remote sensing

MDPI

Article

Exploring Difference Semantic Prior Guidance for Remote Sensing Image Change Captioning

Yunpeng Li [1,2], Xiangrong Zhang [3], Guanchun Wang [3,*] and Tianyang Zhang [3]

[1] School of Electronics & Information Engineering, Wuxi University, Wuxi 214105, China; ypli@cwxu.edu.cn
[2] Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi University, Wuxi 214105, China
[3] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China; xrzhang@mail.xidian.edu.cn (X.Z.); zhangtianyang@xidian.edu.cn (T.Z.)
* Correspondence: wangguanchun@xidian.edu.cn

Highlights

What are the main findings?

- A novel RSICC framework unifies both the difference context and semantic changes on the RSI pair, enhancing cross-modal association with cross refined attention. The network comprises three core modules: the dual-branch difference extraction layer, the difference comprehending module, and the cross refined attention block.

What are the implication of the main finding?

- For the alignment of common/difference context features, a dual-branch difference extraction layer composed of symmetric difference context encoding and shallow difference auxiliary encoding is designed to seek the multi-scale difference features.
- In order to learn semantic knowledge in difference representations, a trainable difference comprehending module can generate high-level linguistic embeddings about appeared changes, and then the detected change content and embedding textual cues are integrated into a novel cross refined attention block for more cross-modal inter-actional information.

Abstract

Understanding complex change scenes is a crucial challenge in remote sensing field. Remote sensing image change captioning (RSICC) task has emerged as a promising approach to translate appeared changes between bi-temporal remote sensing images into textual descriptions, enabling users to make accurate decisions. Current RSICC methods frequently encounter difficulties in consistency for contextual awareness and semantic prior guidance. Therefore, this study explores difference semantic prior guidance network to reason context-rich sentence for capturing appeared vision changes. Specifically, the context-aware difference module is introduced to guarantee the consistency of unchanged/changed context features, strengthening multi-level changed information to improve the ability of semantic change feature representation. Moreover, to effectively mine higher-level cognition ability to reason salient/weak changes, we employ difference comprehending with shallow change information to realize semantic change knowledge learning. In addition, the designed parallel cross refined attention in Transformer decoder can balance vision difference and semantic knowledge for implicit knowledge distilling, enabling fine-grained perception changes of semantic details and reducing pseudochanges. Compared with advanced algorithms on the LEVIR-CC and Dubai-CC datasets, experimental results validate the outstanding performance of the designed model in RSICC tasks. Notably, on the

Check for updates

Academic Editor: Zhenwei Shi

Received: 18 November 2025
Revised: 26 December 2025
Accepted: 2 January 2026
Published: 11 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

LEVIR-CC dataset, it reaches a CIDEr score of 143.34%, representing a 3.11% improvement over the most competitive SAT-cap.

Keywords: change captioning; bi-temporal remote sensing images; semantic comprehending; difference enhancement; hierarchical feature fusion

1. Introduction

In today's digital era, there are massive remote sensing images (RSIs) that conveys complex land-covers in the same locality. The analytics for bi-temporal remotely sensed earth observation are required to develop more efficient information processing networks. However, change detection [1,2] for bi-temporal RSIs is a informative pixel-level task, which outputs a per-pixel "from-to" change map. An intuitive solution to automatically analyse the changed content is to summarize land cover changes with natural language sentences, namely remote sensing image captioning (RSICC). There is no doubt that the change captioning can make intelligent decisions for unprofessional user, showing more potential application value than change detection task in urban planning, agricultural estimation, military intelligence gathering, etc. Note that the bi-temporal RSIs are captured at the same location but at different times, which is also referred to as RSI pair.

RSICC is an extended task from remote sensing image captioning. The RSI captioning is a two-step process that encodes a RSI and then generate relevant textual descriptions. Specifically, RSICC networks start with inputting RSI pair and extract bi-temporal features for bi-temporal RSIs. The genuine changes learning is an indispensable role in RSICC task, which are then processed by caption generator for change captioning. Apparently, RSICC is more challenging than RSI captioning, which understands the independent contents [3–5] of RSI pair and further reasons discrepancy between them with linguistic sentences.

In recent years, several datasets and encoder-decoder networks for RSICC task have been proposed. Each RSI pair is well-aligned from one region, including T1 image (before the change) and T2 image (after the change). Many RSICC methods can be roughly divided into three processes: image encoding, difference encoding and change decoding. The parallel feature encoder is generally adopted in RSICC task based on convolution neural networks (CNNs), which has the ability to capture the independent T1/T2 features. The primary change detection is a subtraction/interaction function from T2 to T1, abbreviated as "T2-to-T1". Then Recurrent Neural Networks (RNNs) or Transformer is built for locating the changed element and captioning changes. The pioneer work [6] released a small Dubai-CC dateset. However, there will be more change scene in the real world, which support in-depth exploration for RSICC task. As such, LEVIR-CC dataset was conducted by Liu et al. [7], the changes were simultaneously extracted by a dual change Transformer extractor. The Transformer framework for RSICC network [8] has significantly exhibited performance superior. Therefore, a novel Chg2Cap Transformer-based architecture [9] was explored, in which a fully hierarchical interaction integrated multi-head attention (MHA) played a crucial role in learning difference features. Sun et al. [10] thought the spatial relationship between bi-temporal pixels was crucial for understanding the complex RSI scene, the designed scene graph and dependency grammar enhanced (SGDCC) RSICC method that jointly optimized scene graph difference encoder and grammar enhanced Transformer decoder. To model reliable interaction alignment between the bi-temporal visual signals, the innovative difference information learning with a single-stream extractor network (SEN) [11] was adopted, in which the single extractor took place Siamese feature encoder and reduced the computational cost. Besides, Sun et al. [12] proposed a sparse focus Transformer (SFT) for reducing the parameter number and computational

complexity within the dual Transformer change encoding network. However, they either suffered from the drawbacks of ignoring the multi-level change context awareness, in which simple multi-stage fusion module introduced limited functionality. For RSI change detection [13–15], multi-scale changes can provide rich content information with multiple receptive fields. Inspired by such designs, the progressive scale-aware network (PSNet) [16] utilized the multi-scale bi-temporal features for difference encoding that contained a progressive change perception layer and scale-aware reinforcement module. The results had garnered increasing attention and witnessed promising advancement for RSICC task, which was also applied to an interactive change-aware Transformer network (ICT-Net) [17]. Such approaches neglect that the knowledge facts can provide the explicit change information between the RSI pair, while measuring semantic relation for appeared changes is capable of distinguishing potential change information. To achieve this goal, Huang et al. [18] proposed a text-augmented change captioning method (TACC), which took advantage of the high-level semantic information for understanding the content relationship and learning changes in text-level. The emergence of Diffusion framework has significantly exhibited computer superior across the natural language processing and the computer vision. A novel diffusion feature extractor [19] is designed for multi-level and multi-time-step bi-temporal features, aiming to obtain most change-aware difference representations. Nevertheless, the designed implicit priors are still difficult to explore potential of cross-stage difference. Recently, to fully exploit foreground visual changes, Li et al. [20] constructed a pipeline framework that utilized explicit change masks combined with multi-temporal features to generate better change captions, which was introduced pixel-level change detection branch into RSICC task. Therefore, an efficient RSICC model should be capable of benefiting the difference context-aware enhancement and reducing the aliasing pseudo changes.

In summary, these endeavors mainly model bi-temporal interaction on deep change features, focusing on the conspicuous changes. There are some weaknesses: (1) Unexciting interaction for bi-temporal deep features with multi-level fusion does not align the similar or dissimilar information. In fact, feature perturbation will be existing in "T2-to-T1" difference encoding, it may overwhelm the seen changes. We argue that "T1-to-T2" difference also stores semantic changes, the combination difference encoding (i.e."T1-to-T2", "T2-to-T1") is helpful to differentiate unchanged/changed context features. (2) Changes with weak features might be disregarded, which is hard to localize at representation level. The more fine-grained change will ensure discriminative changes and supplement deep difference features for better reasoning genuine changes. Such multi-scale difference observation in RSICC task also inspires us to pay more attention to multi-scale difference learning. (3) Directly reasoning the captured difference features fails to sufficiently mine reliable change semantics for the language decoder. Although each RSI pair embodies with five captions, the semantic-guided role is negative for difference encoding module. Due to the complexity of translating from change features to language domains, it is necessary to define some knowledge prompts for extracting text differences, which learns change information of interest in a text manner.

In this paper, a RSICC model based on difference semantic prior guidance has been developed. Figure 1 shows the flowchart of our proposed model, which consists of two level difference learning, namely context-aware difference learning and difference comprehending along with a language generator. For context-aware difference learning module, the input RSI pair is firstly processed by Siamese CNN encoders to extract multi-scale bi-temporal features, these features are then passed into difference learning layer. The context-aware difference learning module dual branches is employed for capturing multi-scale change contents, in which the symmetric difference context encoding directs the focus towards visually relevant information and the shallow difference auxiliary encoding

models changes considering weak features. Based on the richness of change details from shallow branch, the difference comprehending layer learns a knowledge-aware difference representation about shallow changes, this semantic representation performs more implicit reasoning to promote decoder step. In next decoding stage, the multi-aspect difference embeddings are fed into the designed cross refined attention incorporated with single-layer Transformer for a sentence prediction. Extensive experiments show that our model gives a rise compared with state-of-the-art (SOTA) results on two public datasets.

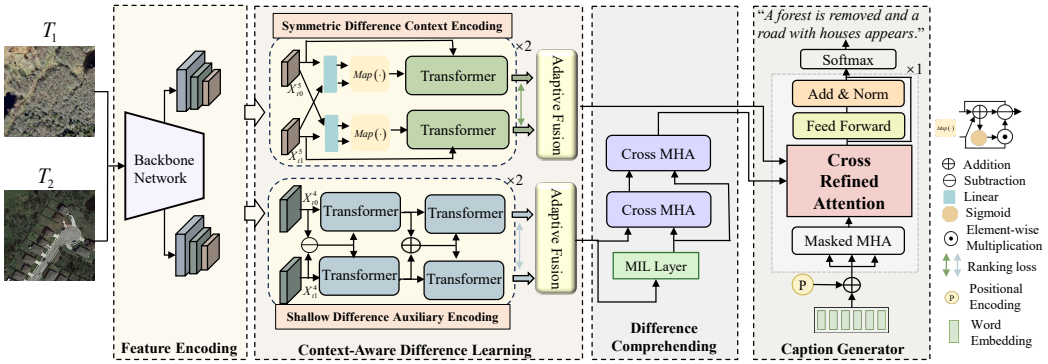


Figure 1. The overview of our proposed RSICC model. It consists of RSI pair encoding, context-aware difference learning, difference comprehending and caption generator. Herein, difference comprehending is the major component to learn semantic difference by shallow context-aware difference distilling.

The major contributions can be summarized as follows:

(1) A novel knowledge-driven RSICC model is proposed to learn higher-level cognition of changes. It efficiently captures the difference context and semantic changes on the RSI pair by encoding deep and shallow bi-temporal representations, giving decoder the benefit of identifying and reasoning changes.

(2) A dual-branch difference extraction layer composed of symmetric difference context encoding and shallow difference auxiliary encoding is proposed. The symmetric branch guarantees the alignment of common/difference context features, the auxiliary branch can distill detail changes with weak feature.

(3) In order to learn semantic knowledge in difference representations, the difference comprehending is designed. It characterizes specific change contents with the output of the auxiliary branch and build reliable cross-modal alignment through MHA layers, thus generating a high-level linguistic embeddings about appeared changes.

(4) The constructed cross refined attention is an extension of MHA, which has been incorporated to Transformer decoder. Therefore, our decoder fully explores the potentials between detected change content and embedding textual cues, benefiting more cross-modal interaction and the change captioning performance.

2. Related Work

RSICC task is new vision-language task that incorporates single RSI captioning [21–23] into change analysis for the RSI pair, which involves the RSI pair difference learning and change captioning. The primary objective of discerning "effective changes" is critical for obtaining change captions. The prevailing difference learning in RSICC methods can be summarized in two mainstreams: cross-temporal change location and pixel-level change mask.

2.1. Cross-Temporal Change Location Methods

In the literature, the Siamese encoder (i.e., CNN) encodes given RSI pair and a cross-temporal interaction module performs feature enhancement and change location, and a decoder (i.e., Transformer) converts change features into natural language. For example, a fully Transformer encoder-decoder RSICC method [24] adopted Siamese features extractor and MCCFormers-D/S interaction module for locating multiple changes. Inspired by MCCFormers, a improved dual-branch Transformer difference encoder [7–9] emerged. Cai et al. [17] treated interactive change-aware module as key changes perception and utilized the predicted difference embeddings to guide the model for more precise change caption generation. Li et al. [25,26] further proposed a cross-spatial Transformer and symmetric difference localization network, the former enhanced spatial-aware feature representations, the latter aimed to learn symmetric differential features. To better capture the spatial relationship of the changing area, Sun et al. [10] proposed a novel scene graph difference encoder to model the spatial relationship, thus providing richer spatial context information. In addition, the decoder with syntax knowledge for revealing the dependency relationship with generated sentence. Furthermore, Huang et al. [18] took advantage of the rich semantic information provided with pre-defined text prompt, the bi-temporal features were firstly extracted by the FastSAM encoder and then enhanced by CLIP encoder, which were fed into text-guidance difference capture module for capturing semantic-level changes. To overcome limitations of the receptive field (i.e., CNN-, Transformer-based extractor) and computational complexity, Liu et al. [27] proposed a novel RSCaMa model, the spatial difference-aware SSM module was designed with multiple CaMa layers, following with the temporal traversing SSM to enhance the temporal information understanding and interaction. Sun et al. [12] implemented a sparse focus attention mechanism with fewer parameters. To improve Siamese feature encoder with more computational costs, Zhou et al. [11] focused on differences and content with a single-stream extractor between the RSI pair. For building all the multi-scale representations, the multi-scale PSNet [16] method that effectively learned multi-scale bi-temporal features, containing a progressive change perception layer and scale-aware reinforcement module. Later, data augmentation for LEVIR-CC dataset had been explored by Karimli et al. [28] on PSNet. Notably, there are multiple land cover changes of more informative contents in the RSI pair. Instead of obtaining single change sentence, Yang et al. [29] tended multiple captions with complex grammars instead of solely describing what changes happen, a novel cascade information network was proposed to model change features, which would be iteratively updated in linguistic modules to with the help of defined probability case and information theory. Compared to existing CNN- and Transformer-based RSICC methods, the diffusion model offers forward process to add corruption noise and reverse steps to denoise features [30,31]. Yang et al. [19] introduced a jointly trained model that combined the diffusion difference-aware module with Transformer caption generation, providing rich semantic change information for the RSICC model. Bai et al. [32] integrated bayesian diffusion prior to deal with uncertainty and noise in bi-temporal features, the semantic-level perception of change regions was modeled by designed double-layer multi-coding and an image-text approach.

Previous approaches treat difference encoding process with cross-temporal interaction involving the multi-scale, linguistic-driven and diffusion-based change perception. Although they attempt to capture discriminative change regions, it still encounters difficulties in the insufficient, unnecessary or misleading difference information. To address these aforementioned challenges, we attempt to describe the content differences between bi-temporal RSIs, which makes use of the multi-scale difference feature guided with the high-level semantic information inherent in changed scenes.

2.2. Multi-Task Change Aggregation Methods

For making change information more granular, robust and precise in RSICC task [33–35], the pixel-level semantic guidance from RSI change detection can alleviate the latent knowledge between temporally distinct RSIs, representing with a binary change mask. The more accurate located visual change areas can confirm the effectiveness of difference encoding and enhance RSICC performance. To achieve comprehensive change interpretation and insightful analysis, the proposed change-agent [36] network balanced pixel-level change detection and semantic-level change captioning, multi-level change interpretation branch for changed masks and a large language model for change description simultaneously. Shi et al. [37] also tried simultaneously to solve both pixel-level change detection and RSICC tasks with the designed CNN-Transformer-based multitask network, including a shared Siamese backbone, difference feature enhancement and fusion and a multitask decoder for generating change pixel-level mask and captioning. Li et al. [20] integrated change detection branch with a binary mask to enhance difference learning, then processing the multi-temporal difference feature fusion (i.e., C-Stream, N-Stream) for distinguishing foreground changes and background changes. To facilitate information interaction between heterogeneous tasks [38], a semantic-cc [39] was gathered a bi-temporal SAM-based encoder, a multi-task semantic aggregation branch, a supervised pixel-level change detection and a change caption decoder based on the large language model. Sun et al. [40] refined the multi-scale change detection results with an well-designed diffusion model that managed high-frequency noise throughout the diffusion process, modeling paradigms more on feature distribution learning.

It is important to highlight the capacity of RSICC model to recognize various changes and maintain high sensitivity across diverse scales and complex semantic difference. In the future work, we plan to develop pixel-level semantic change to filter or correct inaccurate textual guidance to form more powerful difference representation.

3. Method

In this section, we detail our proposed RSICC architecture, which mainly consists of three parts: context-aware difference learning for more powerful difference representation, a difference comprehending for expressing change semantic information and a change caption decoder for decoding all cross-modal difference features. We firstly describe the standard RSICC algorithm, then show we implement a context-aware difference learning which consists of a symmetric difference context encoding and a shallow difference auxiliary encoding, to select crucial information from all multi-scale difference features. In addition, we introduce a difference comprehending procedure and the designed Transformer inference process as well as the model training process. The overall structure of our method is shown in Figure 1.

3.1. Standard Change Captioning

For the RSICC task, a standard encoder-decoder framework with three-steps, it firstly encodes given RSI pair to bi-temporal features X_{t_0} and X_{t_1}, where X_{t_0} and X_{t_1} denotes "before" and "after" RSI features separately. Then, difference encoding layer is inserted following bi-temporal features to calculate the change features D_{dif}. Finally, a change description $y = \{y_1, y_2, ..., y_T\}$ is generated according to previously generated words and change features D_{dif}, where T is the number of words in the generated sentence.

In the training procedure, the RSICC model with parameter θ are generally trained to minimize the conditional cross-entropy loss, where the loss function considers the distribution over possible output sentences y:

$$L_{ce} = \sum_{t=1}^{T} \log p_\theta(y_t | y_{<t}) \tag{1}$$

During inference procedure, a word is produced at t time and then embedded into the decoder with semantic changes to predict the next word. The sequential decoding is tend to copy semantics from the generated phrases to enhance the grammatical accuracy. Once the guidance information is not strong for decoder, the certain inertia will easily cause semantic error and limits the diversity of change descriptions. Therefore, localizing effective change features is a important factor leading to generate a high-level linguistic sentence.

3.2. Image Pair Encoder

RSIs contain richer background and structural information of objects, the small-scale object may be easily overlooked or misjudged based solely on deep features, failing to recapitulate detail of an RSI. Previous RSICC works under multi-scale features have achieved great success. In such a Siamese framework, Siamese CNNs are used to extract multi-scale bi-temporal features for given RSI pair. To be specific, We use Siamese ResNet-101s to extract multi-scale features, which consists of five stages (C1–C5). The C5 feature map is generally considered to capture high-level visual information with large receptive field. In contrast, The C4 feature map may retain objects with week information. To balance between representational capacity and computational efficiency, C4 and C5 feature maps are extracted in our model. After presenting the RSI pair $\{I_{t0}, I_{t1}\}$ to the model, deep bi-temporal features $\{F_{t0}^5, F_{t1}^5\}$ and shallow bi-temporal features $\{F_{t0}^4, F_{t1}^4\}$ are operated with our Siamese encoders.

3.3. Context-Aware Difference Learning

We notice that some existing methods capture differences between RSI pair only based on bi-temporal deep features, while neglecting the use of more comprehensive features like $\{F_{t0}^4, F_{t1}^4\}$. We argue that, to learn unchanged/changed features between bi-temporal content, the context features of commonality and difference should be encapsulated with multi-scale bi-temporal features. To obtain common and difference context features, we propose a context-aware difference learning module, which is composed of a symmetric difference context encoding and a shallow difference auxiliary encoding. Notably, the top/down difference branch has two same layers. Hence, the difference information is generated features at multiple levels, detecting subtle changes and distinguishing genuine alterations from complex scene pairs. Specifically, $\{F_{t0}^5, F_{t1}^5\}$ is the input for top difference branch as well as $\{F_{t0}^4, F_{t1}^4\}$ for down difference branch.

Symmetric Difference Context Encoding: It works interactively to capture the symmetric changes between $\{F_{t0}^5, F_{t1}^5\}$, the symmetric map function is defined with addition, Sigmoid, wise multiplication and subtraction. The flow for map is illustrated in Figure 1. For former three steps in map function, the relevant information between bi-temporal features will be enhanced, then subtraction operation can furthest distinguish changed/unchanged features as well as considering contextual features. Moreover, compared with traditional "T2-to-T1" difference extraction, we think symmetric differences represent the same semantic changes, which should be consistent. Therefore, the symmetric difference learning

can enhance the consistency between invariant and change-aware contextual features. Let $T_{t0}^{0} = F_{t0}^{5}$ and $T_{t1}^{0} = F_{t1}^{5}$. Formally, the formulas are shown as follows:

$$U_0^{l+1} = Map\left(T_{t0}^{l}, T_{t1}^{l}\right) \tag{2}$$

$$U_1^{l+1} = Map\left(T_{t1}^{l}, T_{t0}^{l}\right) \tag{3}$$

where $l = 0, 1$, l represents one layer within the two difference layers. However, the representational power of $\left\{U_0^{l+1}, U_1^{l+1}\right\}$ is quite limited because it suffers from locational variance causing difference accuracy problems. Then, we adopt Transformer structure further enhances the change awareness through the incorporation by considering symmetry for semantic changes. The formulas are shown as follows:

$$T_0^{l+1} = Transformer\left(T_{t0}^{l}, U_0^{l+1}\right) \tag{4}$$

$$T_1^{l+1} = Transformer\left(T_{t1}^{l}, U_1^{l+1}\right) \tag{5}$$

where the top difference context features of each layer are represented as $\left\{T_0^1, T_1^1\right\}$ and $\left\{T_0^2, T_1^2\right\}$, respectively. For guiding Transformer layer to mine difference context features, the consistency constraints are customized between $\left\{U_0^{l+1}, U_1^{l+1}\right\}$. These multi-level features are calculated for mean features, then projected and normalized, generating $\left\{u_0^{l+1}, u_1^{l+1}\right\}$. Next, we introduce the ranking loss [25,41] to optimize their alignment:

$$L_T^{l}\left(u_0^{l+1}, u_1^{l+1}\right) = \frac{1}{2}\sum\left\{\begin{array}{l}\max\left[0, \gamma - s\left\{u_0^{l+1}, u_1^{l+1}\right\} + s\left\{u_0^{l+1}, u_1^{(l+1)-}\right\}\right] \\ +\max\left[0, \gamma - s\left\{u_1^{l+1}, u_0^{l+1}\right\} + s\left\{u_1^{l+1}, u_0^{(l+1)-}\right\}\right]\end{array}\right\} \tag{6}$$

where γ is a pre-defined parameter. Firstly, the similarity matrix will be calculated by cosine similarity for $u_0^{l+1} \to u_1^{l+1}$ and $u_1^{l+1} \to u_0^{l+1}$. The diagonal elements of the similarity matrix represent positive samples $\left\{u_0^{l+1}, u_1^{l+1}\right\}$ and $\left\{u_1^{l+1}, u_0^{l+1}\right\}$, others for the negative samples like $\left\{u_0^{l+1}, u_1^{(l+1)-}\right\}$ and $\left\{u_1^{l+1}, u_0^{(l+1)-}\right\}$. Then, $s\{\cdot\}$ computes the similarity gap between positive and negative samples achieves feature alignment optimization.

Shallow Difference Auxiliary Encoding: Generally, the small-scale objects may be disturbed by the surrounding environment, especially when small-scale objects appear in large-scale RSIs. Once difference encoding is only dependent on $\left\{T_0^{l+1}, T_1^{l+1}\right\}$, there will be a shortage, forming unstable difference. The down difference branch with two layers is the other component in context-aware difference learning, which is introduced to learn difference context feature restoring fine-grained details between the RSI pair. Specifically, a subtraction is firstly employed for the difference. Then, each layer stacks two Transformers to construct cross-temporal interaction for bi-temporal shallow features. Let $D_{t0}^{0} = F_{t0}^{4}$ and $D_{t1}^{0} = F_{t1}^{4}$. The process for the first Transformer can be written as follows:

$$V_0^{l+1} = Transformer\left(D_{t0}^{l}, D_{t1}^{l} - D_{t0}^{l}\right) \tag{7}$$

$$V_1^{l+1} = Transformer\left(D_{t1}^{l}, D_{t1}^{l} - D_{t0}^{l}\right) \tag{8}$$

where $l = 0, 1$. The second Transformer can be defined as:

$$D_{t0}^{l+1} = Transformer\left(V_0^{l+1}, V_0^{l+1} + V_1^{l+1}\right) \tag{9}$$

$$D_{t1}^{l+1} = Transformer\left(V_1^{l+1}, V_0^{l+1} + V_1^{l+1}\right) \tag{10}$$

where the shallow difference context features of each layer are represented as $\{D_0^1, D_1^1\}$ and $\{D_0^2, D_1^2\}$, respectively. It can be noticed that these change features also contain the semantic-change relationship, the symmetric supervision with ranking loss is adopted to pull semantically close information and push away non-related content. We apply an adaptive-avg layer on $\{D_0^1, D_1^1\}$ and $\{D_0^2, D_1^2\}$ extracts global change context $\{d_0^1, d_1^1\}$ and $\{d_0^2, d_1^2\}$:

$$L_D^l\left(d_0^{l+1}, d_1^{l+1}\right) = \frac{1}{2}\sum\left\{\begin{array}{l}\max\left[0, \gamma - s\left\{d_0^{l+1}, d_1^{l+1}\right\} + s\left\{d_0^{l+1}, d_1^{(l+1)-}\right\}\right] \\ +\max\left[0, \gamma - s\left\{d_1^{l+1}, d_0^{l+1}\right\} + s\left\{d_1^{l+1}, d_0^{(l+1)-}\right\}\right]\end{array}\right\} \tag{11}$$

where the parameters are same with Equation (6), $\left\{d_0^{l+1}, d_1^{l+1}\right\}$ and $\left\{d_1^{l+1}, d_0^{l+1}\right\}$ represent positive samples, the negative samples are $\left\{d_0^{l+1}, d_1^{(l+1)-}\right\}$ and $\left\{d_1^{l+1}, d_0^{(l+1)-}\right\}$. Through the consistency supervision, the multi-scale difference semantics can be enforced with aligned distributions. The final consistency constraint loss is defined as following:

$$L_{td} = \beta_1 L_T^l + \beta_2 L_D^l \tag{12}$$

where we set $\beta_1 + \beta_2 = 1$.

Adaptive Fusion: To more effectively couple the multi-scale difference from from the context-aware difference awareness stage, a adaptive fusion block is adopted to adaptively fuse the multi-level change-aware representations, which contains GLU activation function and multi-level additions. For deep fused difference representation, we set $T_1 = \{T_0^1; T_0^2\}$, $T_2 = \{T_1^1; T_1^2\}$ and $T = \{T_1; T_2\}$. Applying $D_1 = \{D_0^1; D_0^2\}$, $D_2 = \{D_1^1; D_1^2\}$ and $D = \{D_1; D_2\}$ serves for shallow fused difference representation, ";" denotes concatenation. The coupling operation of the adaptive fusion can be expressed as follows:

$$Z_T = GLU(T) + W_{t1}T_1 + W_{t2}T_2 \tag{13}$$

$$Z_D = GLU(D) + W_{d1}D_1 + W_{d2}D_2 \tag{14}$$

where $W = \{W_{t1}, W_{t2}, W_{d1}, W_{d2}\}$ are the learnable weights.

3.4. Difference Comprehending

Previous RSICC methods, fused multi-scale change features, supplemented the deep difference features with the shallow difference features. The fusion generally relies heavily on up-sampling and down-sampling operations in visual-spatial level. However, there exists semantic gap in RSICC task, directly modeling visual information and text in the same semantic space is impossible. Therefore, we expect to explore essentially change words for the different content, providing explicit change semantic information. We argue that the deep difference features can correlate and mine locally common/different context features for deducing locally difference features, the shallow difference features can augment the locally difference features to ensure all changes are distilled. Such the shallow difference features are adopted as the input of difference comprehending layer, consisting of a change prediction layer and two cross attention layers.

In specific, some change words are selected and listed in the vocabulary head. These words are encoded with one-hot index matrix, defined as A. Firstly, we model the similarity between the shallow difference features and the change word embedding:

$$P_c = Sigmoid\left((W_a E A)^T \otimes W_z Z_D\right) \tag{15}$$

where P_c denotes the probability of predicting change words, $Sigmoid(\cdot)$ represents activation function, W_a and W_z are trainable parameters, E denotes the embedding matrix, $\otimes$ denotes the matrix multiplication. However, the predicted words may be unrelated to change content, the final probability p_i for containing i^{th} word is calculated as follows:

$$p_i = 1 - \prod_{Z_D} \left(1 - P_c^i\right) \tag{16}$$

In general, given a RSI pair will be associated with multiple change semantics. Thus, top-m probability for p_i are selected and transformed with $S = s_1, s_2, \ldots, s_m$, S are served as primary semantic change cues to assist in optimizing the decoder for optimizing our overall network. Afterward, the S is passed to two cross-attention with features Z_D:

$$\tilde{S} = MHA(MHA(Z_D, S), S) \tag{17}$$

the $\tilde{S}$ ensures the difference comprehending to learn the fine alignment between the change features and the corresponding words. In addition, the focal loss proposed in [20] is leveraged for training the difference comprehending in an end-to-end manner, which can be formulated as:

$$L_f = \sum_m \left\{ -l_i \delta (1 - p_i)^\gamma \log(p_i) - (1 - l_i)(1 - \delta) p_i^\gamma \log(1 - p_i) \right\} \tag{18}$$

where m denotes the number of the predicted words, $l_i = 1$ for i^{th} word in the GT captions, $l_i = 0$ for opposite case, and δ and γ are empirically set to 0.95 and 2.

3.5. Change Caption Decoder

For balancing difference features and change word embeddings during prediction process, a cross refined attention is used to interact cross-modal information to trigger sentence generation. As shown in Figure 2, the attention layer is employed to separately conduct cross-attention over the difference features Z_T and the semantics $\tilde{S}$ depending on the same query form masked MHA. One attention aims to find the associated change details contained in difference features, and the other attention captures the task-relevant semantic change by using inter-semantic information.

Concretely, the GT sentences will be embedded by masked-MHA, obtained the semantic query $\tilde{E}$. Subsequently, the query is used to select and balance the weights learned from Z_T and $\tilde{S}$, yielding the holistic change context Y_t:

$$Y_t = AN(FFN(MA_3(X_3) + MA_2(X_2) + MA_1(X_1))) \tag{19}$$

where $X_1 = \tilde{E} + MHA(\tilde{E}, \tilde{S}, \tilde{S})$, $X_2 = \tilde{E} + MHA(\tilde{E}, Z_T, Z_T)$ and $X_3 = \tilde{E}$, the $FFN(\cdot)$ is feed-forward layer, layer is utilized to process the summed output, $AN(\cdot)$ consists of residual connection and normalization. Such semantic feature has an effective impact on one layer Transformer-based decoder, while predicting the probability distributions of words via a Softmax layer.

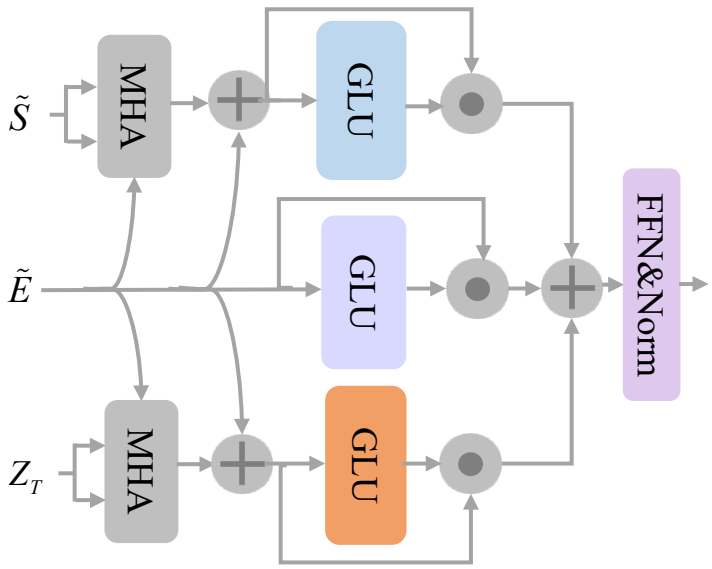


Figure 2. The cross refined attention in our Transformer decoder separates the change semantics and difference feature with parallel MHA, during which the semantic information is aggregated and aligned through the sentence embedding.

3.6. Training Strategy

At training stage, the overall objective of our model is measured as the integration of the ranking loss in context-aware difference learning L_{td}, the focal loss in difference comprehending L_f and the typical cross entropy loss defined in Equation (1):

$$L = \alpha_1 L_{ce} + \alpha_2 L_{td} + \alpha_3 L_f \tag{20}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. In training step, we set $\alpha_1 = 0.8$, $\alpha_2 = \alpha_3 = 0.1$ following [21,25].

4. Experiments and Analysis

In this section, we conduct extensive experiments including comparision, ablation and parameter analysis for our model. Section 4.1 gives the details of datasets, evaluation metrics, model settings and compared methods. Section 4.2 analyses the designed model and state-of-the-art approaches in our experiments.

4.1. Dataset and Setting

4.1.1. Datasets

We conduct all experiments on the widely used datasets, namely LEVIR-CC [7] and Dubai-CC [6]. For the input of RSICC method, the dataset structure is like <imageT1,imageT2,sentences>, each <imageT1,imageT2> is annotated with five different sentences with consistent change semantics. All annotations are divided into changed or unchanged sentences. The unchanged sentence is short, changed sentences are with different lengths.

(1) LEVIR-CC dataset [7]: Each <imageT1,imageT2> focuses on changes across various building instances (i.e., buildings). The dataset consists of 10,077 RSI pairs with span time from 5 to 14 years. Each image has 256 × 256 pixels with a resolution of 0.5 m/pixel. For annotations, the maximum length of change sentence is 39. Based on the official split [7], a split of of 6815, 1333, and 1929 of the <imageT1,imageT2,sentences> is used for training, validation and testing, respectively.

(2) Dubai-CC dataset [6]: The original bi-temporal RSIs are collected about Dubai urban environments, spanning from 19 May 2000 to 16 June 2010. In order to facilitate RSICC

task, each RSI is cropped with a size of 50 × 50 pixels. Therefore, there are 500 sliced <imageT1,imageT2>. For consistent experimental conditions with LEVIR-CC dataset, the RSI will be upsampled to 256 × 256 pixels. Notably, the changing scenarios are fewer than LEVIR-CC dataset. The longest annotated sentence contains 27 words, shorter than Dubai-CC dataset. We also follow the split [7] and use 300, 50 and 150 <imageT1,imageT2,sentences> for training, validation and testing respectively.

4.1.2. Evaluation Metrics

We adopted the standard image captioning assessment metrics for evaluation, i.e., BLEU-n [42], METEOR [43], ROUGE_L [44] and CIDEr [45]. These metrics are determined by assessing the proportion of correctly predicted words and semantics between predictions and annotated GT sentences. The n in BLEU-n represents n-element word group selected from predictions and given references, a weighted geometric mean of n-gram precision evaluates the accuracy and diversity of change captioning. L represents the longest common L-tuple word group calculated by ROUGE_L. In order to consider high-level matches beyond the word group, METEOR metric enforces alignment between generation and reference based on the WordNet synonyms. Compared to other evaluation metrics, CIDEr is a widely favored metric, which is specifically designed to evaluate the performance of image captioning. In detail, the mutual constraint of TF-IDF vectors in CIDEr can feature the semantic similarity and consistency between generated and annotated sentences.

4.1.3. Train Details and Experimental Setup

For bi-temporal RSIs, we set pre-trained Siamese ResNet-101s to obtain multi-scale bi-temporal features with size of $28 \times 28 \times 512$ and $14 \times 14 \times 1024$. All annotated sentences are represented with one-hot mapping and then padded with 0 for same length. Some symbols (i.e., <strat>, <end>) are appeared in each sentence. The maximum length of training and generated sentences is 50. The word embedding for Transformer decoder is with the dimensionality of 512.

For difference comprehending module, the change words are predicted for more semantic change representation. The change targets or phenomenon will be considered as change words, which are noun in the annotated sentences. Then, these selected words are sorted by word frequency. The following other words together form the vocabulary. The number of words for LEVIR-CC and Dubai-CC vocabulary is 463 and 328, respectively. Besides, the number of change words for LEVIR-CC and Dubai-CC is 120 and 60, respectively.

Our proposed model is powered on a single 24 GB NVIDIA GPU (GeForce RTX 4090) with the PyTorch framework. Considering GPU memory, the batch size is set to 16 during training with the maximum epoch 35. For model optimization, the Adam algorithm is chosen for network training. The initial learning rate of the encoder and decoder is equal to 1×10^{-4}, the decay factor is 0.8 which plays a role with no performance increases on the BLEU4 score within three epochs. Based on adjusted learning rate, there is no effective performance growth within ten epochs, our model training is finished.

For context-aware difference learning block, the number of difference encoding layer is 2. In the decoder, the Transformer layer is one, and the word embedding size is 512. In order to get better change captions, the beam search is uniformly set to 3.

4.1.4. Compared Models

To make comparison between our proposed model and other models, the following RSICC methods are chosen to conduct our all experiments. In the following models—except for DUDA—the change caption generator is based on Transformer framework.

DUDA [46]: DUDA taskes subtracted bi-temporal features as difference input for dual spatial attention. A basic LSTM is adopt to output change captions.

MCCFormers [24]: Fully Transformer framework adopts two change encoder (i.e., MCCFormers-D and MCCFormers-S) for multiple changes.

RSICCformer [7]: Compared with MCCFormers, the main difference in RSICCformer is to capture and recognize multiple changes with a dual-branch Transformer encoder.

Chg2Cap [9]: The overall architecture is similar to RSICCformer, but it adopts an attentive encoder in place of the dual-branch change encoder for attending change features.

SEN [11]: In SEN, bi-temporal features are captured by a pre-trained single feature extractor. Furthermore, a designed difference feature encoder enhances difference modeling explicitly, making single word or sentence to completely depict land cover changes.

ICT-Net [17]: The interactive change awareness follows the design of a dual multi-scale feature extractor. ICT-Net extra introduce cross gated-attention into decoder for better capturing the semantic changes.

Diffusion [19]: The diffusion feature extractor is used for RSICC task. Then, the improvement of change captioning is achieved by utilizing the extracted diffusion features from a time-channel-spatial attention mechanism.

SFT [12]: SFT network incorporates a sparse focus attention into Transformer difference encoder, capturing changing information with fewer parameters.

SAT-Cap [8]: The difference encoder is implemented by spatial-channel attention encoder and a difference-guided fusion module is adopted to improve change details preservation.

TACC [18]: TACC can be seen as a text-driven RSICC method considering multi-scale bi-temporal features. The semantic difference learning is modified by text-guided difference capture module.

4.2. Evaluation Results and Analysis

To demonstrate the performance of our proposed method, the quantitative results of our model and the current popular RSICC methods are shown in Tables 1 and 2, which are conducted on LEVIR-CC and Dubai-CC dataset. The performance of those methods is evaluated using a series of metrics including BLEU-n, METEOR, ROUGE_L and CIDEr. The value of each metric ranked in bold is taken as the best value. Some example of the predicted change semantics and the generated change caption for bi-temporal RSIs are shown in Figure 3.

From the results shown in Tables 1 and 2, it can be observed that our model has a higher success in terms of key evaluation indicators except for BLEU4. The BLEU4 obtained by Diffusion is slightly higher than our network and performs best among all compared methods. It is worth noting that the BLEU4 metric has higher persuasiveness than other BLEU-n score, which measures the co-occurrences of 4-gram in generated sentences and GT sentences. The highest BLEU4 score shows that diffusion difference learning can stimulate the improvement of change captions on some metrics. In detail, TACC in the compared SOTA methods utilize linguistic guidance to design a supervised RSICC network for enhanced the identification of changed semantics. The results obtained show that the TACC outperforms traditional difference learning like Chg2Cap and SEN in terms of CIDEr. However, our model achieves 1.4334 on CIDEr, which is 6.17% higher than TACC. It is because the difference comprehending layer models semantic changes effectively, and the key idea extracting difference features at semantic level is essential for sequential sentence generation. In addition to the above comparisons, we also have a comparable performance with the most competitive SAT-Cap model (1.4023 CIDEr), a 3.11% improvement over all compared methods, other metrics also scores also demonstrate better performance.

Therefore, the improvement for our proposed model is particularly obvious. It may be because that the difference context awareness and change semantics have good abilities to make the caption results more coherent, which is more consistent with our motivation.



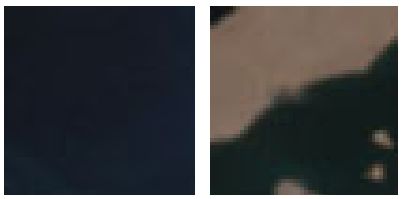
(a)

GT Captions
(1) Some villas are built in the forest.
(2) Villas are built in the woods with a path to the road.
(3) Some separate buildings appear in the forest.
(4) Several buildings have been built in the forest.
(5) Some buildings are constructed in the woods.
A1: houses, road, woods
A2: Some houses are constructed in the woods along the road.



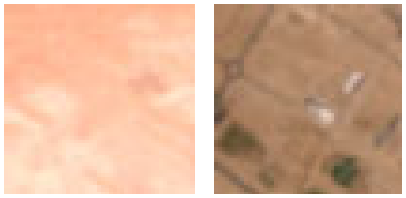
(b)

GT Captions
(1)All the trees disappear and some new houses are built near the corner of the road.
(2)The vegetation has been removed and a road and several houses are constructed on the bareland.
(3) The forest disappears and two roads and a group of houses appear.
(4)Trees are removed and a turning road appears at the top and a row of villas appears at the bottom.
(5)A road has been built at the top and some villas at the bottom replace all the trees.
A1: road, villas, vegetation
A2: The vegetation has been removed and a road is built with villas built along.



(c)

GT Captions
(1) A wide island was realized in the sea.
(2) A big artificial island has been built in the sea.
(3) One big and some small artificial islands appear in the sea.
(4) New artificial islands are present in the sea.
(5) An island has been created in the sea.
A1: sea, island, sand
A2: An artificial island was built in the sea.



(d)

GT Captions
(1) The area has been urbanized.
(2) New roads and constructions are present in the desert.
(3) Some buildings and a roundabout appear in the desert.
(4) The desert has been urbanized in the left.
(5) Some roads along with a roundabout and some green areas appear in the desert.
A1: roads, buildings, desert
A2: Some buildings and roads appear in the desert.

Figure 3. Three change semantics (A1) predicted by difference comprehending and change captioning results (A2) achieved by our proposed model. Two examples, (**a**,**b**), are from LEVIR-CC dataset, (**c**,**d**) for Dubai-CC dataset. Blue words in five GT captions represent change semantics, orange words represent appropriate semantic information and red words represent wrong predictions.

In order to objectively evaluate the model efficiency of our method and some recent methods, the parameters and the inference speed (images per second) are reported shown in Table 3. Due to structural differences of the Diffusion, SFT and SAT-Cap, the comparison is conducted on Ours, RSICCformer, Chg2Cap, SEN, ICT-Net and TACC. Obviously, our method with complex structure has a disadvantage in terms of parameter quantity and inference time, which is also not conducive to industrial implementation. In the future, exploring lightweight RSICC models is one of our main explorations.

Table 1. Comparison scores of our method and other state-of-the-art methods on LEVIR-CC dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
DUDA	0.8144	0.7222	0.6424	0.5779	0.3715	0.7104	1.2432
MCCFormers-S	0.8216	0.7295	0.6542	0.5941	0.3826	0.7210	1.2834
MCCFormers-D	0.8049	0.7111	0.6352	0.5734	0.3823	0.7140	1.2685
RSICCformer	0.8411	0.7540	0.6801	0.6193	0.3879	0.7302	1.3140
Chg2Cap	0.8614	0.7808	0.7066	0.6439	0.4003	0.7512	1.3661
SEN	0.8510	0.7705	0.7001	0.6409	0.3959	0.7457	1.3602
ICT-Net	0.8606	0.7812	0.7145	0.6612	0.4051	0.7521	1.3836
Diffusion	0.8628	0.7750	0.7109	**0.6693**	0.4016	0.7537	1.3861
SFT	0.8456	0.7587	0.6864	0.6287	0.3993	0.7469	1.3705
SAT-Cap	0.8614	0.7819	0.7144	0.6582	0.4051	0.7537	1.4023
TACC	0.8549	0.7741	0.7052	0.6462	0.4007	0.7496	1.3717
Ours	**0.8649**	**0.7881**	**0.7221**	0.6671	**0.4161**	**0.7681**	**1.4334**

Note: The bold numbers indicate the best results.

Table 2. Comparison scores of our method and other state-of-the-art methods on Dubai-CC dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
DUDA	0.5885	0.4359	0.3363	0.2539	0.2205	0.4834	0.6278
MCCFormers-S	0.5297	0.3702	0.2762	0.2257	0.1864	0.4329	0.5381
MCCFormers-D	0.6465	0.5045	0.3936	0.2948	0.2509	0.5127	0.6651
RSICCformer	0.6792	0.5361	0.4137	0.3128	0.2541	0.5196	0.6654
Chg2Cap	0.7204	0.6018	0.5084	0.4170	0.2892	0.5866	0.9249
SEN	0.7095	0.5728	0.4581	0.3625	0.2662	0.5595	0.9177
ICT-Net	0.6938	0.5703	0.4650	0.3617	0.2678	0.5731	0.9297
Diffusion	0.7318	**0.6136**	**0.5225**	**0.4541**	**0.3085**	0.6056	0.9647
SFT	0.7204	0.6018	0.5084	0.4170	0.2892	0.5866	0.9249
SAT-Cap	**0.7348**	0.6098	0.5051	0.4080	0.2962	0.5906	0.9774
TACC	0.7217	0.5865	0.4824	0.3859	0.2817	0.5893	0.9545
Ours	0.7234	0.6052	0.5030	0.4065	0.2940	**0.6266**	**1.0042**

Note: The bold numbers indicate the best results.

Table 3. Comparison of our framework and state-of-the-art methods in terms of Parameters and Inference Speed (images per second).

Methods	Parameters (M)	Inference Speed (s)
RSICCformer	56.20	12.67
Chg2Cap	32.81	39.58
SEN	39.90	3.79
ICT-Net	96.40	10.44
TACC	94.14	1.06
Ours	**110.97**	**12.75**

Note: The bold numbers indicate the parameters and inference speed for our model, while it is better that the number is more small.

As shown in Figure 3, there are four examples and visualization of the change semantics and captioning generation results. Note that the predicted change semantics are set to 3 on two RSICC datasets. We can see some new houses appear in Figure 3a,d scene, the predicted change semantics can cover main content in "after" image, in which "houses" and "buildings" with the same meaning are high-frequency vocabulary. For Figure 3b, the change captions balance change scene information in "before" and "after" image, the "vegetation" is inferred by difference comprehending module and appeared in our captioning results. For Dubai-CC dataset, the annotated change captions are shorter than LEVIR-CC dataset. Therefore, a wrong change semantics "sand" in Figure 3c will

confuse change caption generation. It indicates that difference encoding stage is not good at distinguishing features with high similarity, which is significant challenge in RSICC task. All in all, the change semantics can lead to better comprehension and high quality change sentence generation.

5. Discussion

5.1. Ablation Experiments

To verify the effectiveness of the component design in our approach, we conducted extensive ablation experiments on LEVIR-CC dataset, the three components consist of the context-aware difference learning (D-S-Diff), difference comprehending (DC) and Transformer decoder with cross refined attention (DE). Our Baseline is analogous to [17], which is formed via Siamese ResNet-101s for encoding multi-scale bi-temporal features, multi-scale change-aware branches and one-layer Transformer decoder for change captioning.

The Baseline does not contain any modules. Specifically, D-S-Diff(D)+DC sequentially stack the deep branch in the context-aware difference learning with difference comprehending, while D-S-Diff(D)+DC represents the combination of the shallow branch difference learning with difference comprehending. Note that the standard Transformer is used without DE symbol. The ablation experimental results are given in Table 4. Comparison of the experimental results for Baseline and D-S-Diff shows that adding context-aware difference learning module in the difference encoding stage to introduce the multi-scale difference context features improves the performance. Specifically, the BLEU4 and CIDEr increase by 2.2% and 2.15%, respectively, compared with Baseline. These results show that D-S-Diff can effectively change captioning through reliable difference distillation. The results of D-S-Diff+DE show that adding the cross refined attention in Transformer decoder can select the critical multi-scale representation for better change captioning. To validate the contribution of deep and shallow branch in the context-aware difference learning, we have constructed ablation experiments for D-S-Diff(D)+DC and D-S-Diff(S)+DC. The performance for D-S-Diff(S)+DC is more significant, which directly verifies that the combination of shallow difference features and the difference comprehending is critical to our model's effectiveness. This is because the shallow features contain fine-grained spatial information, which is also beneficial for semantic feature awareness. Furthermore, the D-S-Diff(S)+DC outperforms D-S-Diff in terms of all considered metrics, especially improves the CIDEr score from 1.4015 to 1.4207. The change semantics generated from the difference comprehending can clearly passes useful semantic information to the captioner. This achieves a further RSICC performance boost. Moreover, the caption generator combined with the cross refined attention is an important connection for visual changes, change semantics and text associations. Similarly, the performance of FULL compared with D-S-Diff(S)+DC also improved on all metrics. Importantly, there is clear performance improvement between Baseline and FULL from Table 4, the highest performing compared with baseline is increased by 1.53% on ROUGE_L and 5.34% on CIDEr. Overall, our model is positively affected by each component.

In addition, one GT sentence and sentences generated by baseline (B1) and ablation models (B2–B4) are visualized in Figure 4. Specifically, D-S-Diff (B2) stacked with the difference comprehending is abbreviated as B3, Full model for B4. We can see that the baseline (B1) sometimes fails to generate informative or correct sentences. For example, the sentences generated by B1–B4 in Figure 4d contain different level errors, which are marked red color. We analyze this phenomenon with two reasons, one is that the GT annotation tends to quantity statistics, and the other is due to the similarity of landforms. This indicates that our model has deficiencies in quantity statistics. Besides, the "desert" marked red is generated by baseline in Figure 4c, which also suffers from the problems of information

missing like "a parking lot". By contrast, with the guidance of context-aware difference learning module, B2 model can observe key feature information contained in "before" and "after" RSI. As shown in Figure 5, it can be seen the visualized maps for "before" and "after" RSI, which the weights are from the final MHA in context-aware difference learning module. Upon these examining figures, the appeared changes in "after" RSI can be precisely identified, which can cover the complete change region. Thus, the precise interpretation of complex visual content can be more accurate and fluent compared with B1 model. On the other hand, sentences generated by B2 closely relate to the changed image content and add details to change captions. With the assistance of the difference comprehending process, our model leverages the predicted change semantics to generate a more natural and informative caption. Concretely, B3 model can better understand more change content with the benefit of the change semantics, such as "a parking lot" in Figure 4b. It also shows that change semantic information provides semantic explainability of the RSI pair since it summarizes high-level image content. Further, the change captions generated by full model (B4) are more informative and well-ordered, basically following an extend Transformer structure. Our decoder helps the captioning model to make visual-semantic associations accurately. Above all, the full model not only focuses on comprehensive difference information, but also improves the overall quality of generated change description.



Figure 4. Change captioning results of our ablated models. There is one GT annotated sentence and four predicted change sentences (B1–B4). Green words are more accurate prediction compared with GT, wrong words marked red appear in sentences. (**a**–**d**) are examples from LEVIR-CC dataset, the first column represents "before" image, "after" image is listed in the second column.

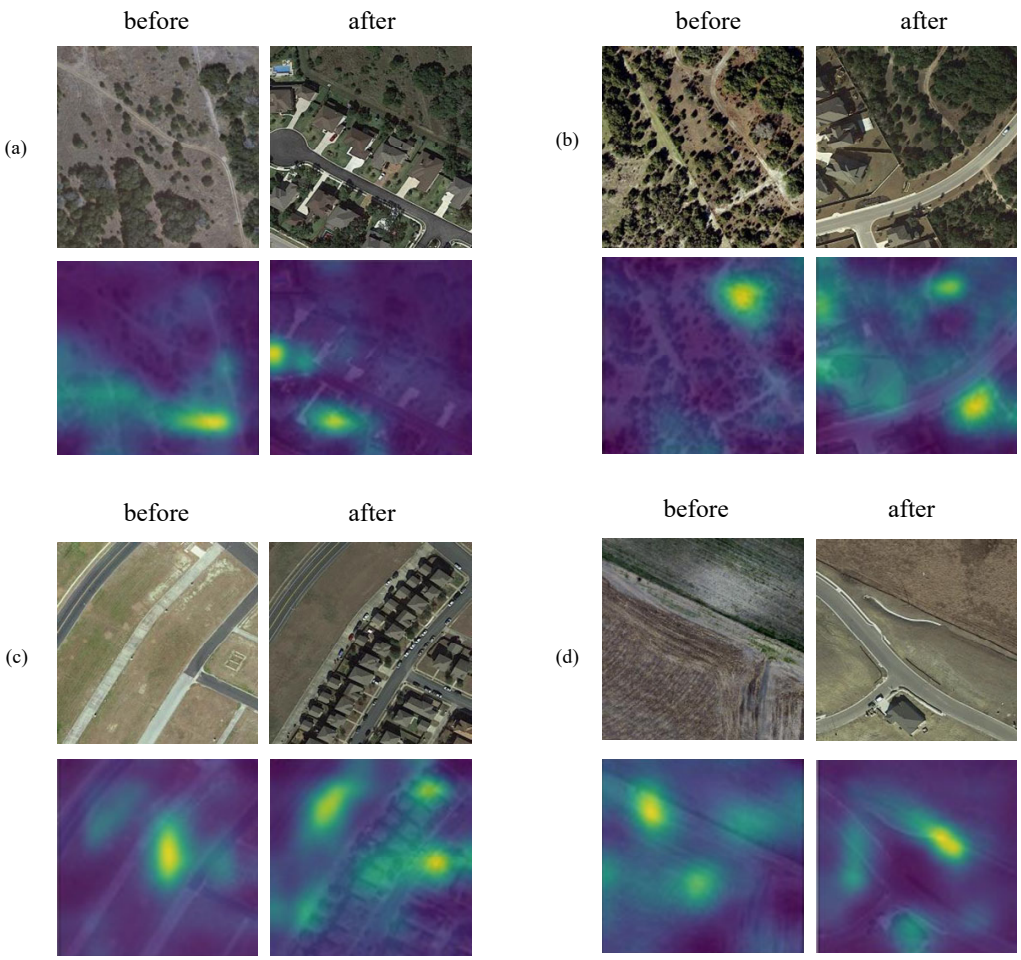


Figure 5. Visualization of the attention maps for the context-aware difference learning module. (**a**–**d**) represent four change scenes, the "before" and "after" RSIs are with the corresponding attention map.

Table 4. Ablation performance of our designed model on LEVIR-CC dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
Baseline	0.8606	0.7812	0.7145	0.6612	0.4051	0.7521	1.3836
D-S-Diff	0.8625	0.7855	0.7212	0.6699	**0.4170**	0.7624	1.4015
D-S-Diff+DE	0.8652	**0.7909**	**0.7286**	**0.6780**	0.4159	0.7612	1.4152
D-S-Diff(D)+DC	0.8643	0.7894	0.7255	0.6735	0.4166	0.7612	1.4087
D-S-Diff(S)+DC	0.8629	0.7824	0.7109	0.6513	0.4083	0.7629	1.4207
FULL(DE)	**0.8649**	0.7881	0.7221	0.6671	0.4161	**0.7681**	**1.4334**

Note: The bold numbers indicate the best results.

5.2. Parameter Analysis

To evaluate the impact of the ratio parameter for $\{\beta_1, \beta_2\}$ in difference encoding module and the top-m with different values in the difference comprehending module, we set different m value and $\{\beta_1, \beta_2\}$. When the value of m changes, $\{\beta_1, \beta_2\}$ and other parameters remain fixed. For evaluating the impact of $\{\beta_1, \beta_2\}$, we also ensure consistency of other parameters. In order to highlight the trend of the results in parametric study, which are multiplied by 100.

For dual ranking losses, trade-off parameters are set to balance the contributions to perform well on multi-scale difference learning. The shallow difference branch typically learns low-level generic features, the deep difference branch tends to have strong generalization capability. It is common to weaken the ratio of the shallow layers to prevent more

attention on low-level features, and to tune lager ratio of the deeper branch to adapt to the downstream tasks. We examined how the performance of the model is affected by fine-tuning varying different ratios on dual ranking losses. Both ratio parameters $\{\beta_1, \beta_2\}$ are discussed on LEVIR-CC dataset, the results shown in Figure 6a,b. In Figure 6, the "top ratio/down ratio" is labeled for specific parameter setting. Notably, 3 in "3/7" represents 0.3, 7 for 0.7. It can be observed that the difference context with "3/7" setting achieves the best performance for the model. By fine-tuning the deep branch ratio about 0.5, the model can learn more shallow difference features and further restrict its performance on the RSICC tasks. Furthermore, it can be observed that when the deep branch ratio is less than 0.5, the performance of the model on RSICC is significantly poor. The results imply that high-level difference features primarily contribute to the performance of the model, it hampers the model's capability to adapt excessive reliance on low-level features.

The parameter top-m in the difference comprehending module plays a crucial role in the performance of our model. With appropriate $\{\beta_1, \beta_2\}$, both deep and shallow difference branch can have a positive impact on the model's ability to accurately generate descriptions. We explored different m value, ranging from 1 to 4. From Figure 7a,b, we observe that the model performs best when $m = 3$, as indicated by the highest values of 0.7681 for ROUGE_L and 1.4334 for CIDEr. From Figure 7a,b, we observe that as the m increases from 1 to 3, the model's performance gradually improves.However, m reaches 4, the performance of the model declines. This may be because a large m value will cause strong semantic interference, causing a negative impact on the model's performance. The small m is less applicable to downstream tasks, where our model has not learned enough semantic change features from the shallow difference features.

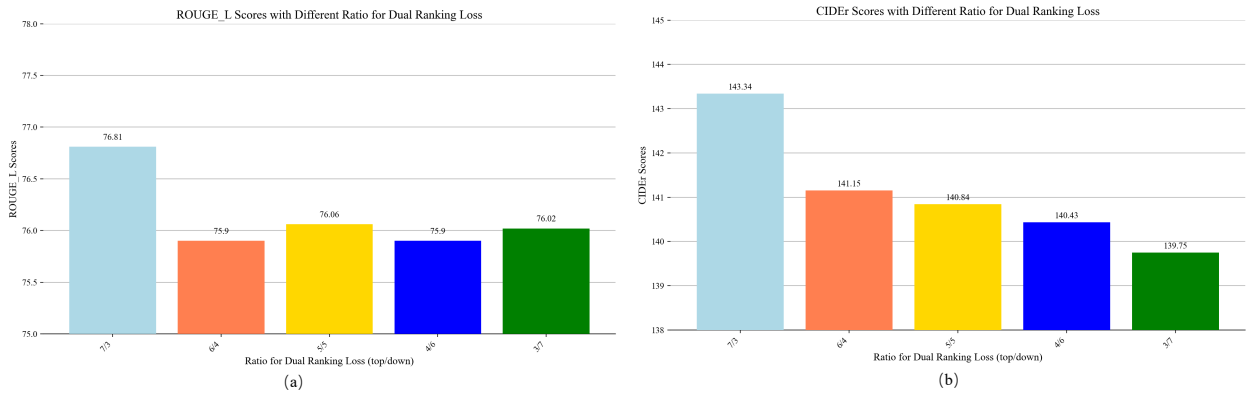


Figure 6. Effects of the ratio parameter for the context-aware difference learning module in terms of ROUGE_L (**a**) and CIDEr (**b**) metric on LEVIR-CC dataset. For example, "3/7" represents $\beta_1 = 0.3$ for the top difference branch, $\beta_1 = 0.7$ for the down difference branch.

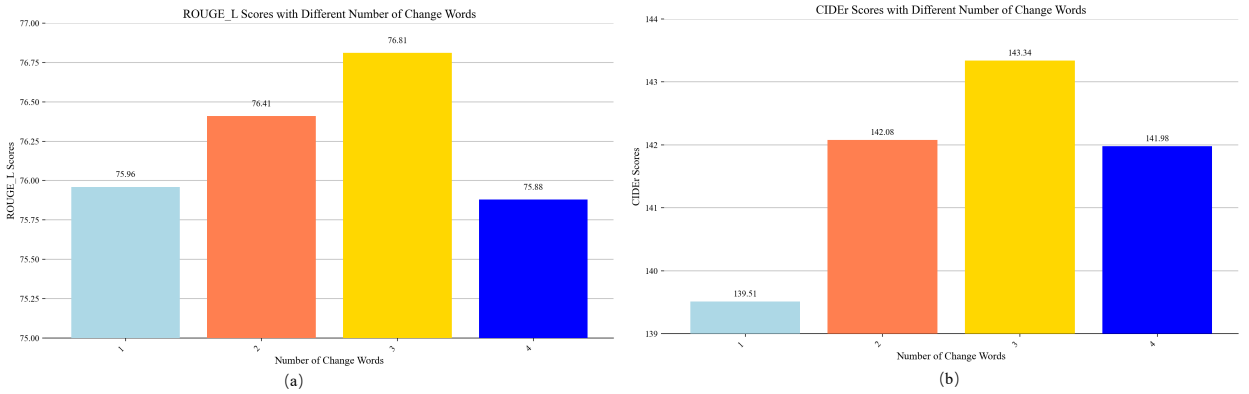


Figure 7. Effects of the top-m parameter for the difference comprehending module in terms of ROUGE_L (**a**) and CIDEr (**b**) metric on LEVIR-CC dataset.

6. Conclusions

This article produces a semantic-guided change captioning designed for RSICC task. It features a symmetric difference context encoding, a shallow difference auxiliary encoding, a difference comprehending layer and change captioning generator. At first, the Siamese encoder, built upon ResNet-101s, provides deep and shallow bi-temporal features, supporting the symmetric difference context branch and shallow difference auxiliary branch. The context-aware difference learning module utilizes the multi-scale different features distilled salient/ weak changes across complex RSI pairs, enhancing the distinction between the changed and unchanged information. Moreover, we believe that some expected improvement can be achieved when high-level semantics are then transmitted to the decoder. The difference comprehending layer can model semantic changes with shallow change information. Such textual cues and deep difference features are leveraged to inform and refine the change captioning process, preserving as much high-quality change content within the difference context feature as possible. Compared to SOTA approaches, our model demonstrates superior performance on popular benchmark datasets (i.e., LEVIR-CC dataset). Particularly, when B2 and B3 are utilized jointly, efficiency analysis with ablation studies further highlight the effectiveness in integrating semantic and change information.

Future work should focus on improving model accessibility with multi-task encoder, exploring knowledge graph with diffusion decoder to capture long-range dependencies and enhance predictive accuracy.

Author Contributions: Conceptualization, Y.L.; funding acquisition, Y.L. and G.W.; methodology, Y.L., X.Z., and G.W.; software, Y.L. and T.Z.; supervision, X.Z., G.W. and T.Z.; writing—original draft, Y.L.; writing—review and editing, X.Z., Y.L., G.W. and T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Wuxi Innovation and Entrepreneurship Fund "Taihu Light" Science and Technology (Fundamental Research) Project under Grant K20241045, the National Natural Science Foundation of China under Grant 62506285, and the Start-up Fund for Introducing Talent of Wuxi University under Grant 2024r011.

Data Availability Statement: The LEVIR-CC and Dubai-CC datasets can be obtained from (https://github.com/Chen-Yang-Liu/RSICC (accessed on 1 November 2022), https://disi.unitn.it/~melgani/datasets.html (accessed on 1 August 2022).).

Acknowledgments: The authors would like to express their gratitude to the editors and the anonymous reviewers for their insightful comments.

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

Abbreviations

The following abbreviations are used in this manuscript:

RSICC	Remote sensing image change captioning.
RSI	Remote sensing image.
CNN	Convolutional neural networks.
LSTM	Long short-term memory.
VGG	Visual geometry group.
ResNet	Residual network.
BLEU	Biingual evaluation understudy.
ROUGE-L	Recall-oriented understudy for gisting evaluation—Longest.
METEOR	Metric for Evaluation of translation with explicit ordering.
CIDEr	Consensus-based image description evaluation.
SOTA	State-of-the-art models.
MHA	Multi-head attention.

TACC	Text-augmented change captioning method.
$\{F_{t0}^4, F_{t1}^4\}$	The shallow bi-temporal features.
$\{F_{t0}^5, F_{t1}^5\}$	The deep bi-temporal features.
$\{T_0^{l+1}, T_1^{l+1}\}$	The top difference context features of l layer.
$\{D_0^{l+1}, D_1^{l+1}\}$	The shallow difference context features of l layer.
Z_T	The fused difference feature from the top difference encoding module.
Z_D	The fused difference feature from the down difference encoding module.
S	The predicted change semantic words.
$\tilde{S}$	The alignment feature between the change features and semantic words.
T	The max length in ground-truth sentence.
y_t	The generated word at t time.

References

1. Zhang, X.; He, L.; Qin, K.; Dang, Q.; Si, H.; Tang, X.; Jiao, L. SMD-Net: Siamese Multi-Scale Difference-Enhancement Network for Change Detection in Remote Sensing. *Remote Sens.* **2022**, *14*, 1580. [CrossRef]
2. Wang, G.; Zhang, X.; Peng, Z.; Tian, S.; Zhang, T.; Tang, X.; Jiao, L. OraL: An Observational Learning Paradigm for Unsupervised Hyperspectral Change Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 5380–5393. [CrossRef]
3. Zhang, X.; Hong, W.; Li, Z.; Cheng, X.; Tang, X.; Zhou, H.; Jiao, L. Hierarchical Knowledge Graph for Multilabel Classification of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
4. Zhang, X.; Chen, Y.; Wang, G.; Zhang, Y.; Jiao, L. EDDA: An Efficient Divide-and-Conquer Domain Adapter for Automatics Modulation Recognition. *IEEE J. Sel. Top. Signal Process.* **2025**, *19*, 140–153. [CrossRef]
5. Zhang, X.; Li, Y.; Wang, X.; Liu, F.; Wu, Z.; Cheng, X.; Jiao, L. Multi-Source Interactive Stair Attention for Remote Sensing Image Captioning. *Remote Sens.* **2023**, *15*, 579. [CrossRef]
6. 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]
7. 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]
8. Wang, Y.; Yu, W.; Ghamisi, P. Change Captioning in Remote Sensing: Evolution to SAT-Cap—A Single-Stage Transformer Approach. *arXiv* **2025**, arXiv:2501.08114.
9. Chang, S.; Ghamisi, P. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Trans. Image Process.* **2023**, *32*, 6047–6060. [CrossRef]
10. Sun, Q.; Wang, Y.; Song, X. Scene Graph and Dependency Grammar Enhanced Remote Sensing Change Caption Network (SGD-RSCCN). In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 2121–2130.
11. 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]
12. Sun, D.; Bao, Y.; Liu, J.; Cao, X. A lightweight sparse focus transformer for remote sensing image change captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 18727–18738. [CrossRef]
13. 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]
14. De Bem, P.P.; de Carvalho Junior, O.A.; Fontes Guimarães, R.; Trancoso Gomes, R.A. Change detection of deforestation in the Brazilian Amazon using landsat data and convolutional neural networks. *Remote Sens.* **2020**, *12*, 901. [CrossRef]
15. 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]
16. 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.
17. Cai, C.; Wang, Y.; Yap, K.H. Interactive change-aware transformer network for remote sensing image change captioning. *Remote Sens.* **2023**, *15*, 5611. [CrossRef]
18. Hang, R.; Luo, J.; Lin, H.; Liu, Q. Text-Augmented Semantic Feature Extraction and Difference Information Learning for Remote Sensing Image Change Captioning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5636112. [CrossRef]
19. Yang, Y.; Liu, T.; Pu, Y.; Liu, L.; Zhao, Q.; Wan, Q. Remote sensing image change captioning using multi-attentive network with diffusion model. *Remote Sens.* **2024**, *16*, 4083. [CrossRef]

20. Li, X.; Sun, B.; Wu, Z.; Li, S.; Guo, H. Cd4c: Change detection for remote sensing image change captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 9181–9194. [CrossRef]
21. Li, Y.; Zhang, X.; Gu, J.; Li, C.; Wang, X.; Tang, X.; Jiao, L. Recurrent attention and semantic gate for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
22. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description generation for remote sensing images using attribute attention mechanism. *Remote Sens.* **2019**, *11*, 612. [CrossRef]
23. Zhang, Z.; Diao, W.; Zhang, W.; Yan, M.; Gao, X.; Sun, X. LAM: Remote sensing image captioning with Label-Attention Mechanism. *Remote Sens.* **2019**, *11*, 2349. [CrossRef]
24. 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 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1971–1980.
25. Li, Y.; Zhang, X.; Cheng, X.; Chen, P.; Jiao, L. Inter-temporal interaction and symmetric difference learning for remote sensing image change captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13.
26. Wu, R.; Ye, H.; Liu, X.; Li, Z.; Sun, C.; Wu, J. A Cross-Spatial Differential Localization Network for Remote Sensing Change Captioning. *Remote Sens.* **2025**, *17*, 2285. [CrossRef]
27. Liu, C.; Chen, K.; Chen, B.; Zhang, H.; Zou, Z.; Shi, Z. Rscama: Remote sensing image change captioning with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
28. Karimli, O.; Mustafazade, I.; Karaca, A.C.; Amasyalı, F. Data augmentation in remote sensing image change captioning. In Proceedings of the 2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkiye, 26–28 October 2024; pp. 287–292.
29. Yang, K.; Wei, J.; Chen, C.; Wang, Z.; Lan, J.; Li, X.; Hua, D.; Xue, D.; Wu, Y. Restricted supervised Cascade Information Network for remote sensing change captioning with serial sentences. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104686. [CrossRef]
30. Wang, Z.; Wang, M.; Xu, S.; Li, Y.; Zhang, B. Ccexpert: Advancing mllm capability in remote sensing change captioning with difference-aware integration and a foundational dataset. *arXiv* **2024**, arXiv:2411.11360. [CrossRef]
31. Yu, X.; Li, Y.; Ma, J.; Li, C.; Wu, H. Diffusion-rscc: Diffusion probabilistic model for change captioning in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5617013. [CrossRef]
32. Bai, Q.; Wang, X. Cross-Temporal Remote Sensing Image Change Captioning: A Manifold Mapping and Bayesian Diffusion Approach for Land Use Monitoring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 14406–14415. [CrossRef]
33. Li, X.; Sun, B.; Li, S. Detection assisted change captioning for remote sensing image. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 10454–10458.
34. 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.
35. Karaca, A.C.; Ozelbas, E.; Berber, S.; Karimli, O.; Yildirim, T.; Amasyali, M.F. Robust change captioning in remote sensing: Second-cc dataset and mmodalcc framework. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 21494–21513. [CrossRef]
36. 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]
37. 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]
38. Yang, C.; Li, Z.; Jiao, H.; Gao, Z.; Zhang, L. Enhancing perception of key changes in remote sensing image change captioning. *IEEE Trans. Image Process.* **2025**, *34*, 7378–7390. [CrossRef] [PubMed]
39. Zhu, Y.; Li, L.; Chen, K.; Liu, C.; Zhou, F.; Shi, Z. Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5648916. [CrossRef]
40. Sun, D.; Yao, J.; Xue, W.; Zhou, C.; Ghamisi, P.; Cao, X. Mask approximation net: A novel diffusion model approach for remote sensing change captioning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5652311. [CrossRef]
41. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2121–2129.
42. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification 2010. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information System, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
43. Banerjee, S.; Lavie, A. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2005; pp. 65–72.
44. Lin, C. *ROUGE: A Package for Automatic Evaluation of Summaries*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004.

45. Vedantam, R.; Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. *arXiv* **2015**, arXiv:1411.5726. [CrossRef]
46. Park, D.H.; Darrell, T.; Rohrbach, A. Robust change captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4624–4633.

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