different frequencies. High-frequency features typically contain edge, texture, and detail information in the image, which exhibit rapid changes and significant variations in grayscale values. In RS images, high-frequency features are often associated with noise and subtle changes that may not represent genuine ground object variations but rather pseudo-changes caused by factors such as shadows, variations in illumination conditions, and vegetation color changes. Low-frequency features represent regions in the image where color changes gradually, such as large-scale structures. Low-frequency components generally contain the overall structural information of the image, which is relatively stable and less susceptible to factors like illumination and shadows. Consequently, low-frequency features exhibit fewer pseudo-changes and more accurately reflect genuine surface changes in RS images. To address the low-frequency and high-frequency characteristics of RS images, we designed two modules to enhance low-frequency components and suppress high-frequency components, thereby mitigating the interference of pseudo-changes.

3.2.1. Frequency Selection

We aim to enhance the low-frequency features in RS images that contain more genuine surface changes while suppressing high-frequency features associated with noise. To achieve this, frequency selection is performed on the transformed frequency-domain features. First, different masks are applied to decompose the features into distinct frequency bands:

$$F' = \mathbb{F}^{-}(MF) \tag{4}$$

where $\mathbb{F}^{-}$ denotes the Inverse Fast Fourier Transform (IFFT) reconstructing filtered spatial features, and M represents the learnable frequency mask with binary constraints.

$$M(u, v) = \begin{cases} 1 & \text{if } \theta_n \leq max(F(u, v)) \leq \theta_{n+1} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where F represents a predefined frequency threshold, and we divide the frequency domain into four distinct bands $[0, \frac{1}{16})$, $[\frac{1}{16}, \frac{1}{8})$, $[\frac{1}{8}, \frac{1}{4})$, $[\frac{1}{4}, \frac{1}{2})$. Subsequently, the selection mask is applied to the decomposed frequency bands for spatial dynamic reweighting.

$$F^{s}(h, w) = \sum_{i=0}^{n} S(h, w)F'(h, w) \tag{6}$$

where $F^{s}(h, w)$ represents the frequency features after frequency selection, and $S(h, w)$ denotes the selection map for the corresponding frequency band. This method enhances low-frequency features and attenuates high-frequency features in the frequency domain through spatial dynamic reweighting. It facilitates the extraction of genuine surface features that reflect more structural changes while mitigating the impact of pseudo-change features caused by factors such as illumination on change captioning.

3.2.2. Adaptive Conventional Kernel

After extracting frequency-domain features, convolutional kernels are applied for further processing to understand complex image distributions. Traditional methods employ

the same convolutional kernels to operate on features across different frequency bands. To further distinguish between low-frequency and high-frequency features, we dissect the convolutional kernel parameters into low-frequency and high-frequency constituents.

$$w = w_l + w_h \tag{7}$$

where w_l represents the average convolutional kernel parameters used to process low-frequency features, which is analogous to a low-pass filter, while $w_h = w - w_l$ constitutes the remaining portion, serving as a high-pass filter to handle high-frequency features. The decomposed convolutional kernels dynamically process low-frequency and high-frequency features as follows:

$$w' = \alpha w_l + \beta w_h \tag{8}$$

where α and β represent dynamic weights for each channel, computed through a global pooling operation followed by convolution. This enables the module to adaptively process different frequency components in the feature map, better understanding the image distribution and capturing genuine surface change features.

Following the frequency selection and convolutional operations, as demonstrated in Figure 3, the resultant frequency feature maps are integrated with the initial feature maps. This fusion yields composite feature maps that encapsulate information from both the frequency and spatial domains. This fusion provides a richer spatial–frequency representation.

$$F^{f+s} = F_f + F \tag{9}$$

where F denotes the original feature map, F_f represents the frequency feature map, and F_{f+s} signifies the fused feature map that combines both frequency and spatial information.

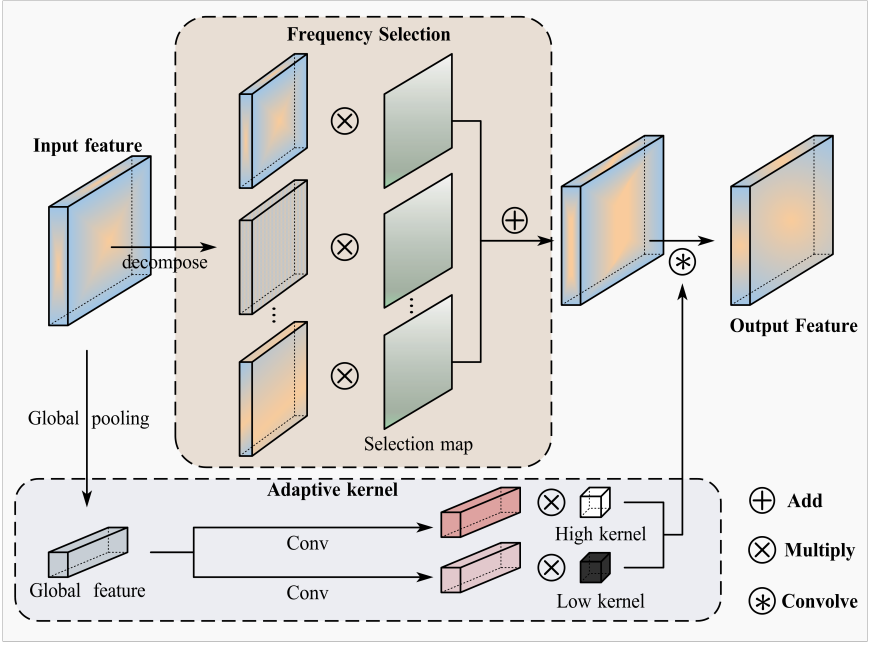

Figure 3. Structure of the Frequency–Spatial Fusion module.

3.3. Spatial–Temporal Modeling Module

The State Space Model (SSM) is a mathematical framework used to describe the evolution of a dynamic system's state over time. SSM characterizes the system's progression through time steps using a set of matrices and state variables. The model typically consists

of state space equations and output equations, which can be computed in either continuous or discrete time. In deep learning, its representation is

$$x'(t) = Ax(t) + Bu(t) \tag{10}$$

$$y(t) = Cx(t) \tag{11}$$

where $x(t)$ denotes the state variable at the current time step, $u(t)$ represents the input variable introduced as an external influence to the system, and $y(t)$ signifies the output variable, which corresponds to the system's observed value. Here, A is the state matrix that governs the evolution of the state vector and influences the state update over time, B is the control matrix, and C is the output matrix. The state $x(t+1)$ at the next time step is computed iteratively, enabling the dynamic modeling of the system's temporal evolution.

Mamba enhances the processing of sequence data by incorporating a selective state space mechanism, which enables the parameter matrices to dynamically adapt in response to the input sequence. This innovation significantly improves the model's ability to selectively process information across sequences and perform context-aware operations on the input. Mamba employs a specialized scanning mechanism to process sequential data, sequentially scanning text or image tokens to capture sequence features. However, the standard Mamba model utilizes a causal scanning mechanism. At any time step, the model only uses current and past information for predictions without future information. While this mechanism is particularly effective for causal sequences such as language modeling, it is not well suited for spatial perception and temporal understanding in visual tasks.

Therefore, by integrating Mamba's temporal scanning mechanism with the temporal attributes of RS image pairs, we design the CC-SSM (Change Captioning State Space Model) spatial-temporal scanning module. As can be seen from Figure 4, this module employs three distinct scanning mechanisms to sequentially process a sequence of image tokens arranged in a column, enhancing the understanding of image change features from both temporal and spatial dimensions.

The sequential scanning mechanism primarily focuses on spatial change perception in images. It arranges the prechange and post-change image tokens in chronological order and scans them sequentially:

$$I_s = [I_{i,j}^{T_1}(0), I_{i,j}^{T_1}(1), \ldots, I_{i,j}^{T_1}(\frac{HW}{2^j}), I_{i,j}^{T_2}(0), I_{i,j}^{T_2}(1), \ldots, I_{i,j}^{T_2}(\frac{HW}{2^j})] \tag{12}$$

The parallel scanning mechanism performs spatiotemporal joint modeling on the image pair by concatenating the tokens of both images along the channel dimension and scanning them simultaneously:

$$I_p = [I_{i,j}^{T_1}(0), I_{i,j}^{T_2}(0), \ldots, I_{i,j}^{T_1}(\frac{HW}{2^j}), I_{i,j}^{T_2}(\frac{HW}{2^j})] \tag{13}$$

The cross-scanning mechanism focuses on the interaction of temporal information between bi-temporal images. It interleaves the tokens from both temporal phases and then scans them:

$$I_c = contact_c(I_{i,j}^{T_1}, I_{i,j}^{T_2}) \tag{14}$$

The operational mechanism of the CC-SSM module is

$$I_i^s = \phi_{SSM}[\sigma(\phi_L(\phi_{Norm}(I_i)))] \tag{15}$$

$$I_d = \sigma\phi_L(I_{i_2} - I_{i_1}) \tag{16}$$

$$I_i^F = I_i^s \cdot I_d \tag{17}$$

$$I_i^O = \phi_L(I_i^F) + I_i \tag{18}$$

where σ represents the SiLU activation function, ϕ_{Norm} denotes the normalization layer, ϕ_L is the linear projection layer, and ϕ_{SSM} signifies the scanning layer. The arranged bi-temporal image tokens first undergo linear projection, normalization, and activation using the SiLU function, followed by spatial-temporal scanning. The resulting output is then multiplied by the original features, further enhancing the model's spatial perception and temporal interaction capabilities.

The Spatial–Temporal Modeling of change features fuses temporal and spatial dimensions, thereby augmenting the model's capacity for temporal comprehension and spatial discernment. Additionally, Mamba's global receptive field and linear complexity contribute to enhanced efficiency and accuracy of the model.

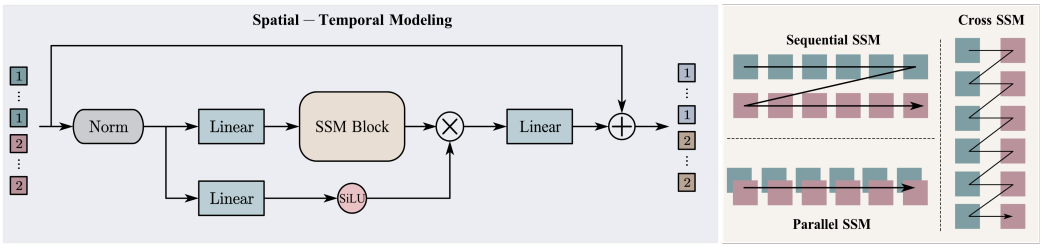

Figure 4. Structure of the Spatial–Temporal Modeling module. Patches labeled '1' represent the pre-change image sequence, while '2' denotes the post-change sequence.

3.4. Dual-Task Decoder

We propose a dual-task decoder that jointly optimizes change detection and change captioning tasks. The pixel-level mask predicted by the change detection decoder further guides the change captioning decoder to generate precise captions of the changes.

3.4.1. Change Detection Decoder

The change detection decoder utilizes multi-level features that incorporate both frequency and spatial information to predict change masks. It fuses bi-temporal features at each layer and then progressively integrates these features from bottom to top through deconvolution. This method integrates high-level features, which are abundant in semantic information, with low-level features that capture the detailed characteristics of ground objects, enhancing the change detection model's discriminative ability and improving its accuracy. The fusion process for bi-temporal features at each layer is:

$$D = conv_{3\times3}(F_{T_1} - F_{T_2}) + cos(F_{T_1}, F_{T_2}) \tag{19}$$

$$F_{contact} = contact(F_{T_1}, D, F_{T_2}) \tag{20}$$

$$F_{fusion} = conv_{1\times1}(RELU(BN(conv_{3\times3}(F_{contact})))) \tag{21}$$

where $cos(\cdot)$ is the computation of cosine similarity.

3.4.2. Change Captioning Decoder

The change captioning decoder utilizes high-level features that integrate frequency, spatial, and temporal information. The fused feature maps output by the Spatial–Temporal Modeling module are fed into the Transformer decoder for captions prediction.

Bi-temporal RS images contain both real changes, such as variations in roads and buildings, and task-irrelevant pseudo-changes, such as illumination variations. To mitigate the impact of pseudo-changes on the change captioning decoder's predictions, it is necessary to guide the model to focus on regions of genuine changes. The output of the change detection decoder, which is a segmentation mask containing change and no-change categories, is thus used to guide the change captioning decoder in generating captions.

$$Mask_D = Downsample(Mask, H, W) \tag{22}$$

$$E_{text} = \phi_{decoder}(F_{cap} \odot Mask_D) \tag{23}$$

where H and W represent the dimensions of the fused feature map, and $Mask$ is the predicted mask from the change detection branch. First, the change mask is downsampled and then element-wise multiplied with the fused feature map. The resulting features are fed into the decoder for caption prediction. This change-detection-assisted captioning method discards image tokens corresponding to unchanged regions in the predicted mask and only processes effective tokens from the changed regions in the subsequent language decoder, generating accurate and task-specific captions.

3.5. Training Objective

For the change detection task, we employ the standard cross-entropy loss to measure the discrepancy between the predicted change mask and the ground truth:

$$\mathfrak{L}_{CD} = -\frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \sum_{j=1}^{C} y_i(j) log(p_i(j)) \tag{24}$$

where C represents the number of categories, N_{train} denotes the total number of pixels, $p_i(j)$ is the predicted probability for the i pixel, and $y_i(j)$ is the ground truth label for the i pixel.

For the change captioning task, following current RSICC practices, we use the cross-entropy loss to measure the discrepancy between the predicted change captions and the ground truth captions:

$$\mathfrak{L}_{CC} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} y_i(t) log(p_i(t)) \tag{25}$$

where M represents the length of the sentence, T denotes the vocabulary size $p_i(j)$ is the probability of the i-th word being assigned to vocabulary index t, and $y_i(t)$ indicates whether the i-th word belongs to vocabulary index t as the ground truth label. To harmonize the contributions of the two tasks to model training, we implement a dual-task learning strategy by normalizing the losses associated with both tasks. This approach enables effective joint optimization for CD and CC. The total loss is formulated as follows:

$$\mathfrak{L} = \mathfrak{L}_{CC} + \frac{detach(\mathfrak{L}_{CC})}{detach(\mathfrak{L}_{CD})} \cdot \mathfrak{L}_{CD} \tag{26}$$

where the *detach* operation separates the tensor from the computation graph, halting gradient computation.

4. Experiments

4.1. Dataset and Evaluation Metrics

4.1.1. LEVIR-MCI Dataset

The LEVIR-MCI dataset [39] is an extension of the current largest change captioning dataset LEVIR-CC [8], containing 10,077 pairs of bi-temporal RS images primarily sourced from the change detection dataset LEVIR-CD [40]. Each image has a resolution of 256 × 256 pixels with a spatial resolution of 0.5 m per pixel. For each image pair, the dataset includes sentences provided by five different annotators, totaling 50,385 description sentences, along with ground truth change masks.

The LEVIR-MCI dataset categorizes image pairs with only irrelevant changes (e.g., illumination, shadow variations) as no-change pairs, comprising 5038 change pairs and 5039 no-change pairs, ensuring an almost balanced distribution. The sentence length distribution in the dataset shows that most sentences range between 5 to 15 words. Sentences for no-change pairs are relatively shorter, with an average length of 5 words, while sentences for change pairs have an average length of approximately 11 words. In the experiments, the dataset is divided into training, validation, and testing sets, which include 6815, 1333, and 1929 pairs of images, respectively.

4.1.2. WHU-CDC Dataset

Derived from the change detection dataset WHU-CD [41], this dataset [42] focuses on post-earthquake urban reconstruction in Christchurch, New Zealand, following the February 2011 seismic event, with imagery spanning from 2011 to 2016. The WHU-CDC dataset targets five change categories: buildings, parking lots, roads, vegetation and water, comprising 7434 high-resolution bi-temporal image pairs (256 × 256 pixels, 0.075 m/pixel resolution). The WHU-CD dataset has 37,170 captions in total with a vocabulary of 327 unique words. The dataset is randomly divided into training, validation, and testing sets, which contain 5947, 743, and 744 pairs of images, respectively.

4.1.3. Evaluation Metrics

The detection masks in the LEVIR-MCI dataset contain three categories: building, road, and no change, and the masks in the WHU-CDC dataset contain two categories: change and no change. To evaluate the performance of FST-Net on the change detection task, we employ the Mean Intersection over Union (MIoU)[43]. And we use intersection over union (IoU) of the change category on WHU-CDC with binary change masks. For assessing FST-Net's performance on the change captioning task, we utilize commonly used metrics: BLEU-n [44] (n = 1, 2, 3, 4), METEOR [45], ROUGE [46], and CIDEr-D [47].

4.2. Experimental Setup

Our proposed model was implemented within the PyTorch 2.1.1 deep learning framework and trained on an NVIDIA GTX 4090 GPU. We employed the pretrained Segformer-B1 as the backbone network for extracting features from bi-temporal images. During training, we utilized the Adam optimizer with an initial learning rate of 0.0001 to optimize the model parameters, and we apply data augmentation through random resizing, flipping, and cropping. The maximum epoch was set to 50, with a word embedding dimension of 512 and a batch size set at 32. Training was halted if the sum of the BLEU-4 score and mIoU score did not increase for 10 consecutive epochs. For inference, the beam search size was fixed as 3. The vocabulary size was set at 501 for LEVIR-MCI and 334 for WHU-CDC, while the max sequence length was set to 41 for LEVIR-MCI and 26 for WHU-CDC.

4.3. Comparison with State-of-the-Art Methods

The LEVIR-MCI dataset and the WHU-CDC dataset is utilized to compare our proposed method with SOTA RSICC methods for the change captioning task, including Capt-Dual-Att, DUDA [32], MCCFormers-S [34], MCCFormers-D [34], PSNet [9], RSICCFormer [8], PromptCC [16], Chg2Cap [10], and SEN [17]. For the change detection task, comparisons are made with SOTA RSICD methods such as BiT [18], ChangeFormer [48], SNUNet [49], and Siam-Diff [25]. The following provides a concise introduction of these methods.

RSICC Methods:

- Capt-Dual-Att [32]: Capt-Dual-Att utilizes dual CNNs to generate spatial attention weight maps for bi-temporal images. The resulting differential features are fed into an LSTM decoder to generate change descriptions.
- DUDA [32]: This method implements a dual-attention mechanism with parallel decoders—one focusing on change regions via spatial-channel attention, and the other modeling contextual stability through global average pooling.
- MCCFormers-S [34]: It is a single-stream Transformer architecture that processes concatenated bi-temporal features through shared encoders.
- MCCFormers-D [34]: MCCFormers-D employs dual independent Transformer encoders to extract hierarchical features from each temporal phase.
- PSNet [9]: It is a pure Transformer-based framework incorporating stacked difference-aware layers to compute multi-scale feature discrepancies. Scale-aware enhancement modules adaptively reweight spatial–frequency components, prioritizing structural changes over high-frequency noise.
- RSICCFormer [8]: In this method, a siamese dual-branch Transformer where cross-encoder modules compute differential attention maps to highlight change regions. Multi-level Bi-temporal Fusion modules hierarchically aggregate features from early to late encoding stages, capturing both coarse and fine-grained change semantics.
- PromptCC [16]: This method integrates a frozen LLM with visual encoders by injecting task-specific prompts and change category embeddings. This approach bridges the domain gap between low-level visual features and high-level linguistic priors, enabling human-aligned caption generation.
- Chg2Cap [10]: Chg2Cap combines a Siamese CNN backbone with a hybrid attention mechanism. Spatial-differential attention dynamically weights change-related regions, while the Transformer decoder iteratively refines descriptions using both visual evidence and lexical context, achieving state-of-the-art performance on fine-grained change categories.
- SEN [17]: SEN leverages a pretrained single-stream feature extractor with shallow feature encoding (SFE) modules to preserve high-resolution details. Cross-attention guided difference (CAGD) modules compute channel-wise discrepancy maps, enhancing sensitivity to subtle changes while suppressing misaligned registration artifacts.

RSCD Methods:

- ChangeFormer [48]: It is a Transformer-based change detection framework that leverages self-attention mechanisms to model global feature dependencies and contextual change relationships.
- SNUNet [49]: SNUNet combines Siamese encoders with a nested U-Net architecture to preserve high-resolution spatial details. The integrated Enhanced Channel Attention Module performs multi-scale channel recalibration, while deep supervision through skip connections ensures fine-grained feature retention for sub-meter change detection.

- BiT [18]: This method employs a bidirectional temporal convolutional network to process temporal features, enabling enhanced capture of change patterns in time-series data.
- Siam-Diff [25]: Siam-Diff is a Siamese network variant that computes absolute difference maps between same-scale encoder features. Multi-level difference maps are concatenated and progressively refined through decoder blocks, emphasizing scale-consistent change propagation for robust binary change prediction.

In change captioning tasks, BLEU-n primarily evaluates the accuracy and fluency of generated text, while METEOR comprehensively assesses text quality with a focus on semantic alignment closer to human understanding. ROUGE emphasizes the coverage of reference text information, and CIDEr-D prioritizes the diversity and precision of generated captions. For change detection tasks, the primary metric employed is mIoU, which quantifies the similarity between model predictions and ground truth labels.

Table 1 presents a comprehensive performance comparison between the proposed FST-Net and state-of-the-art methods on the LEVIR-MCI dataset. The results demonstrate that FST-Net achieves superior performance across all evaluation metrics, establishing new benchmarks for both change detection and captioning tasks. Specifically, compared with the previous SOTA model Chg2Cap, FSTNet exhibits significant improvements of 3.65% in BLEU-4, 1.20% in METEOR, 1.37% in ROUGE-L, and 4.08% in CIDEr-D. The selection of the largest public dataset, LEVIR-MCI, for model evaluation was necessitated by FST-Net's multi-task learning framework, which requires change detection masks. The performance enhancement in these tasks is attributed to the integration of multi-domain information (temporal, spatial, and frequency), enabling enhanced focus on changes of interest. Additionally, the joint optimization framework bridges pixel-level change detection with semantic caption generation, improves localization accuracy and comprehensiveness of captions for change regions.

Table 1. Benchmarking results of state-of-the-art methods on the LEVIR-MCI dataset for change detection and captioning tasks. Higher metric values denote better performance, with best results highlighted in bold.

Task	Method	Metrics							
		mIoU	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr-D**
Change Detection	ChangeFormer [48]	78.50	–	–	–	–	–	–	–
	SNUNet [49]	82.76	–	–	–	–	–	–	–
	BiT [18]	83.14	–	–	–	–	–	–	–
	Siam-Diff [25]	83.75	–	–	–	–	–	–	–
Change Captioning	Capt-Dual-Att [32]	–	79.51	70.57	63.23	57.46	36.56	70.69	124.42
	DUDA [32]	–	81.44	72.22	64.24	57.79	37.15	71.04	124.32
	MCCFormers-S [34]	–	79.90	70.26	62.68	56.68	36.17	69.46	120.39
	MCCFormers-D [34]	–	80.42	70.87	62.86	56.38	37.29	70.32	124.44
	PSNet [9]	–	83.86	75.13	67.89	62.11	38.80	73.60	132.62
	RSICCFormer [8]	–	84.72	76.12	68.87	62.77	39.61	74.12	134.12
	PromptCC [16]	–	83.66	75.73	69.10	63.54	38.82	73.72	136.44
	Chg2Cap [10]	–	86.14	76.08	70.66	63.36	40.03	75.12	134.55
	SEN [17]	–	85.10	77.05	70.10	64.09	39.59	74.57	136.02
	FST-Net (Ours)	**85.32**	**86.76**	**78.82**	**71.71**	**65.67**	**40.51**	**76.15**	**140.04**

Note: "–" indicates the metric is not applicable for the task.

The experimental results on the WHU-CDC dataset, presented in Table 2, also highlight the competitive performance of our proposed FST-Net. FST-Net achieves the highest scores on all evaluation metrics, BLEU-4 at 76.78, METEOR at 48.78, ROUGE at 82.91, CIDEr at 160.01, demonstrating its effectiveness in generating accurate and semantically meaningful captions. Compared with the LEVIR-MCI dataset with multi-label masks, the WHU-CDC dataset employs binary change masks focusing on building change detection. FST-Net also demonstrates excellent performance in change detection with an IoU of 87.36. Above all,

our method achieves state-of-the-art results across different datasets, proving the robustness and adaptability, which underscore the superiority of FST-Net in addressing the challenges of RSICC.

Table 2. Benchmarking results of methods on the WHU-CDC dataset for change detection and captioning tasks. Best results are highlighted in bold.

Task	Method	Metrics							
		IoU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Change Detection	ChangeFormer [48]	74.81	–	–	–	–	–	–	–
	SNUNet [49]	77.88	–	–	–	–	–	–	–
Change Captioning	Capt-Dual-Att [32]	–	77.15	68.47	60.81	55.42	34.22	68.04	121.45
	DUDA [32]	–	79.04	69.53	61.57	55.64	34.29	68.98	121.85
	MCCFormers-S [34]	–	77.18	70.48	65.45	61.76	43.17	74.97	138.99
	MCCFormers-D [34]	–	82.28	76.64	72.05	68.80	43.79	76.48	139.79
	PSNet [9]	–	85.80	80.63	76.41	73.75	46.40	79.85	149.03
	RSICCFormer [8]	–	82.61	77.25	72.72	69.52	45.56	78.23	147.43
	PromptCC [16]	–	82.68	76.94	72.20	68.99	43.78	76.14	137.10
	Chg2Cap [10]	–	86.33	82.07	78.54	76.16	48.29	81.44	156.10
	SEN [17]	–	85.12	80.28	76.84	74.71	46.81	79.45	148.70
	FST-Net (Ours)	**87.36**	**88.15**	**83.46**	**79.55**	**76.78**	**48.78**	**82.91**	**160.01**

Note: "–" indicates the metric is not applicable for the task.

4.4. Ablation Studies

This section presents ablation studies to validate the effectiveness of the proposed modules: Frequency–Spatial Fusion (FSF) module, Spatial–Temporal Modeling (STM) module, and dual-task decoder. Table 3 demonstrates the contributions of these components in FST-Net. It is evident that both FSF and STM contribute to the improvement of model performance. The dual-task decoder alone also outperforms the baseline model, owing to the mutual reinforcement between change detection and change captioning tasks, which leads to more accurate localization of changed regions. When FSF and STM are jointly employed, the model captures multi-domain (frequency, spatial, and temporal) change information, achieving the best result. Notably, the STM + Dual-Decoder configuration slightly outperforms the full FSF + STM + Dual-Decoder model in BLEU-2, BLEU-3, and BLEU-4. The BLEU-n metric primarily measures the degree of n-gram overlap between the generated text and the reference text. During feature fusion, information from different domains may conflict, leading to a decrease in n-gram overlap for some generated texts. However, the overall performance of FST-Net still improves, indicating that the model has advantages in semantic accuracy and diversity.

Table 3. Ablation study of FST-Net with different module configurations on the LEVIR-MCI dataset. Symbols "×" and "✓" indicate the exclusion and inclusion of specific modules, respectively. Higher metric values denote better performance, with best results highlighted in bold.

FSF	STM	DTD	B-1	B-2	B-3	B-4	M	R	C
×	×	×	82.09	73.77	66.46	60.51	37.49	72.04	126.51
✓	×	×	84.97	76.75	69.71	63.98	39.62	74.65	135.61
×	✓	×	85.79	77.99	71.04	65.24	39.91	75.24	136.56
×	×	✓	85.84	77.67	70.60	64.97	40.10	75.10	137.76
✓	✓	×	84.96	76.99	70.34	64.81	40.19	74.81	137.07
✓	×	✓	85.98	78.10	71.35	65.16	40.26	75.29	138.77
×	✓	✓	86.48	78.95	**71.99**	**66.10**	40.26	75.75	138.43
✓	✓	✓	**86.76**	**78.82**	71.71	65.67	**40.51**	**76.15**	**140.04**

In Table 4, we further evaluate the model's performance by determining whether changes exist between bi-temporal RS images and assessing the effectiveness of the generated captions. The test set is divided into three parts: (1) image pairs with no changes,

(2) image pairs with changes, and (3) image pairs containing both changed and unchanged areas. It can be observed that the model incorporating both FSF and STM modules achieves better results compared with the baseline model. This demonstrates that capturing more change-related feature information and excluding the interference of irrelevant changes in caption generation can enhance both the accuracy and diversity of the captions. The model with the dual-task decoder outperforms the model with a single-task decoder, which proves that the change detection task can further enhance the performance of the change captioning task. Therefore, the proposed modules contribute to both distinguishing whether changes have occurred and describing the changes.

Table 4. Ablation studies on the FSF, STM, and DTD modules on the test sets with only no changes and only changes and the entire test set. Symbols "×" and "✓" indicate the exclusion and inclusion of specific modules, respectively. Higher metric values denote better performance, with best results highlighted in bold.

Test Range	FSF	STM	DTD	Metrics						
				B-1	B-2	B-3	B-4	M	R	C
Only no-change	×	×	×	97.19	96.75	96.55	96.44	74.89	97.34	–
	✓	×	✓	97.80	97.45	97.29	97.17	76.58	97.85	–
	×	✓	✓	97.59	97.21	97.00	96.86	76.10	97.79	–
	✓	✓	✓	**98.96**	**98.78**	**98.67**	**98.59**	**80.24**	**99.01**	–
Only change	×	×	×	69.96	54.99	41.12	29.53	21.69	46.70	42.46
	✓	×	✓	76.40	62.32	50.03	39.98	25.39	52.70	64.74
	×	✓	✓	77.34	**63.96**	**51.20**	**40.40**	25.65	53.68	63.19
	✓	✓	✓	**78.06**	63.49	50.59	39.88	**25.81**	**54.17**	**64.75**
Entire set	×	×	×	82.09	73.77	66.46	60.51	37.49	72.04	126.51
	✓	×	✓	85.98	78.10	71.35	65.16	40.26	75.29	138.77
	×	✓	✓	86.48	**78.95**	**71.99**	**66.10**	40.26	75.75	138.43
	✓	✓	✓	**86.76**	78.82	71.71	65.67	**40.51**	**76.15**	**140.04**

4.4.1. Parameter and Complexity Analysis

In the comparison of total parameters, FLOPs, and inference time across different methods for 256 × 256 pixel images (as shown in Table 5), FST-Net demonstrates clear advantages. While ranking first in caption accuracy, FST-Net exhibits strong parameter efficiency and computational complexity. For instance, it reduces parameter counts by half compared with Chg2Cap. Our method also achieves competitive inference speed due to the low linear complexity of Mamba in the STM module, ensuring that it is practical for real-world applications. Notably, its significantly lower FLOPs stem from the element-wise multiplication with change masks in the captioning decoder, which filters out unchanged regions and reduces computational overhead. This efficiency highlights FST-Net's ability to achieve superior performance with reduced resource demands. Overall, these result highlight the advantages of FST-Net in memory optimization and computational acceleration, positioning it as a compelling solution for resource-constrained environments or scenarios demanding high-speed processing.

Table 5. Comparison of parameters and complexity with image size 256 × 256.

Method	Params (M)	FLOPs (G)	Inference (Sample/ms)
DUDA	80.31	20.28	30.18
MCCFormer-S	162.55	25.09	36.11
MCCFormer-D	162.55	25.09	35.71
RSICCFormer	172.80	27.10	44.30
PromptCC	408.58	19.88	57.59
Chg2Cap	231.49	37.36	49.10
FST-Net	111.76	24.13	31.90

4.4.2. Frequency Extraction Module

Tables 3 and 4 demonstrate the effectiveness of the FSF module. In the FSF module, we propose frequency selection and adaptive convolutional kernels to enhance low-frequency components containing more relevant change information while suppressing high-frequency noise. Table 6 presents ablation studies of these components on the LEVIR-MCI dataset. To eliminate interference from other modules, we conducted independent experiments for change captioning and change detection rather than using the dual-task decoder for joint optimization. The results demonstrate that both components contribute to performance improvement in change interpretation by effectively mitigating pseudo-change interference. Notably, frequency selection yields greater performance gains in both change detection and captioning tasks. This superior performance stems from its ability to sharpen the model's focus on regions of genuine interest, while the combined two-stage FSF architecture achieves optimal results.

Table 6. Ablation study of the FSF module with frequency selection and adaptive convolutional kernels on the LEVIR-MCI dataset. "FS" represents frequency selection, and "AK" represents adaptive convolutional kernel. Symbols "×" and "✓" indicate the exclusion and inclusion of specific modules, respectively. Higher metric values denote better performance, with best results highlighted in bold.

FS	AK	mIoU	B-1	B-2	B-3	B-4	M	R	C
×	×	82.01	82.09	73.77	66.46	60.51	37.49	72.04	126.51
✓	×	84.16	83.49	75.55	68.61	62.80	38.69	73.49	132.85
×	✓	83.42	82.86	74.97	67.71	61.84	37.98	72.81	128.56
✓	✓	84.99	**84.97**	**76.75**	**69.71**	**63.98**	**39.62**	**74.65**	**135.61**

To further validate the impact of the frequency selection mechanism on model performance and demonstrate that enhancing low-frequency components can effectively suppress pseudo-change interference, we conduct ablation experiments on different frequency band partitioning strategies and dynamic reweighting mechanisms. The experiments systematically evaluated four configurations: low-frequency band ($[0, \frac{1}{16})$), mid-low frequency band ($[0, \frac{1}{8})$), full frequency band ($[0, \frac{1}{2})$), and four partitioned bands ($[0, \frac{1}{16})$, $[\frac{1}{16}, \frac{1}{8})$, $[\frac{1}{8}, \frac{1}{4})$, $[\frac{1}{4}, \frac{1}{2})$) with both dynamic and fixed weighting approaches. The results in Table 7 demonstrate that dynamic weighting across all four frequency bands achieves optimal performance, significantly outperforming other configurations. The dynamic reweighting mechanism contributes substantial improvements of 2.74 mIoU and 6.34 CIDEr, confirming its critical role in pseudo-change suppression through adaptive weight learning. Notably, the mid-low frequency band ($[0, \frac{1}{8})$) exhibits the best balance by preserving essential low-frequency change features while avoiding high-frequency noise interference, whereas using only the lowest frequency band ($[0, \frac{1}{16})$) leads to a 0.62 mIoU degradation, indicating the importance of mid-frequency information for accurate detection. Collectively, these findings prove

that the frequency selection mechanism effectively enhances the model's capability to identify genuine changes by amplifying critical frequency components and suppressing noise interference.

Figure 5 visualizes three types of pseudo-changes and the frequency map generated by the FSF module, where high-frequency components are color-coded in red and low-frequency components in blue. The enhanced low-frequency features significantly emphasize regions of interest where the yellow box marks, such as emerging roads and buildings, while attenuating background elements like static vegetation. Case (1) demonstrates pseudo-changes caused by illumination variations. Chg2Cap is influenced by lighting angles and erroneously concludes that "a building appears in the woods". In comparison, FST-Net addresses this by leveraging frequency-domain analysis to emphasize building regions in both prechange and post-change images, correctly identifying no actual structural changes. In Case (2), seasonal effects introduce discrepancies in vegetation shape and color between images. FST-Net suppresses background vegetation to focus on road networks. Case (3) shows interference from tree shadows. With frequency information, FST-Net accurately detects houses within tree-covered areas across both temporal images. This frequency domain processing mechanism directs the model's attention to semantically relevant changes of interest, effectively mitigating pseudo-change interference caused by transient factors. The visual evidence confirms that FSF improves localization precision of ROIs by adaptively filtering task-irrelevant frequency bands, thereby enhancing the robustness of change interpretation systems.

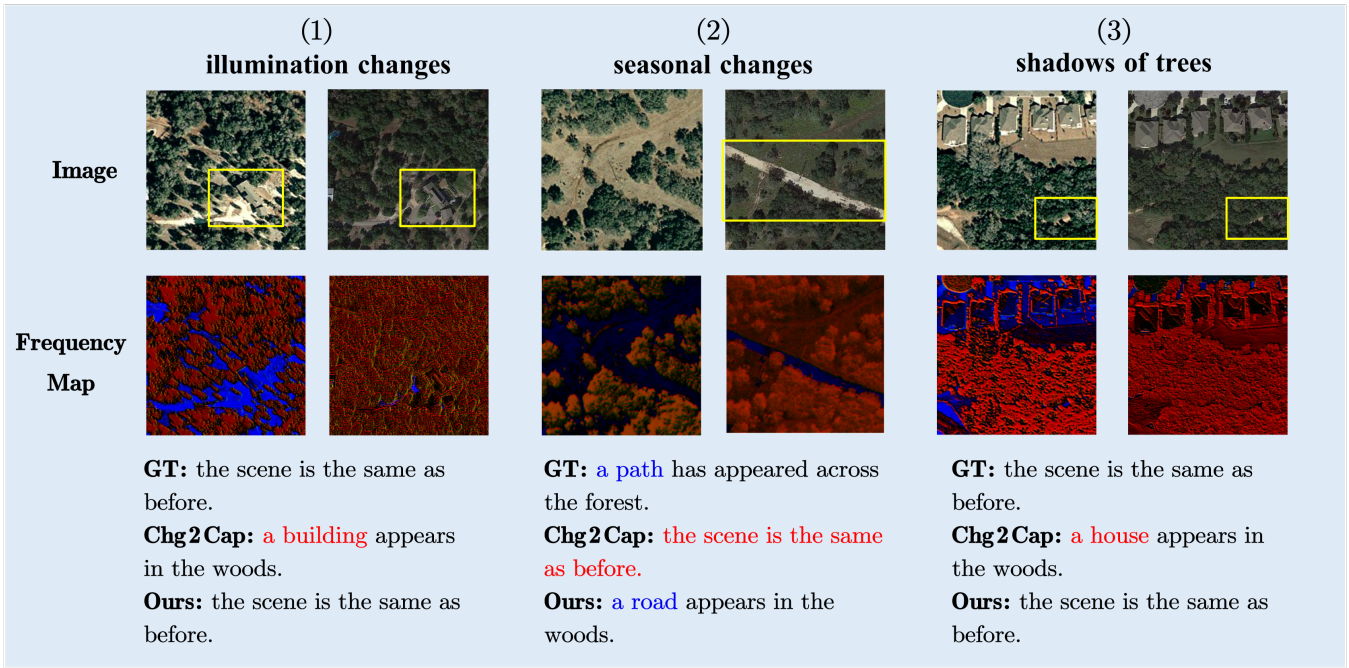

Figure 5. Visualization of the FSF module suppressing various types of pseudo-change interference, where high-frequency components of the frequency map are color-coded in red and low-frequency components in blue. Blue words highlight the correctly predicted change objects for our method, and red words highlight the wrong captions from Chg2Cap. The yellow box marks the enhanced region.

Table 7. Ablation study of frequency band selection and dynamic weighing mechanisms in the FSF module on the LEVIR-MCI dataset. Higher metric values indicate better performance, with best results highlighted in bold.

Frequency Band	Dynamic Weighting	mIoU	B-4	M	R	C
Full	×	82.23	60.51	37.49	72.04	126.51
0-1/2	✓	82.50	60.75	37.60	72.25	127.80
0-1/8	✓	83.77	62.05	38.20	73.12	131.10
0-1/16	✓	83.15	61.20	37.95	72.88	129.42
Full	✓	**84.97**	**62.80**	**38.69**	**73.49**	**132.85**

4.4.3. Spatial–Temporal Modeling Module

The STM module integrates three scanning mechanisms—sequential scanning, cross scanning, and parallel scanning—to jointly facilitate the Spatial–Temporal Modeling of bi-temporal features. As demonstrated in Table 8, the synergistic combination of these mechanisms achieves optimal performance, with quantitative results validating their significant contributions to improve the accuracy of change captioning. Specifically, sequential scanning strengthens spatial change perception by progressively aggregating local-to-global context, while cross scanning enables effective temporal interaction through token-wise interleaving of bi-temporal features. Parallel scanning further reinforces joint spatiotemporal dependency learning by concurrently processing aligned feature pairs. This hierarchical architecture ensures comprehensive capture of both transient and persistent changes, effectively addressing challenges such as occlusions and pseudo-changes in complex geospatial scenarios.

Table 8. Ablation study of the STM module with different combinations of scanning mechanisms. The baseline model includes only the FSF module and the DTD module. Symbols "×" and "✓" indicate the exclusion and inclusion of specific modules, respectively. Higher metric values denote better performance, with best results highlighted in bold.

Method	S	P	C	B-1	B-2	B-3	B-4	M	R	C
Baseline	×	×	×	85.98	78.10	71.35	65.16	40.26	75.29	138.77
-	✓	×	×	86.13	78.19	71.44	65.43	40.28	75.58	139.01
-	✓	✓	×	86.32	77.89	71.27	65.29	40.31	75.79	139.83
Dual Task	✓	✓	✓	**86.76**	**78.82**	**71.71**	**65.67**	**40.51**	**76.15**	**140.04**

4.4.4. Dual-Task Decoder

To balance the losses between dual tasks in the dual-task decoder, we employ a gradient detach operation to ensure both losses operate at comparable magnitudes. As demonstrated in Table 9, the integration of the change detection decoder significantly enhances the performance of the change captioning model. This loss balancing strategy harmonizes the optimization of both tasks, providing comprehensive and interpretable results for bi-temporal change analysis.

Table 9. Quantitative comparison between dual-task decoder and single-task decoder. Higher metric values denote better performance, with best results highlighted in bold.

Method	mIoU	B-1	B-2	B-3	B-4	M	R	C
Only CC	–	86.06	77.44	69.87	63.65	39.81	74.63	136.85
Only CD	84.99	–	–	–	–	–	–	–
Dual Task	**85.32**	**86.76**	**78.82**	**71.71**	**65.67**	**40.51**	**76.15**	**140.04**

4.5. Qualitative Analysis

Figures 6 and 7 present the captioning results and detection masks for the change in the LEVIR-MCI and WHU-CDC datasets. For each image pair, we provide one of the five ground-truth sentences alongside captions generated by Chg2Cap and our proposed FST-Net. Accurately predicted change-related terms are highlighted in blue, while erroneous captions are marked in red. The results demonstrate that FST-Net generates more comprehensive and precise captions compared with existing methods, effectively summarizing temporal object changes while producing finer-grained and more accurate change detection masks.

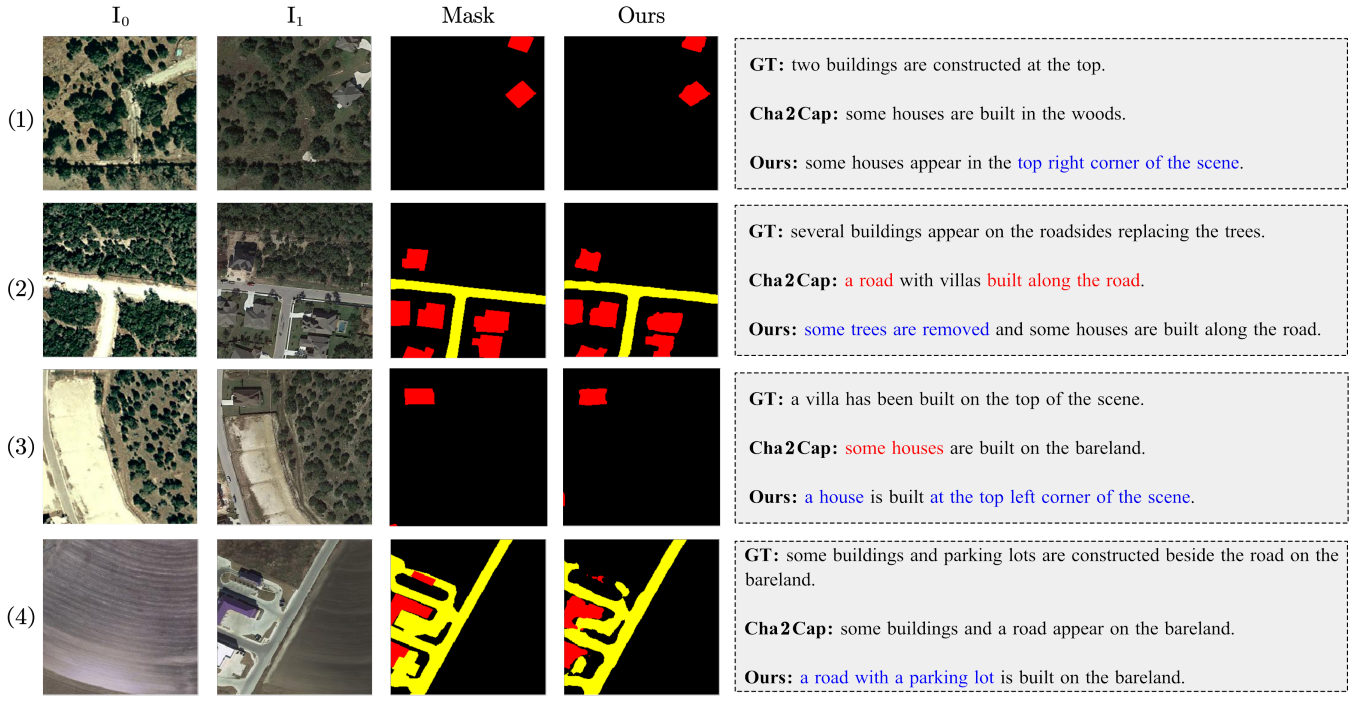

Figure 6. Qualitative results on the LEVIR-MCI dataset. GT represents the ground truth caption. Mask represents the ground truth change mask. Red words emphasized instances where Chg2Cap erroneously predicts change targets. Blue words highlighted the correctly predicted change objects for our FST-Net.

On the LEVIR-MCI dataset, FST-Net precisely localizes change directions, as shown in pairs (1) and (3), whereas Chg2Cap provides vague positional references (e.g., "some houses are built in the woods" vs. "some houses appear in the top right corner of the scene"). Furthermore, FST-Net captures temporal evolution, as illustrated in pair (2), where it describes "some trees are removed and some houses are built along the road", while existing methods fail to articulate sequential changes. The integration of change detection masks enables FST-Net to localize regions and identify change types with superior precision. In pair (4), FST-Net correctly identifies "a parking lot", a detail omitted by baselines.

On the WHU-CDC dataset, FST-Net also shows more accurate prediction results, as shown in pairs (1) and (2), while Chg2Cap provides wrong change direction or even ignores the change region. In pair (3), FST-Net correctly identifies "a building with a parking lot", a detail omitted by Chg2Cap. Pair (4) shows that FST-Net accurately identifies the location of house construction. These examples collectively validate that our method not only generates holistic change analyses but also achieves higher accuracy and robustness by synergizing multi-domain features and task-specific optimization.

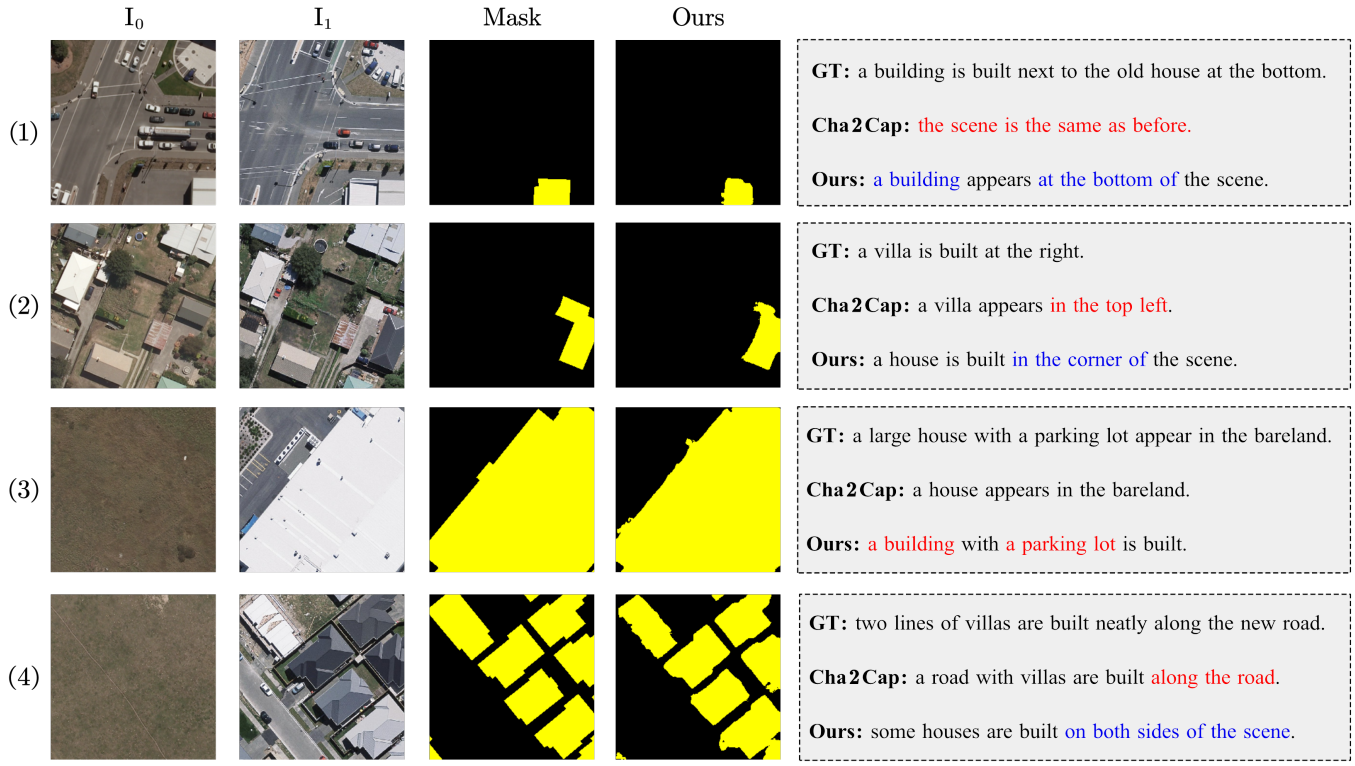

Figure 7. Qualitative results on the WHU-CDC dataset. GT represents the ground truth caption. Mask represents the ground truth change mask. Red words emphasized instances where Chg2Cap erroneously predicts change targets. Blue words highlighted the correctly predicted change objects for our FST-Net.

However, our method still has certain limitations. As shown in Figure 8, FST-Net primarily focuses on newly emerged objects while sometimes overlooking disappeared ones. It also fails to accurately describe the extent of changes. For instance, in Case 2, it did not recognize the road widening. Additionally, FST-Net occasionally makes errors in quantifying changed objects, as seen in Case 3, where it only described building changes while missing the newly appeared road. These observations indicate that FST-Net's comprehensiveness in change captioning needs further improvement.

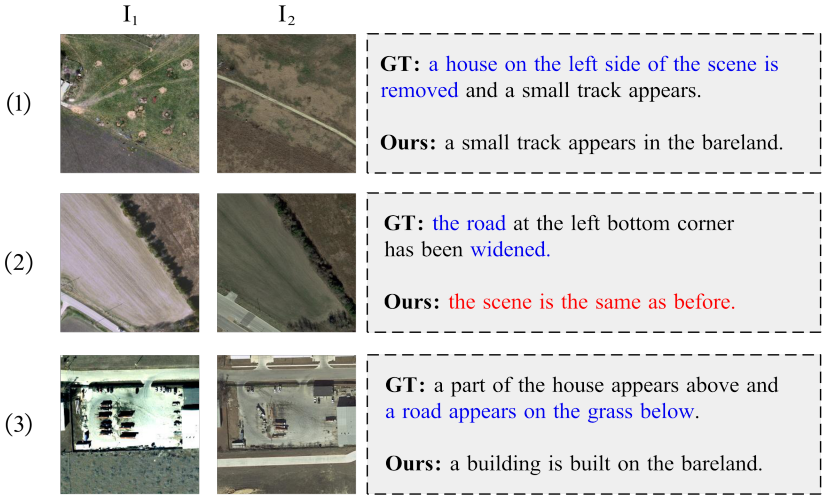

Figure 8. Some failure cases of FST-Net. GT represents the ground truth caption. Red words emphasized instances where FST-Net erroneously predicts change targets. Blue words highlighted the change objects ignored by FST-Net in ground truth.

5. Conclusions

This paper presents a Frequency–Spatial–Temporal Fusion Network (FST-Net), a novel framework addressing critical limitations in RSICC by integrating multi-domain features to enhance both accuracy and robustness. FST-Net synergizes frequency-domain filtering, Spatial–Temporal Modeling, and dual-task learning to achieve comprehensive change interpretation. FSF adaptively suppresses high-frequency noise while amplifying structural changes through frequency feature fusion, reducing pseudo-change interference. STM employs sequential, cross, and parallel scanning mechanisms to capture temporal dependencies and spatial interactions. Coupled with a dual-task decoder that jointly optimizes pixel-level change detection and semantic-level captioning, FST-Net achieves remarkable performance on the LEVIR-MCI dataset (mIoU: 85.32, BLEU-4: 65.67, METEOR: 40.51, ROUGE: 76.15, CIDEr-D: 140.04), demonstrating its capability to generate precise, context-aware descriptions while localizing fine-grained changes such as urban infrastructure developments and vegetation dynamics. This work is able to achieve a good balance between RSCD and RSCC tasks.

Nevertheless, it should be acknowledged that there is still room for further improvement in our work. First, given the diverse types of pseudo-changes, we should continue to investigate frequency-domain features and explore better fusion methods between frequency and spatial features, aiming to fundamentally address the impact of pseudo-changes on the model's understanding of the data distribution in bi-temporal image pairs. Additionally, further reducing the number of parameters and model training time will facilitate the deployment and application of the RSICC model on mobile devices. Moreover, in terms of generated captions, our model primarily focuses on objects that appear in the post-change images while neglecting targets that disappear after the change, resulting in incomplete semantic coverage. This is another direction we need to explore in the future. We will continue to refine our approach to address some of the existing issues in the current model.

Author Contributions: Conceptualization, S.Z.; methodology, S.Z.; software, S.Z. and Y.X.; validation, S.Z.; formal analysis, S.Z.; investigation, S.Z. and Y.X.; resources, S.Z.; data curation, S.Z.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z., Y.X., Y.W. and X.L.; visualization, S.Z.; supervision, Y.X., Y.W. and X.L.; funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Provincial Natural Science Foundation Project (No.2023JJ30082).

Data Availability Statement: The LEVIR-MCI dataset can be obtained from https://github.com/Chen-Yang-Liu/Change-Agent, accessed on 25 December 2024. The WHU-CDC dataset can be obtained from https://huggingface.co/datasets/hygge10111/RS-CDC, accessed on 1 March 2025. Further inquiries can be directed to the corresponding author.

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

References

1. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 811–815. [CrossRef] [CrossRef]
2. Sakurada, K.; Okatani, T. Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015 . [CrossRef]
3. Khan, S.H.; He, X.; Porikli, F.; Bennamoun, M. Forest Change Detection in Incomplete Satellite Images With Deep Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5407–5423. [CrossRef] [CrossRef]

4. Dong, D.; Zhang, R.; Guo, W.; Gong, D.; Zhao, Z.; Zhou, Y.; Xu, Y.; Fujioka, Y. Assessing Spatiotemporal Dynamics of Net Primary Productivity in Shandong Province, China (2001–2020) Using the CASA Model and Google Earth Engine: Trends, Patterns, and Driving Factors. *Remote Sens.* **2025**, *17*, 488. [CrossRef] [CrossRef]
5. Gong, D.; Huang, M.; Ge, Y.; Zhu, D.; Chen, J.; Chen, Y.; Zhang, L.; Hu, B.; Lai, S.; Lin, H. Revolutionizing ecological security pattern with multi-source data and deep learning: An adaptive generation approach. *Ecol. Indic.* **2025**, *173*, 113315. [CrossRef] [CrossRef]
6. Chouaf, S.; Hoxha, G.; Smara, Y.; Melgani, F. Captioning Changes in Bi-Temporal Remote Sensing Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2891–2894. [CrossRef]
7. Hoxha, G.; Chouaf, S.; Melgani, F.; Smara, Y. Change Captioning: A New Paradigm for Multitemporal Remote Sensing Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef] [CrossRef]
8. Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; Shi, Z. Remote Sensing Image Change Captioning With Dual-Branch Transformers: A New Method and a Large Scale Dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef] [CrossRef]
9. Liu, C.; Yang, J.; Qi, Z.; Zou, Z.; Shi, Z. Progressive Scale-Aware Network for Remote Sensing Image Change Captioning. In Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6668–6671. [CrossRef]
10. Chang, S.; Ghamisi, P. Changes to Captions: An Attentive Network for Remote Sensing Change Captioning. *IEEE Trans. Image Process.* **2023**, *32*, 6047–6060. [CrossRef] [CrossRef]
11. Liu, C.; Chen, K.; Qi, Z.; Liu, Z.; Zhang, H.; Zou, Z.; Shi, Z. Pixel-Level Change Detection Pseudo-Label Learning For Remote Sensing Change Captioning. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 8405–8408. [CrossRef]
12. Zhong, Y.; Li, B.; Tang, L.; Kuang, S.; Wu, S.; Ding, S. Detecting Camouflaged Object in Frequency Domain. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4494–4503. [CrossRef]
13. Yang, Y.; Yuan, G.; Li, J. SFFNet: A Wavelet-Based Spatial and Frequency Domain Fusion Network for Remote Sensing Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef] [CrossRef]
14. Miao, Z.; Zhao, M. Time–space–frequency feature Fusion for 3-channel motor imagery classification. *Biomed. Signal Process. Control* **2024**, *90*, 105867. [CrossRef] [CrossRef]
15. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024. [CrossRef]
16. Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; Shi, Z. A Decoupling Paradigm With Prompt Learning for Remote Sensing Image Change Captioning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–18. [CrossRef] [CrossRef]
17. Zhou, Q.; Gao, J.; Yuan, Y.; Wang, Q. Single-Stream Extractor Network with Contrastive Pre-Training for Remote-Sensing Change Captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef] [CrossRef]
18. Li, Z.; Tang, C.; Wang, L.; Zomaya, A.Y. Remote Sensing Change Detection via Temporal Feature Interaction and Guided Refinement. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef] [CrossRef]
19. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7296–7307. [CrossRef] [CrossRef]
20. Liu, M.; Shi, Q.; Chai, Z.; Li, J. PA-Former: Learning Prior-Aware Transformer for Remote Sensing Building Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef] [CrossRef]
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
22. Chen, H.; Li, W.; Shi, Z. Adversarial Instance Augmentation for Building Change Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef] [CrossRef]
23. ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 228–239. [CrossRef] [CrossRef]
24. Rahman, F.; Vasu, B.; Cor, J.V.; Kerekes, J.; Savakis, A. Siamese network with multi-level features for patch-based change detection in satellite imagery. In Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, 26–28 November 2018; pp. 958–962. [CrossRef]
25. Caye Daudt, R.; Le Saux, B.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [CrossRef]
26. Hou, B.; Liu, Q.; Wang, H.; Wang, Y. From W-Net to CDGAN: Bitemporal Change Detection via Deep Learning Techniques. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1790–1802. [CrossRef] [CrossRef]

27. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1194–1206. [CrossRef] [CrossRef]
28. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [CrossRef]
29. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015. [CrossRef]
30. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10575–10584. [CrossRef]
31. Zhong, Y.; Wang, L.; Chen, J.; Yu, D.; Li, Y. Comprehensive Image Captioning via Scene Graph Decomposition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 211–229. [CrossRef]
32. Park, D.H.; Darrell, T.; Rohrbach, A. Robust Change Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4623–4632. [CrossRef]
33. Xiangxi, S.; Yang, X.; Gu, J.; Joty, S.; Cai, J. Finding It at Another Side: A Viewpoint-Adapted Matching Encoder for Change Captioning. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 574–590. [CrossRef]
34. Qiu, Y.; Yamamoto, S.; Nakashima, K.; Suzuki, R.; Iwata, K.; Kataoka, H.; Satoh, Y. Describing and Localizing Multiple Changes with Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 1951–1960. [CrossRef]
35. Jhamtani, H.; Berg-Kirkpatrick, T. Learning to Describe Differences Between Pairs of Similar Images. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; pp. 4024–4034. [CrossRef]
36. Kim, H.; Kim, J.; Lee, H.; Park, H.; Kim, G. Viewpoint-Agnostic Change Captioning with Cycle Consistency. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 2075–2084. [CrossRef]
37. Tu, Y.; Li, L.; Yan, C.; Gao, S.; Yu, Z. R^3Net:Relation-embedded Representation Reconstruction Network for Change Captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, online, Punta Cana, Dominican Republic, 7–11 November 2021; Moens, M.F., Huang, X., Specia, L., Yih, S.W.t., Eds.; pp. 9319–9329. [CrossRef] [CrossRef]
38. Yu, X.; Wang, J.; Zhao, Y.; Gao, Y. Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognit.* **2023**, *135*, 109131. [CrossRef] [CrossRef]
39. Liu, C.; Chen, K.; Zhang, H.; Qi, Z.; Zou, Z.; Shi, Z. Change-Agent: Toward Interactive Comprehensive Remote Sensing Change Interpretation and Analysis. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef] [CrossRef]
40. 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] [CrossRef]
41. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef] [CrossRef]
42. Shi, J.; Zhang, M.; Hou, Y.; Zhi, R.; Liu, J. A Multitask Network and Two Large-Scale Datasets for Change Detection and Captioning in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef] [CrossRef]
43. Garcia-Garcia, A.; Orts, S.; Oprea, S.; Villena-Martinez, V.; Rodríguez, J.G. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
44. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; Isabelle, P., Charniak, E., Lin, D., Eds.; pp. 311–318. [CrossRef]
45. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81. [CrossRef]
46. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005 ; Goldstein, J., Lavie, A., Lin, C.Y., Voss, C., Eds.; pp. 65–72. [CrossRef]
47. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. [CrossRef]

48. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210. [CrossRef]
49. 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.* **2022**, *19*, 1–5. [CrossRef] [CrossRef]