remote sensing

MDPI

Article

Interactive Change-Aware Transformer Network for Remote Sensing Image Change Captioning

Chen Cai [1], Yi Wang [2] and Kim-Hui Yap [1,*]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore; e190210@e.ntu.edu.com

[2] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong; yi-eie.wang@polyu.edu.hk

* Correspondence: ekhyap@ntu.edu.sg

Abstract: Remote sensing image change captioning (RSICC) aims to automatically generate sentences describing the difference in content in remote sensing bitemporal images. Recent works extract the changes between bitemporal features and employ a hierarchical approach to fuse multiple changes of interest, yielding change captions. However, these methods directly aggregate all features, potentially incorporating non-change-focused information from each encoder layer into the change caption decoder, adversely affecting the performance of change captioning. To address this problem, we proposed an Interactive Change-Aware Transformer Network (ICT-Net). ICT-Net is able to extract and incorporate the most critical changes of interest in each encoder layer to improve change description generation. It initially extracts bitemporal visual features from the CNN backbone and employs an Interactive Change-Aware Encoder (ICE) to capture the crucial difference between these features. Specifically, the ICE captures the most change-aware discriminative information between the paired bitemporal features interactively through difference and content attention encoding. A Multi-Layer Adaptive Fusion (MAF) module is proposed to adaptively aggregate the relevant change-aware features in the ICE layers while minimizing the impact of irrelevant visual features. Moreover, we extend the ICE to extract multi-scale changes and introduce a novel Cross Gated-Attention (CGA) module into the change caption decoder to select essential discriminative multi-scale features to improve the change captioning performance. We evaluate our method on two RSICC datasets (e.g., LEVIR-CC and LEVIRCCD), and the experimental results demonstrate that our method achieves a state-of-the-art performance.

Keywords: image change captioning; remote sensing; multi-layer change awareness; transformer

check for updates

Citation: Cai, C.; Wang, Y.; Yap, K.-H. Interactive Change-Aware Transformer Network for Remote Sensing Image Change Captioning. *Remote Sens.* **2023**, *15*, 5611. https://doi.org/10.3390/rs15235611

Academic Editor: John Trinder

Received: 7 November 2023
Revised: 27 November 2023
Accepted: 30 November 2023
Published: 3 December 2023

1. Introduction

Recently, deep-learning-based sensing image change captioning (RSICC) technologies have demonstrated their effectiveness in observing and analyzing the change in the earth's surface [1,2]. They take advantage of multitemporal images acquired by sensors onboard satellites or aerial platforms for continual observation and tracking of environmental changes. RSICC is an evolving field of research that aims to understand the changes in input bitemporal remote sensing (RS) images and generate descriptive natural language sentences that accurately describe the differences between them. It analyzes and illustrates the differences between bitemporal scenes, significantly deepening our understanding of the dynamic changes in the environment and landscape. RSICC has a broad range of applications, including landscape damage examination, city planning, environmental monitoring, and land planning [2–4].

RSICC involves interpreting change regions between two RS images captured at the same location but at different times (as shown in Figure 1). It requires a deep understanding of the semantic meaning of these changes in the complex environment and a detailed analysis of the evolved scene. Like recent image captioning works [5–7], RSICC adopts an encoder–decoder-based architecture, where a visual encoder extracts discriminative features and captures the difference between bitemporal images, while the language decoder generates descriptive sentences that articulate these differences. Chouaf et al. [1] are pioneers in the RSICC task; they used a CNN as a visual encoder to capture the temporal changes in scenes and adopted an RNN as a decoder to generate descriptions of the changes. Liu et al. [3] adopted a Transformer-based [8] encoder–decoder framework for the RSICC task, which has achieved a great performance.

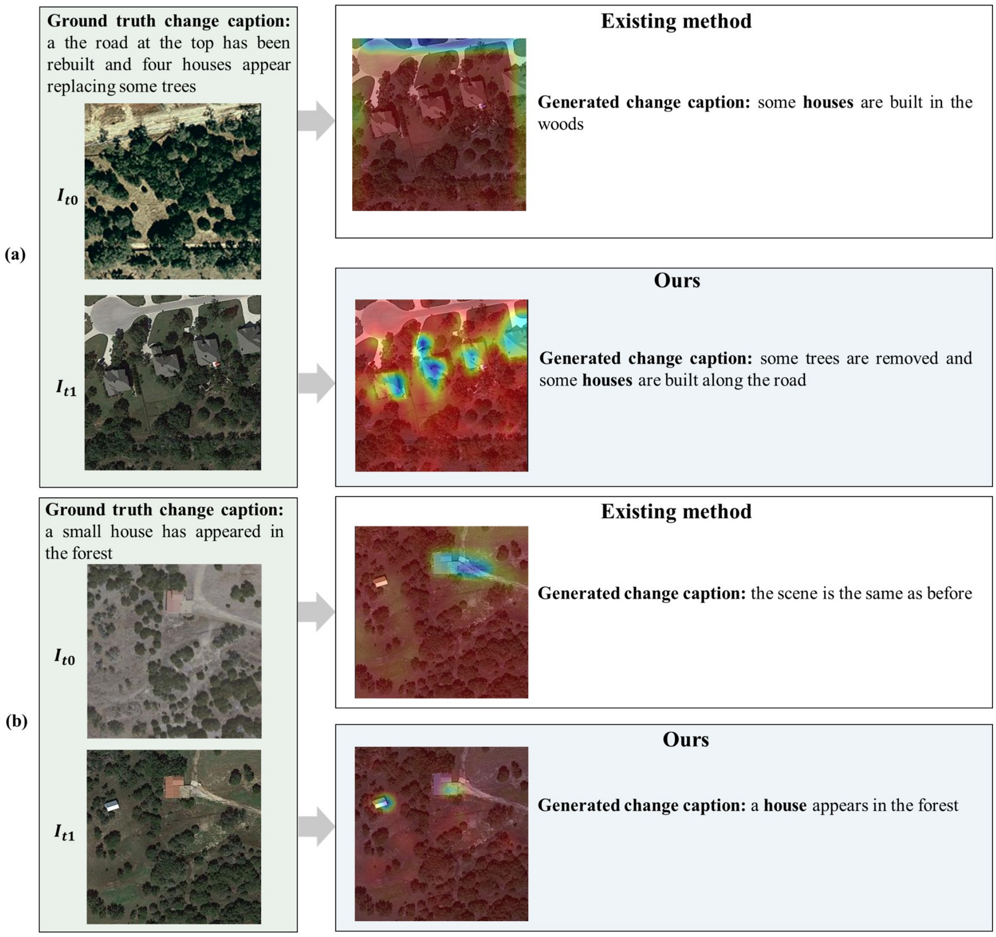

(a)
Ground truth change caption: a the road at the top has been rebuilt and four houses appear replacing some trees
I_{t0}
I_{t1}
Existing method
Generated change caption: some houses are built in the woods
Ours
Generated change caption: some trees are removed and some houses are built along the road
(b)
Ground truth change caption: a small house has appeared in the forest
I_{t0}
I_{t1}
Existing method
Generated change caption: the scene is the same as before
Ours
Generated change caption: a house appears in the forest

Figure 1. A visualization of the existing method and our proposed method. (**a**) The existing method [3] uses a hierarchical approach that tends to integrate the unchanged focused information from each encoder layer, disrupting the change feature learning in the decoder and generating inferior change descriptions. Our proposed method attentively aggregates the essential features for more informative caption generation. (**b**) Existing methods [2–4,9] overlook the change in objects with various scales, generating inferior change descriptions. Ours can extract discriminative information across various scales (e.g., a small scale) for change captioning. Blue indicates that the word "house" is attended to the particular region in the image, while reddish colors suggest a lower level of focus on it. The bluer the color, the higher the attention value.

Recent RSICC methods [3,9] proposed capturing the changes in each encoder layer and gradually concatenating low-level and high-level change-aware semantic features in all the layers to support the change caption decoder in generating more accurate captions. Nevertheless, these approaches are prone to incorporating redundant features from each encoder layer into the caption decoder, thereby adversely affecting the change caption generation process. For instance, in Figure 1a, these methods tend to fuse non-change-focused features and propagate them to the decoder, which causes disruption in word and feature attention in the decoder. Consequently, this interference produces less accurate change descriptions with the absence of the "tree" and "road" compared to the ground truth caption. Moreover, most existing methods [2–4,9] overlook the distinctive characteristics between natural images and remote sensing images, which consequently limits the model's ability to effectively capture the changes in the objects at a different scale (e.g., small-scale objects) and leads to generating inferior sentences to describe the changes, as shown in Figure 1b. Illustrated in Figure 1, it is evident that a significant challenge in remote sensing image change captioning research lies in effectively filtering out noisy feature representations [10,11]. In addition, the diversity of scales in images is a natural characteristic resulting from variations in camera-to-object distances and causes differences in scale among the objects within the image. Hence, it is crucial to be aware of the presence and absence of the objects across different regions with varying scales in bitemporal images and provide comprehensive descriptions of these changes.

In this paper, we proposed an Interactive Change-Aware Transformer Network (ICT-Net) to alleviate the above-mentioned problems. ICT-Net excels in extracting and integrating the most pivotal changes of interest within each encoder layer, thereby enhancing the generation of more effective change descriptions. In the encoder, ICT-Net utilizes an Interactive Change-Aware Encoder (ICE) to capture change information between bitemporal features extracted from the backbone network (e.g., ResNet [12]). Specifically, the ICE leverages the Cross Multihead Attention (Cross-MHA) mechanism [8] in difference and content attention encoding modules to learn the most discriminative representations and recognize the changes of interest between paired features. Moreover, the Multi-Layer Adaptive Fusion (MAF) module is introduced to effectively integrate relevant low- and high-level semantic change-aware features in each ICE layer. MAF utilizes an attention design to filter out irrelevant change information from integrated visual features. In addition, we expand the ICE to extract multi-scale change-aware features, aiming to overcome the challenges of recognizing changes in objects at various scales. In the change caption decoder, we propose a Cross Gated-Attention (CGA) module to generate a change description by considering the relationship of the words and each scale of the features. CGA employs a gated attention structure, enhancing the decoder's capability to utilize crucial features for more precise change caption generation.

To summarise, in the proposed ICT-Net, we utilize an ICE to capture multi-scale discriminative change-aware information between bitemporal features, followed by an MAF module to integrate the most relevant change information in each layer for the change caption decoder. A CGA module is adopted in the decoder to model the relationships between semantic and multi-scale change-aware features to enhance the change captioning performance. A comprehensive set of experiments is conducted on two remote sensing image change caption datasets. The results of these experiments demonstrate that our proposed model achieves superior performance compared to the state-of-the-art approaches across all evaluation metrics. Our contributions are summarized in the following:

1. We propose an Interactive Change-Aware Transformer Network (ICT-Net) to accurately capture and describe changes in objects in remote sensing bitemporal images.
2. We introduce the Interactive Change-Aware Encoder (ICE) equipped with the Multi-Layer Adaptive Fusion (MAF) module. It effectively captures change information from bitemporal features and extracts essential change-aware features from each encoder layer, contributing to improved change caption generation.
3. We present the Cross Gated-Attention (CGA) module, a novel module designed to effectively utilize multi-scale change-aware representations during the sentence-generation process. This module empowers the change caption decoder to explore the relationships between words and multi-scale features, facilitating the discernment of critical representations for better change captioning.

Section 2 provides a summary of previous work in the field of remote sensing image captioning, remote sensing change detection, and neutral change image captioning. In Section 3, we present our proposed ICT-Net in detail. Next, Section 4 presents the experimental results and analysis. Finally, in Section 5, we conclude this work.

2. Related Works

2.1. Remote Sensing Image Change Captioning

The objective of remote sensing image change captioning (RSICC) is to analyze and illustrate the differences between bitemporal scenes using natural language. Chouaf et al. [1] are pioneers in the RSICC task; they used a CNN as a visual encoder to capture the temporal changes between scenes and adopted an RNN as a decoder to generate descriptions of the changes. Hoxha et al. [2] proposed early and late feature fusion strategies to fuse the bitemporal visual features and utilizes an RNN and a multi-class Support Vector Machine (SVM) decoder to generate change captions. More recently, Liu et al. [3] adopted a Transformer-based [8] encoder–decoder framework for the RSICC task, in which they used a dual-branch Transformer encoder to identify the changes between the scenes and proposed a multistage fusion module to fuse multi-layer features for change description generation. Liu et al. [9] further improved the method by utilizing progressive difference perception Transformer layers to capture the high-level and low-level semantic change information. Liu et al. [4] proposed a prompt-based method that uses pre-trained large language models (LLMs) for RSICC tasks, where they used visual features, change classes, and language representation as input prompts to a frozen LLM for change caption generation. Nevertheless, current methods tend to incorporate irrelevant change information into the model, resulting in an inferior performance. Hence, we propose to capture more change-aware discriminative information with the attention structure to enhance the model's ability to illustrate the changes in scenes.

2.2. Remote Sensing Image Captioning

Remote sensing image captioning (RSIC) aims to generate sentences that describe the contents of the given RS image with natural language. Recently, most of the RSIC works [10,13–21] have used deep learning techniques and adopted an encoder–decoder framework for caption generation. The visual encoder utilizes a CNN [12] or a Vision Transformer [22] pre-trained network to extract the visual features from the input image, then injects the features into the RNN-based [23] or Transformer-based [8] decoder to generate the descriptive sentences. Lu et al. [24] explored an encoder–decoder-based method for RSIC that utilizes CNN models to extract the remote sensing image features and uses a recurrent neural network (RNN) to generate the sentence. Li et al. [25] introduced a novel truncation cross entropy (TCE) loss for RSIC, which aims to solve the overfitting issue and facilitates the model to generate more concise RS image descriptions. Sumbul et al. [14] proposed a summarization-driven RSIC method, which implements an adaptive weighting strategy to effectively integrate the summarized ground truth captions into the captioning model to improve performance. RS images may contain objects of different sizes. Some RSIC methods aim to improve the visual representation modeling abilities of the captioning model and aim to describe the objects with various scales in the RS image. Wang et al. [15] proposed a multi-scale multi-interaction method to connect multi-scale image features at different levels, allowing for more efficient visual representation interaction. Ma et al. [26] introduced scene-level feature extraction and target-level feature extraction modules to capture more fine-grained visual representations for RSIC. The aforementioned RSIC methods aim to generate descriptive sentences of an object in a single image. In contrast, RS image change captioning is focused on capturing and describing the differences in bitemporal remote sensing images.

2.3. Remote Sensing Change Detection

The objective of Remote Sensing Image Change Detection (RSICD) [27–32] is to detect the change regions between bitemporal images and generate a pixel-level change map that illustrates the changed areas. Chen et al. [28] introduced a Siamese Transformer-based [8] framework to improve the model's context and identify the change of interest between given bitemporal images. Bao et al. [33] utilized a Convolutional Neural Network (CNN)-based dual structure to extract and detect the difference between multi-scale features of bi-temporal images and employed a Feature Pyramid Network (FPN) [34] fusion module to fuse information over layers to enhance the detection performance. Peng et al. [27] proposed a dense attention architecture for change detection to improve texture and detail extraction of the visual representations. Saha et al. [35] proposed unsupervised learning techniques for RSICD, combining the proposed deep change vector analysis methods with the extracted spatial contextual information to determine changed pixels. Tang et al. [36] further explored the graph convolutional network (GCN) [37] and metric learning algorithm method that captures rich contextual information from the visual representations. In contrast to RSCD tasks that aim to recognize pixel-level changes of interest, RS image change captioning concentrates on detecting and describing the changes of interest between two images at the semantic level.

2.4. Natural Image Captioning

Natural image captioning (NIC) is a fundamental multimodal task at the intersection of computer vision [38–43] and natural language processing [8,23,44,45], which aims to identify objects within images and describe recognized objects with language. Similar to RSIC, most recent NIC methods utilize an encoder–decoder framework. Xu et al. [5] proposed to use a CNN encoder to extract the natural image features and utilize an RNN network as the language decoder to generate natural language words in sequences. Subsequently, spatial [6] and Transformer multi-head attention [8] mechanisms have then been explored with the intention of enhancing the performance of image captioning tasks. Cornia et al. [7] developed a Transformer-based framework incorporating meshed memory to exploit low-level and high-level visual features for caption generating. Besides NIC, several methods [46–49] have been introduced to solve natural scene, 3D scene, and synthetic image change captioning tasks. Qiu et al. [47] proposed understanding and describing the change in 3D scenes from different viewpoints. Tu et al. [48] introduced a method for learning semantic relation-aware difference representations, which effectively localizes semantic changes and captures the semantic relationships across two images. In contrast, in this work, our objective is to describe the change in real RS scenes, which contain many different object categories with multiple scales and complex ground details.

3. Methodology

The ICT-Net utilizes a CNN and a Transformer-based encoder–decoder framework. The overall structure is shown in Figure 2, and is composed of three main elements: (1) A multi-scale feature extractor to extract pairs of RS visual features from different stages of the backbone CNN network; (2) The proposed Interactive Change-Aware Encoder (ICE) with a Multi-Layer Adaptive Fusion (MAF) module to adaptively capture the semantic discrimination information from each pair of multi-scale features; (3) A multi-scale change caption decoder that utilizes a Cross Gated-Attention (CGA) module to select crucial information from all multi-scale change-aware features generated by the MAF module for change captioning.

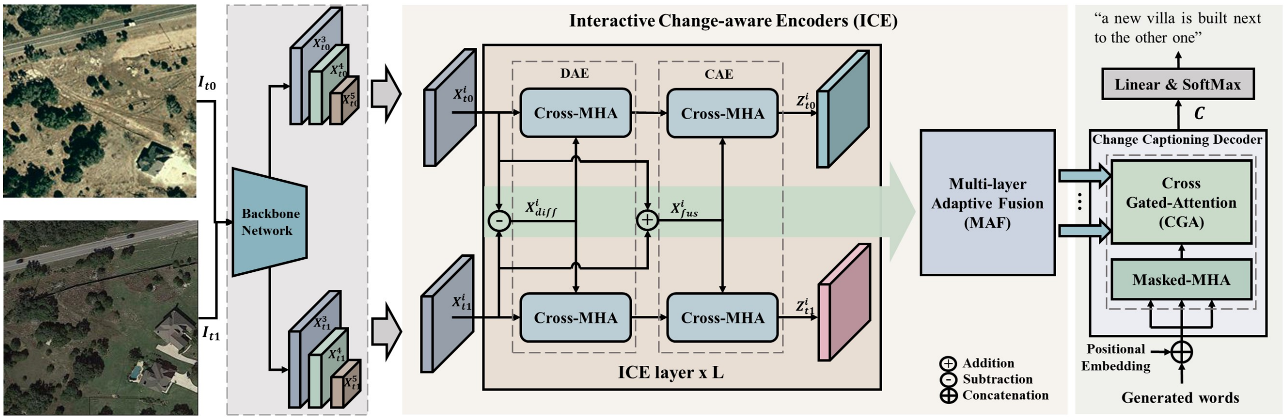


Figure 2. Overview of the proposed ICT-Net. It consists of three components: a multi-scale feature extractor to extract visual features, an Interactive Change-Aware Encoder (ICE) with a Multi-Layer Adaptive Fusion (MAF) module to capture the semantic changes between bitemporal features, and a change caption decoder with a Cross Gated-Attention (CGA) module to generate change descriptions.

3.1. Multi-Scale Feature Extraction

We extract multi-scale features using different convolutional stages in the ResNet [12] backbone to enable the model to capture objects with different scales. As illustrated in Figure 2, given a pair of input images I_{t0} and I_{t1}, the backbone network extracts multi-scale features and uses a transformation function (e.g., 1×1, 3×3 convolutional layers) to transform them to the same dimension, D. We use X^i_{t0} (e.g., $X^5_{t0} \in \mathbb{R}^{H \times W \times D}$, $X^4_{t0} \in \mathbb{R}^{2H \times 2W \times D}$, $X^3_{t0} \in \mathbb{R}^{4H \times 4W \times D}$, where H and W denote the height and width of the feature) and X^i_{t1} to represent the multi-scale feature pairs, where i = {3, 4, 5} denotes the features extracted from the respective stage in the ResNet.

3.2. Interactive Change-Aware Encoder

Obtaining different information that reflects the change regions between bitemporal RS images is essential for RSICC. In this paper, we propose an Interactive Change-Aware Encoder (ICE) that aims to interactively extract highly discriminative features between each pair of input bitemporal features $X^i_{t0} \in \mathbb{R}^{N \times D}$ and $X^i_{t1} \in \mathbb{R}^{N \times D}$ interactively, where $N = W \times H$. As shown in Figure 2, each ICE layer comprises difference attention encoding (DAE) and content attention encoding (CAE) modules. These modules work interactively to capture the changes between bitemporal features by utilizing different features denoted as $\mathbf{X}^i_{diff} \in \mathbb{R}^{N \times D}$, and further enhance the change awareness through the incorporation of aggregated features represented as $\mathbf{X}^i_{fus} \in \mathbb{R}^{N \times D}$. Specifically, DAE first extracts the difference between paired bitemporal features and subsequently models the discriminative representations with these features using the Cross Multihead Attention (Cross-MHA) mechanism. Then, CAE further constructs the output content of DAE through Cross-MHA with aggregated bitemporal features. This process models the long-range dependency of discriminative representations with aggregated features, emphasizing the critical dissimilarities between bitemporal features X^i_{t0} and X^i_{t1}. The DAE process can be represented as follows:

$$\bar{\mathbf{Z}}^i_{tj} = \text{Cross-MHA}(\bar{\mathbf{Q}}, \bar{\mathbf{K}}, \bar{\mathbf{V}}) \in \mathbb{R}^{N \times D}, \tag{1}$$

$$\bar{\mathbf{Q}} = \mathbf{X}^i_{tj}\mathbf{W}^q, \bar{\mathbf{K}} = \mathbf{X}^i_{diff}\mathbf{W}^k, \bar{\mathbf{V}} = \mathbf{X}^i_{diff}\mathbf{W}^v, \tag{2}$$

$$\mathbf{X}^i_{diff} = X^i_{t1} - X^i_{t0}, \tag{3}$$

and the CAE process can be expressed as follows:

$$\mathbf{Z}^i_{tj} = \text{Cross-MHA}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) \in \mathbb{R}^{N \times D}. \tag{4}$$

$$\hat{\mathbf{Q}} = \bar{\mathbf{Z}}^i_{tj}\hat{\mathbf{W}}^q, \hat{\mathbf{K}} = \mathbf{X}^i_{fus}\hat{\mathbf{W}}^k, \hat{\mathbf{V}} = \mathbf{X}^i_{fus}\hat{\mathbf{W}}^v, \tag{5}$$

$$\mathbf{X}^i_{fus} = X^i_{t1} + X^i_{t0}, \tag{6}$$

where $\mathbf{W}^q$, $\mathbf{W}^k$, $\mathbf{W}^v$, $\hat{\mathbf{W}}^q$, $\hat{\mathbf{W}}^k$ and $\hat{\mathbf{W}}^v$ are trainable weight matrices, and j = (0, 1). To ease the representation, we assume that position encoding (PE) is added to with bitemporal features. A Feed-Forward Network (FFN) and Layer Normalization (LN) are included in Cross-MHA, similar to the Transformer block.

Furthermore, we introduce a Multi-Layer Adaptive Fusion (MAF) module to adaptively fuse the change-aware multi-level representations obtained from each layer within the preceding ICE. Each ICE layer can encompass distinct meaningful change representations. By leveraging the MAF module, our model can acquire these distinct features from all ICE layers, allowing it to concentrate on the relevant change representations while filtering out irrelevant changes. As illustrated in Figure 3, we first concatenate all the bitemporal change-aware representations from each ICE layer in the channel dimension. Subsequently, we incorporate a gated attention mechanism that allows the model to filter the irrelevant information and determine the essential change-aware representations from concatenated features. The process of MAF can be formulated as follows:

$$\mathbf{Z}^i = (\bar{\mathbf{Z}}^i\mathbf{W}_a) \odot \sigma(\bar{\mathbf{Z}}^i\mathbf{W}_b) + \bar{\mathbf{Z}}^i_{t1}\mathbf{W}_c + \bar{\mathbf{Z}}^i_{t0}\mathbf{W}_d \in \mathbb{R}^{N\times 2D}. \tag{7}$$

$$\bar{\mathbf{Z}}^i = [\bar{\mathbf{Z}}^i_{t0}; \bar{\mathbf{Z}}^i_{t1}], \tag{8}$$

$$\bar{\mathbf{Z}}^i_{t1} = [\mathbf{Z}^i_{t1,0}; \mathbf{Z}^i_{t1,l-1}; \ldots; \mathbf{Z}^i_{t1,l}], \tag{9}$$

$$\bar{\mathbf{Z}}^i_{t0} = [\mathbf{Z}^i_{t0,0}; \mathbf{Z}^i_{t0,l-1}; \ldots; \mathbf{Z}^i_{t0,l}], \tag{10}$$

where [;] denotes concatenation, $\mathbf{W}_a, \mathbf{W}_b, \mathbf{W}_c$ and $\mathbf{W}_d$ are the learnable weights, and σ and $\odot$ represent the sigmoid activation and element-wise multiplication, respectively. Sigmoid activation and element-wise multiplication serve as a gate to bypass the redundant information from multiple ICE layers. l represents the number of layers in ICE. Subsequently, we can obtain the filtered change-aware features $\mathbf{Z}^i$ (i = 3, 4, 5 with respect to the scale of features) through the MAF module, where $\mathbf{Z}^i$ are down-sampled to a consistent spatial size $N = H \times W$. These features are injected into the decoder for caption prediction.

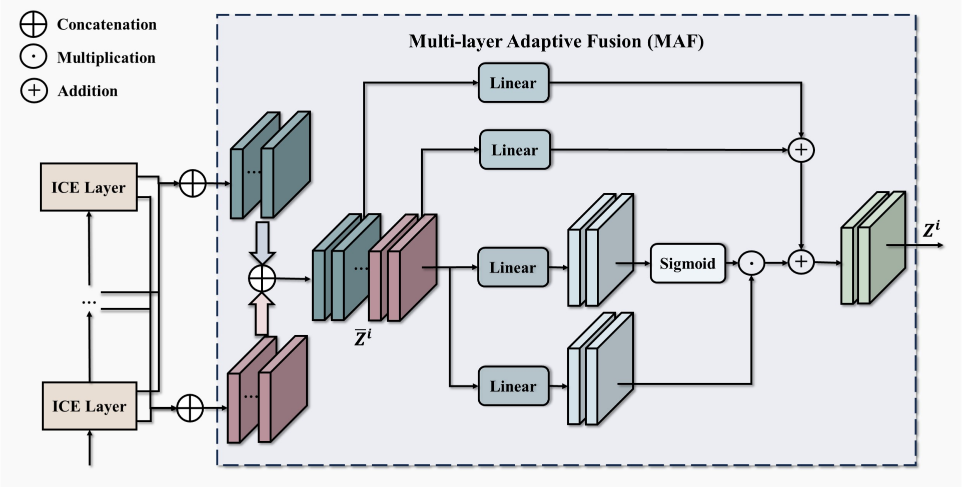


Figure 3. Structure of the Multi-Layer Adaptive Fusion module.

3.3. Multi-Scale Change Caption Decoder

We leverage the previously generated multi-scale change-aware representations modeled from the MAF modules while constructing a multi-layered decoder architecture for change caption generation. To achieve this, we introduce a novel Cross Gated-Attention (CGA) module, which is in contrast to the cross-attention operator used in the original Transformer decoder network [8]. The CGA module allows us to effectively utilize all the multi-scale change-aware representations during the sentence-generation process. Furthermore, it allows the change decoder to attend to and select essential change-aware multi-scale representations for change caption generation with the help of the gated structure. The proposed change caption decoder is composed of three sub-modules: Masked-Multihead Attention (Mask-MHA), Cross Gated-Attention (CGA), and the Feed-Forward Network (FFN), as illustrated in Figure 4. The residual connection and Layer Normalization (LN) operation are adopted for each sub-module.

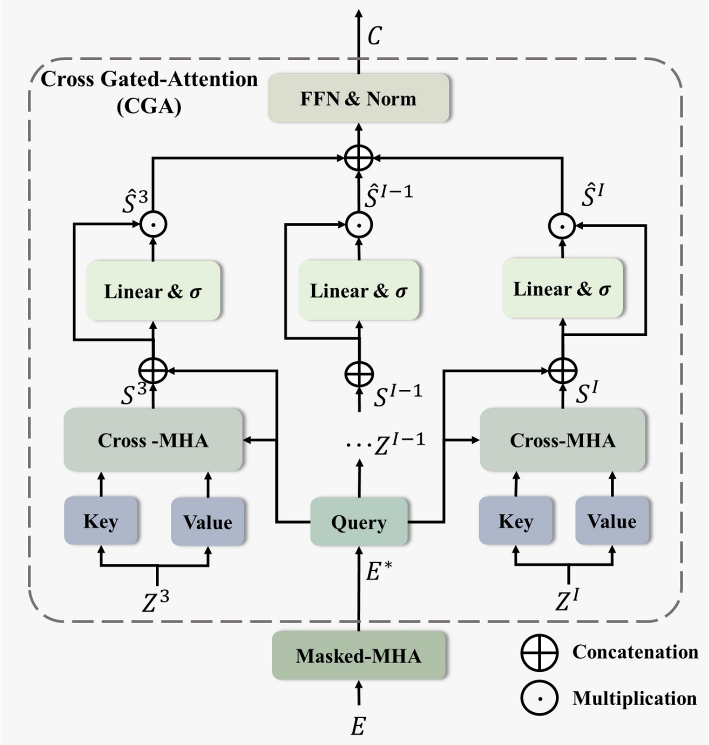


Figure 4. Structure of the Cross Gated-Attention module.

At the training stage, given a sequence of word embeddings $\mathbf{E} = \{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_L\} \in \mathbb{R}^{L\times d}$ as inputs, Mask-MHA masks the subsequent position embeddings at time step t and learns to predict the the word features $\mathbf{e}_t^*$, where L denotes the length of the sentence and d is the word embedding dimension. The process can be written as follows:

$$\mathbf{E}^* = [head_1; head_2; \ldots; head_h]\mathbf{W}_o \in \mathbb{R}^{L\times d}, \tag{11}$$

where

$$head_i = \text{Masked-Attention}(\mathbf{E}\mathbf{W}_i^Q, \mathbf{E}\mathbf{W}_i^K, \mathbf{E}\mathbf{W}_i^V), \tag{12}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ are the learnable projection matrices for query, key, and value of the word embedding at the i-th head and $\mathbf{W}_o$ is the projection matrix that aggregates the information for h number of heads. [;] represents the concatenation operation.

Subsequently, the CGA module is introduced to connect the generated sequence of word features $\mathbf{E}^*$ with all multi-scale change-aware representations $\mathbf{Z}^i$. Hence, instead of focusing on one single scale of the change-aware features, we compute the long-range dependencies across all multi-scale features. The process of computing sentence representations can be written as follows:

$$\mathbf{S}^i = [head_1; head_2; \ldots; head_h]\mathbf{W}_o \in \mathbb{R}^{L\times d}, \tag{13}$$

where

$$head_i = \text{Cross-Attention}(\mathbf{E}^*\mathbf{W}_i^Q, \mathbf{Z}^i\mathbf{W}_i^K, \mathbf{Z}^i\mathbf{W}_i^V), \tag{14}$$

Then, gated attention is introduced to focus on relevant changes of interest in the multi-scale-dependent sentence features $\mathbf{S}^i$ for change caption generation, and it can be computed as:

$$\hat{\mathbf{S}}^i = \mathbf{g}_s \odot \mathbf{S}^i, \tag{15}$$

$$\mathbf{g}_s^i = \sigma(\mathbf{W}_s[\mathbf{E}^*; \mathbf{S}^i] + b_s), \tag{16}$$

Finally, these multi-scale contributed sentence features $\hat{\mathbf{S}}^i \in \mathbb{R}^{L \times d}$ are then summed together:

$$\mathbf{C} = \text{LN}(\mathbf{W}_c[\hat{\mathbf{S}}^3; \ldots, \hat{\mathbf{S}}^I; \mathbf{E}^*] + b_c). \tag{17}$$

where $\mathbf{W}_s$ and $\mathbf{W}_c$ denote learnable projection matrices, and b_s and b_c represent a learnable bias vector. σ and $\odot$ denote sigmoid activation and element-wise multiplication that are used to select and balance the weights learned from each multi-scale feature-dependent word representation $\mathbf{S}^i$, respectively. The output of the caption decoder $\mathbf{C} \in \mathbb{R}^{L \times d}$ is then fed into a linear projection layer and a softmax layer for the prediction of caption word probabilities in the vocabulary:

$$P = \text{Softmax}(\mathbf{C}\mathbf{W}_p + b_p) \tag{18}$$

where L is the length of the sentence, d is the embedding dimension, $\mathbf{W}_p \in \mathbb{R}^{d \times \Sigma}$ are the weight parameters to be learned and Σ denotes the vocabulary size.

The procedure of our proposed model is shown in Algorithm 1 as follows:

Algorithm 1: ICTNet

```
Input: $I \leftarrow (I_{t0}, I_{t1})$
Output: change caption
Step1: Feature extraction
for i in $(t0, t1)$ do
    $\mathbf{X}_{tj}^i \leftarrow$ Backbone($I$)
end
Step2: Interactive Change-Aware Encoder (ICE)
for $l$ in $(1 \backsim L)$ do
    $\bar{\mathbf{Z}}_{tj}^i \leftarrow$ DAE($\mathbf{X}_{tj}^i$)
    $\mathbf{Z}_{tj}^i \leftarrow$ CAE($\mathbf{X}_{tj}^i$, $\bar{\mathbf{Z}}_{tj}^i$)
    $\mathbf{Z}^i \leftarrow$ MAF($\mathbf{Z}_{tj}^i$)
end
Step3: Multi-scale change caption decoder
for $l$ in $(1 \backsim L)$ do
    $\mathbf{E}^* \leftarrow$ Masked-Attention($\mathbf{E}$)
    $\hat{\mathbf{S}}^i \leftarrow$ CGA($\mathbf{E}^*$, $\mathbf{Z}^i$)
    $\mathbf{C} \leftarrow$ LN(Linear($\hat{\mathbf{S}}^i$; $\mathbf{E}^*$))
end
Step 4: Predict change caption
P $\leftarrow$ Softmax(Linear($\mathbf{C}$))
Use probability P to predict caption words y in vocabulary
```

3.4. Training Objective

During the training stage, similar to the existing RSICC [2,3] model, we adopt the widely used cross-entropy (CE) loss to optimize the change caption model, which can be written as follows:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{L} \log(p_\theta(\mathbf{y}_t^* | \mathbf{y}_{1:t-1}^*, I_{t0}, I_{t1})). \tag{19}$$

The model is trained to predict the target ground truth caption $\mathbf{y}_t^*$ with the previous words $\mathbf{y}_{1:t-1}^*$, and the given images I_{t0} and I_{t1}.

4. Experiments

4.1. Dataset and Evaluation Metrics

LEVIR-CC dataset. We conduct experiments on the recently published large-scale LEVIR-CC dataset [3]. The LEVIR-CC dataset contains 10,077 bitemporal remote sensing image pairs, where 5038 image pairs have changed regions and another 5039 image pairs are without changes. The dataset contains 50,385 associate ground truth sentences describing changes between image pairs, whereas 25,190 sentences describe image pairs with changes and the remaining 25,195 sentences express image pairs without changes. The size of the images is 256 × 256 pixels. The dataset has been split into 6815, 1333, and 1929 image pairs for training, validation, and testing, respectively.

LEVIRCCD dataset. We further verify the performance of the proposed method on the LEVIRCCD dataset [2]. It consists of 500 bitemporal images that were originally used for building change detection (CD). The images are cropped into 256 × 256 pixel size. Each image has been annotated with five remote sensing change descriptions, resulting in 2500 change descriptions in total. A split of of 60%, 10%, and 30% of the image and change caption pairs is used for training, validation, and testing, respectively.

Evaluation Metrics. Evaluation metrics measure the accuracy of the generated description with respect to the annotated reference description. Similar to existing works [3,9], we automatically evaluate the change caption performance with four different widely used evaluation metrics, which include BLEU-N (B-N, N = 1, 2, 3, 4) [50], ROUGE-L (R) [51], METEOR (M) [52], and CIDEr-D (C) [53].

The BLEU evaluation metric is used to evaluate the precision accuracy between the candidate and reference sentences, where N represents the n-gram precision between sentences.

METEOR evaluates the uni-gram precision and recall probabilities, and ROUGE-L measures the similarity, calculating the longest common subsequence between two sentences. METEOR and ROUGE-L account for sentence fluency by involving a penalty factor.

CIDEr-D calculates the cosine similarity of the Term Frequency Inverse Document Frequency (TF-IDF). It takes into account both precision and recall, and it reports the real values that exceed 100% [53].

For all these metrics, the higher the metric scores, the higher the accuracy of the generated change description.

4.2. Experimental Setup

We utilized pre-trained ResNet101 [12] as the backbone network for bitemporal remote sensing image feature extraction. The initial learning rate was set to 0.0001 and decays by a weight of 0.7 as the training steps increase by three epochs. The maximum training epoch was set to 40, and the training was discontinued when there was no improvement in the BLEU-4 score for five consecutive epochs. We utilized two Transformer encoder layers and one decoder layer with eight attention heads to achieve the best change caption performance. The model was optimized through the Adam optimizer [54]. Like existing works [3,4], the beam search size was set to 3 for inference. The model was implemented in the PyTorch framework.

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

In Table 1, we compare the remote sensing image change caption performance with state-of-the-art methods on the LEVIR-CC dataset, which include Capt-Dual-Att [55], DUDA [55], MCCFormers$_s$ [49], MCCFormers$_d$ [49], RSICCFormer [3], PSNet [9] and PromptNet [4]. Capt-Dual-Att [55] combines two convolutional layers with spatial attention to attend to important bitemporal visual features. DUDA [55] introduces a dynamic speaker, allowing the model to adaptively attend to visual representations. MCCFormers$_s$ [49] flattens and concatenates the bitemporal feature maps, then injects the fused features into a Transformer network for captioning. MCCFormers$_d$ [49] introduce a Siamese Transformer encoder design to model the relationships between bitemporal visual features and capture the changes. Most of the methods compared utilize the same ResNet-101 backbone, except for PSNet and PromptNet, which use a VIT [22] and CLIP [38] backbone, respectively. B-N helps assess the presence of n-gram words in a sequence. The widely used CIDEr score evaluates the generation of global semantic words in the caption. We can observe that the proposed model presented a superior performance in all of the metrics. These performance improvements shown in the table have proven the effectiveness of our proposed method. We further validate our change caption performance on the LEVIRCCD dataset in Table 2. We compared a method that uses the same backbone network (ResNet50) and has the same settings as our method in Table 2 for a fair comparison. In addition, we selected methods that achieve state-of-the-art performance on the Levir-CC dataset for comparison. We can see that our proposed method achieves a better performance compared with other methods, which further demonstrates the effectiveness of the proposed method. The results can be attributed to the fact that the proposed method has the ability to recognize multi-scale object changes and is able to adaptively fuse multi-layer semantic information for better change caption decoding.

Table 1. Comparison of our proposed method and other state-of-the-art image change caption methods on the Levir-CC dataset. The higher the score, the better the captioning performance. Bold numbers indicate the best result.

Method	B-1	B-2	B-3	B-4	M	R	C
Capt-Dual-Att [55]	79.51	70.57	63.23	57.46	36.56	70.69	124.42
DUDA [55]	81.44	72.22	64.24	57.79	37.15	71.04	124.32
MCCFormer$_s$ [49]	79.90	70.26	62.68	56.68	36.17	69.46	120.39
MCCFormer$_d$ [49]	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 [3]	84.72	76.12	68.87	62.77	39.61	74.12	134.12
PromNet [4]	83.66	75.73	69.10	63.54	38.82	73.72	136.44
Ours	**86.06**	**78.12**	**71.45**	**66.12**	**40.51**	**75.21**	**138.36**

Table 2. Comparisons on the LevirCCD dataset. Our model achieved higher scores, where the metrics in bold have the best performance. Bold numbers indicate the best result.

Method	B-1	B-2	B-3	B-4	M	R	C
CNN-RNN [1]	71.85	60.40	52.18	45.94	27.43	54.13	71.64
MCCFormer$_d$ [49]	66.81	56.89	48.57	41.53	26.16	54.63	78.58
RSICCFormer [3]	69.02	59.78	52.42	46.39	28.18	56.81	80.08
Ours	**72.40**	**62.62**	**55.03**	**48.92**	**30.22**	**58.52**	**85.93**

4.4. Ablation Studies

In this section, we present the numerical results of our ablation studies that validate the effectiveness of the following proposed modules: the Interactive Change-Aware Encoder (ICE), Multi-Layer Adaptive Fusion (MAF), and Cross Gated-Attention (CGA). The ablation models are based on the ResNet101 backbone and were evaluated on the LEVIR-CC dataset.

Table 3 demonstrates the effectiveness of including different components in the proposed method. Difference attention encoding (DAE) and content attention encoding (CAE) are two sub-modules in the ICE module. A tick in the table denotes that the module is included in the model. We observed that the model demonstrates superior performance through the integration of DAE or CAE in the change-aware encoder, surpassing the baseline model utilizing the original Transformer encoder [8]. The proposed method can achieve better results when utilizing both DAE and CAE in the model. Furthermore, the performance is further enhanced by adopting the MAF module that adaptively fuses the change-aware multi-level semantic information obtained from each layer of the ICE. Moreover, the result is further improved with the inclusion of the CGA module that enables the decoder to select the critical multi-scale representation for better change caption generation.

Table 3. Performance of the model with various settings in the Levir-CC dataset. A tick means the module was included for training, whereas a cross denotes the module was not included. Bold numbers indicate the best result.

DAE	CAE	MAF	CGA	B-1	B-2	B-3	B-4	M	R	C
✗	✗	✗	✗	76.43	66.36	58.00	49.50	33.34	69.53	124.74
✓	✗	✗	✗	78.93	68.61	58.88	49.84	34.15	71.81	129.15
✗	✓	✗	✗	79.84	70.40	61.13	53.08	34.43	71.45	128.35
✓	✓	✗	✗	80.74	72.57	65.91	56.84	36.16	72.76	132.67
✓	✓	✓	✗	84.43	76.88	70.46	65.36	39.81	74.69	135.25
✓	✓	✓	✓	**86.06**	**78.12**	**71.45**	**66.12**	**40.51**	**75.21**	**138.36**

In Table 4, we evaluated the model's abilities to determine whether changes exist between bitemporal remote sensing images and whether it was able to describe them with a caption. Hence, we tested the performance with different settings by (1) testing image pairs with no changes, (2) testing image pairs with changes, and (3) testing the overall test set. We can see that the model with an ICE performs better in all three settings as compared to the Transformer network baseline, demonstrating that the ICE effectively captures the change-aware features in bi-temporal remote sensing images. The proposed model with the MAF module achieved a higher evaluation performance compared to only utilizing ICE. This shows the effectiveness of the MAF module in interpreting and filtering the semantic information extracted from different encoder layers to capture multiple changes of interest for better caption generation. Furthermore, the overall model, which incorporates a CGA module, can significantly improve the model performance in all settings. It is designed to exploit word and multi-scale feature relationships and facilitate the selection of essential features to benefit change captioning. Table 4 showcases the significant enhancement brought by our method in terms of both change discrimination and sentence generation performance.

Table 4. Ablation studies on the ICE, MBF, and CGA modules on the test sets with only no changes and only changes and the entire test set. A tick means the module was included for training, whereas a cross denotes the module was not included. Bold numbers indicate the best result.

Test Range	ICE	MAF	CGA	B-1	B-2	B-3	B-4	M	R	C
Test Set (only no-change)	✓	✗	✗	95.15	94.38	94.14	93.48	73.91	95.67	-
	✓	✓	✗	95.80	95.36	95.13	94.97	74.62	96.24	-
	✓	✓	✓	**97.43**	**97.03**	**96.80**	**96.97**	**75.96**	**97.24**	-
Test Set (only change)	✓	✗	✗	71.61	57.40	45.44	36.19	23.21	48.17	55.50
	✓	✓	✗	74.42	60.57	48.36	38.72	25.02	53.12	60.30
	✓	✓	✓	**76.50**	**62.21**	**49.95**	**40.37**	**25.78**	**52.85**	**89.82**
Test Set (entire set)	✓	✗	✗	80.74	72.57	65.91	56.84	36.16	72.76	132.67
	✓	✓	✗	84.43	76.88	70.46	65.36	39.81	74.69	135.25
	✓	✓	✓	**86.06**	**78.12**	**71.45**	**66.12**	**40.51**	**75.21**	**138.36**

Furthermore, this paper investigates the effectiveness of capturing and describing multi-scale object changes between bitemporal remote sensing images. Hence, it is essential to experiment utilizing different stage features (e.g., Stage-3, Stage-4, Stage-5) from the backbone ResNet to localize multi-scale object changes in images. In Table 5, we show the performance of the proposed model after adopting different scales of features for capturing change-aware features. It was observed that the model achieved the best performance when using the Stage-3 and Stage-4 multi-scale features to localize the differences in the two images and describe them with captions. This observation also implies that bitemporal remote sensing images in the dataset tend to contain small- to medium-scale objects, while our proposed model is able to extract and make use of the captured multi-scale change features to improve caption generation. Subsequently, in the experiment, we mainly showcase the outcomes of employing Stage-3 and Stage-4 features as inputs to the model.

Table 5. Performance of the model when utilizing different CNN stages on the Levir-CC dataset. A tick means the module was included for training, whereas a cross denotes the module was not included. Bold numbers indicate the best result.

Conv5	Conv4	Conv3	B-1	B-2	B-3	B-4	M	R	C
✓	✗	✗	73.33	65.24	60.18	56.90	33.47	65.94	112.95
✗	✓	✗	85.02	77.16	70.89	65.61	39.32	74.68	134.91
✗	✗	✓	81.57	75.41	69.24	64.89	37.91	73.09	127.20
✓	✓	✗	84.29	76.08	69.58	64.59	39.96	74.30	132.35
✓	✓	✓	84.36	76.09	69.41	64.29	39.53	73.74	133.53
✗	✓	✓	**86.06**	**78.12**	**71.45**	**66.12**	**40.51**	**75.21**	**138.36**

4.4.1. Interactive Change-Aware Encoder

Tables 3 and 4 provide evidence of the effectiveness of the ICE modules, illustrating their ability to enhance the model's performance. In addition, it is worth paying attention to the change regions located by the ICE between the two images (images taken "before" and "after"). In Figure 5, we visualize and compare the change attention obtained using the DAE module only and the DAE + CAE modules in the ICE. We captured the output attention maps at the last layer of the ICE with different scale input features (Stage-4 and Stage-3), where M_{large} and M_{small} denote attention maps for large and small changes captured between RS image features, respectively. We compare the attention maps generated only using the DAE module with the combination of the DAE and CAE modules to observe and test the effectiveness of these two modules. In role (a), given the two images with only small changes (a small house), we can see that the small-scale object change attention map (M_{small}) generated using DAE + CAE is able to attend to the small house more accurately compared the model only using the DAE module. Similarly, in (b), M_{small} using DAE + CAE is able to focus on the changes in both small houses. We visualize a somewhat large change in (c). M_{large} with DAE + CAE more accurately attends to the change in the large buildings. In (d), we capture the changes in both the large buildings and the narrow load, and we can see that M_{large} with DAE + CAE highly attends to the group of buildings, and M_{small} focuses more on the changes in the narrow load. With these visualizations, we can conclude that DAE and CAE enhance the discriminative feature learning ability of the ICE.

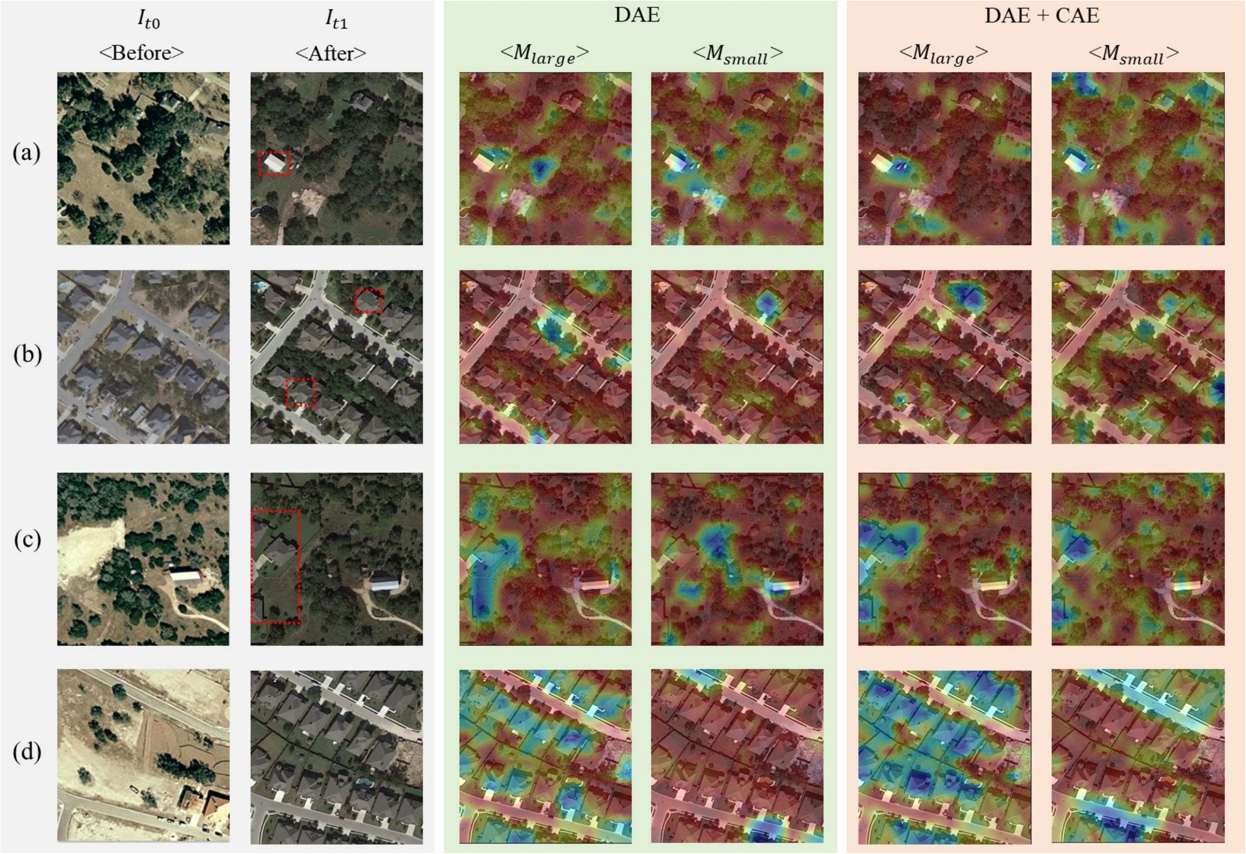


Figure 5. Comparison of attention maps generated using DAE and DAE + CAE. M_{larger} and M_{small} denote the attention maps for large and small changes captured between bitemporal image features, respectively. I_{t_0} and I_{t_1} denote input RS images. Note that regions appearing more blue indicate higher levels of attention. We use the red dotted box to ground the small change areas to ease the visualization.

4.4.2. Multi-Layer Adaptive Fusion module

As shown in Tables 3 and 4, utilizing the MAF module to integrate multi-level semantic feature representation from each layer of the ICE would allow the change caption decoder to explore the relationship between words and each change of interest, which improves the change caption performance. In Figure 6, we visualize the decoder attention between words and integrated change-aware features from the MAF module. The top row are the input bitemporal remote sensing images. The middle image is the word feature attention map computed obtained using MBF [3], and the bottom image is the attention map computed with the proposed MAF. Both MBF and MAF modules are designed to integrate multi-level change aware semantic feature representation. However, MBF lacks a gating design, which may lead to introductions of irrelevant features into the decoder and result in an inferior change description. The MBF module utilizes a gated attention mechanism to select the essential change-aware representations from multi-layer semantic information. For instance, in (a), we can observe that the attention map of the word “house” computed with the MBF module focuses more on the other places instead of the “house” in the image, whereas the attention map captured using the proposed MAF module accurately attends to the “house”. Furthermore, in (b), the model with the MAF module tends to focus on the “road” in the image and is able to generate a more accurate change caption with respect to the ground truth (GT) caption. This visualization demonstrates the effectiveness of incorporating the MAF module, which is beneficial in word visual relationship modeling and allows the model to generate better change captions.

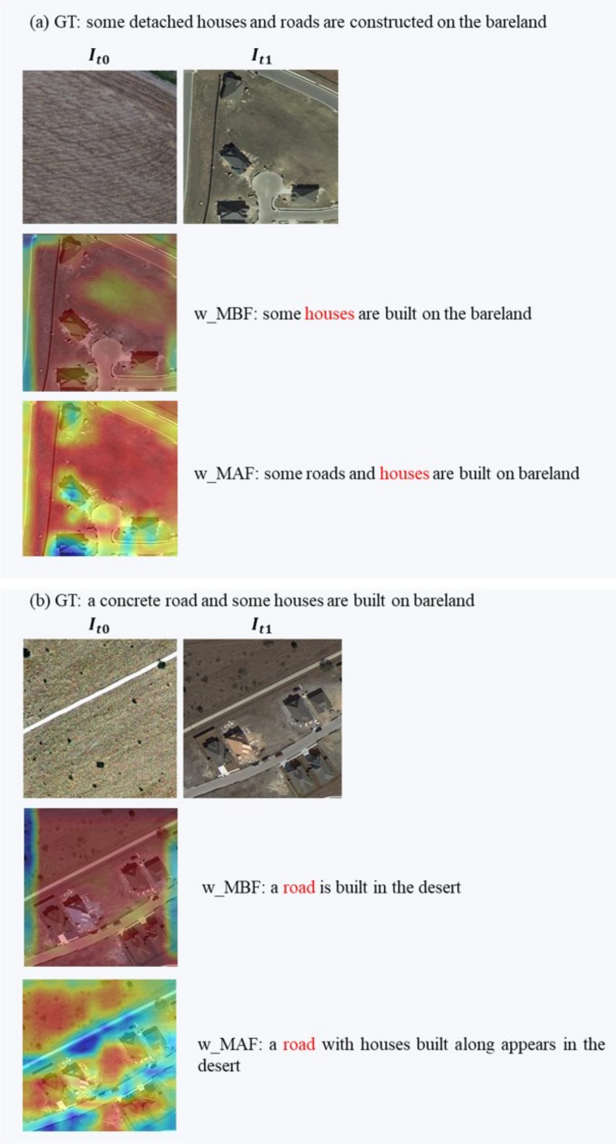
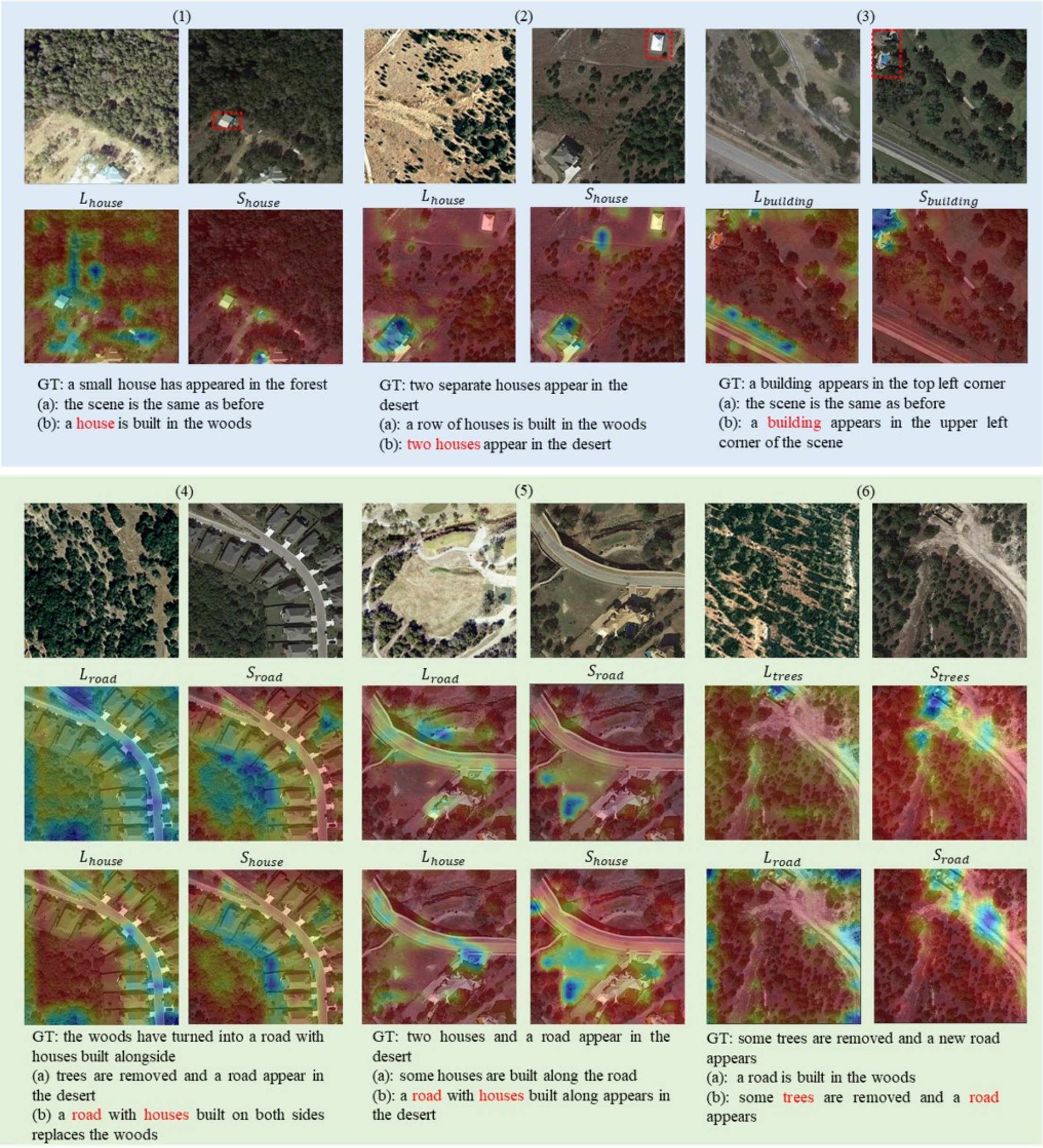

(a) GT: some detached houses and roads are constructed on the bareland
I_{t0}
I_{t1}
w_MBF: some houses are built on the bareland
w_MAF: some roads and houses are built on bareland
(b) GT: a concrete road and some houses are built on bareland
I_{t0}
I_{t1}
w_MBF: a road is built in the desert
w_MAF: a road with houses built along appears in the desert

Figure 6. Visualization of the generated attention map of the caption decoder using the existing MBF [3] method and the proposed MAF. The word highlighted in red in the caption corresponds to the blue region in the generated attention map. Note that regions appearing more blue indicate higher levels of attention.

4.4.3. Cross Gated-Attention Module

Besides observing the relation between words and single-scale change representations in Figure 6, it is also worth paying attention to the relationships between words and multi-scale change-aware features and discovering the ability of GCA that allows the change caption decoder to effectively utilize and select the useful change-aware representations for change description generation with a gated structure. Figure 7 shows the captured multi-scale word and feature attention maps, where L_{words} and S_{words} denote the attention that captures large and small changes for object words (in red) in the generated change caption, respectively. The top three pairs of results (1), (2), and (3) show the abilities of GCA in capturing the small object changes in two images, whereas (4), (5), and (6) demonstrate the capability of GCA to attend to multi-scale objects. For each set of examples, (a) is the generated change caption that only uses a single scale of features as an input and (b) is the proposed method using multi-scale features, where GT denotes the ground truth caption. In (1), we can observe that the attention map L_{house} is not able to attend to the change in the small "house" in the images, while S_{house} is able to capture it. Hence, by selecting the information of both attention weights from S_{house} using GCA, the proposed method is able to accurately generate a change caption (b) to describe the small "house" as compared to (a), which fails to capture the changes in the images. GCA is able to locate small changes in (2)(b) and (3)(b) and allows the model to generate more accurate captions as compared to the GT. In (4) and (5), we can see that GCA assists the change decoder to attend to the larger changes in the "road" and "house" for change caption generation. In (6), the decoder attends to both the larger changes in the "trees" that have been removed and the narrow/small changes in the new "road" that has been built. As a result, we can conclude that the GCA module is critical in the change caption model for identifying and selecting the essential changes of interest in the image for better change description generation.



GT: a small house has appeared in the forest
(a): the scene is the same as before
(b): a house is built in the woods

GT: two separate houses appear in the desert
(a): a row of houses is built in the woods
(b): two houses appear in the desert

GT: a building appears in the top left corner
(a): the scene is the same as before
(b): a building appears in the upper left corner of the scene

GT: the woods have turned into a road with houses built alongside
(a) trees are removed and a road appear in the desert
(b) a road with houses built on both sides replaces the woods

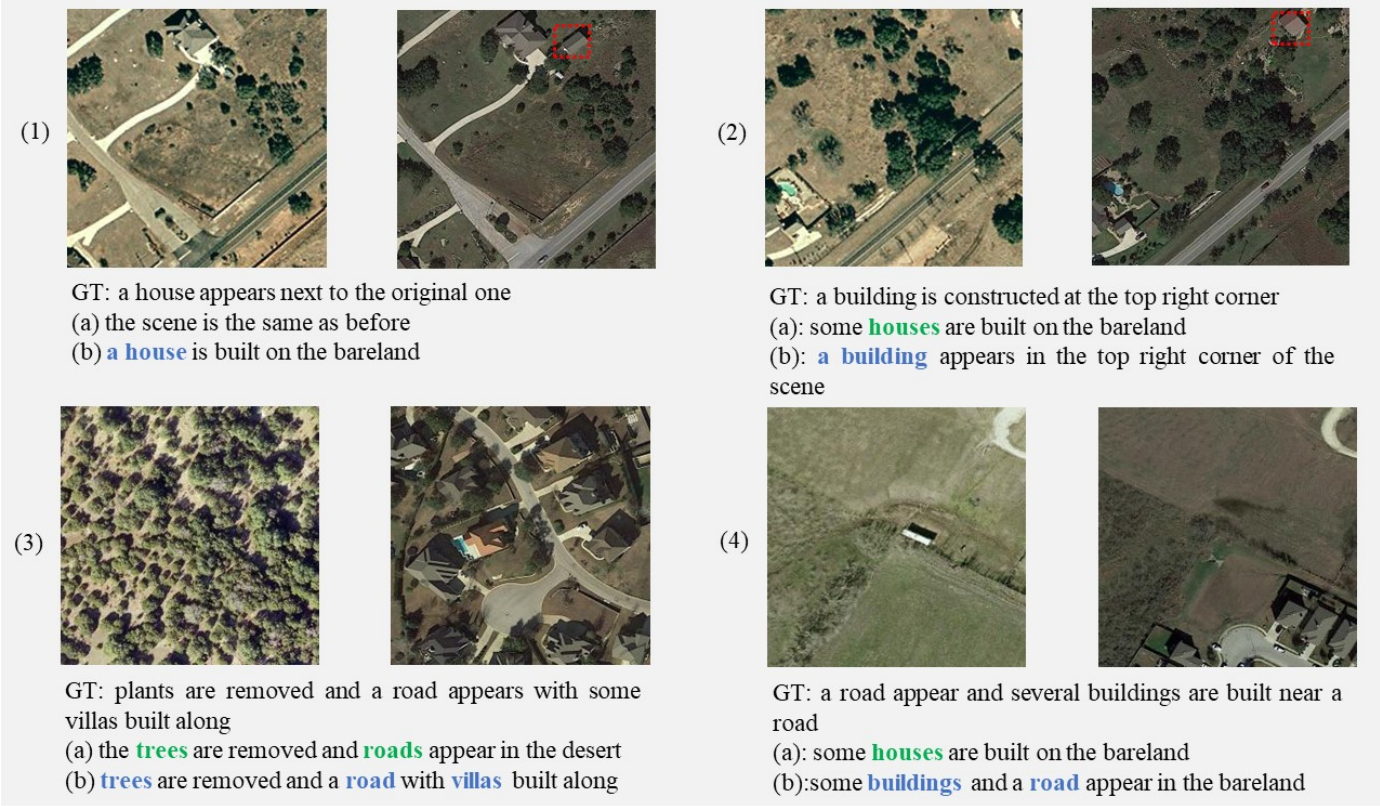
GT: two houses and a road appear in the desert
(a): some houses are built along the road
(b): a road with houses built along appears in the desert

GT: some trees are removed and a new road appears
(a): a road is built in the woods
(b): some trees are removed and a road appears

Figure 7. Visualization of captured multi-scale word and feature attention maps in the change caption decoder of the GCA module, where L_{words} and S_{words} denote the attention maps that capture large and small object changes for each object word (highlighted in red) in the generated change caption, respectively. We use the red bounding boxes to indicate the small-scale object change regions for image pairs (1), (2), and (3). (4), (5), and (6) include middle to large-scale changes. The regions appearing more blue indicate higher levels of attention.

4.5. Qualitative Analysis

Figure 8 shows the change captioning results on the LEVIR-CC dataset. For each image pair, we provide one of the five ground truth sentences and the sentences generated by an existing method [3] in (a) and our proposed method in (b). The accurately predicted change object words by our method (b) are highlighted in blue. It is observed that the proposed method generates change descriptions that are more precise and accurate compared to the existing method. For instance, our method is able to identify and describe the change in the small-scale "house" in the woods, as shown in image pairs (1) and (2), whereas the baseline method tends to predict no change or inferior results. Our method can simultaneously recognize and describe multiple changes in the objects at different scales in the bitemporal images. For instance, it accurately recognizes "trees", "villas", and the "road" in image pairs in (3)(b) rather than just the "trees" and "road" (highlighted in green) in (3)(a). Similarly, in image pair (4)(b), our method can describe the change more informatively compared to the caption generated by the baseline model in (4)(a). Our proposed method can effectively leverage distinct scale information for more precise recognition of changes in bitemporal remote sensing images and generate a more informative and accurate change description.



GT: a house appears next to the original one
(a) the scene is the same as before
(b) **a house** is built on the bareland

GT: a building is constructed at the top right corner
(a): some **houses** are built on the bareland
(b): **a building** appears in the top right corner of the scene

GT: plants are removed and a road appears with some villas built along
(a) the **trees** are removed and **roads** appear in the desert
(b) **trees** are removed and a **road** with **villas** built along

GT: a road appear and several buildings are built near a road
(a): some **houses** are built on the bareland
(b):some **buildings** and a **road** appear in the bareland

Figure 8. Qualitative results on the LEVIR-CC dataset. The I_{t0} image was captured "before", and the I_{t1} was captured "after". GT represents the ground truth caption. We use red bounding boxes to indicate the small-scale object change regions for image pairs (1) and (2). (3) and (4) include middle to large-scale changes. Green and blue words highlighted the correctly predicted change objects for the existing method (a) and ours (b), respectively.

4.6. Parametric Analysis

There are multi-layers that can be stacked in the proposed ICE and change the caption decoder. These layers of the network are essential hyperparameters that can significantly influence the performance of the model for generating change descriptions. In Table 6, we show the performance of the models when adopting different numbers of layers in both the encoder (E.L.) and decoder (D.L.). We observe that the model achieves the best performance when E.L. is equal to 2 and D.L. is equal to 1. Our ICE is composed of DAE and CAE sub-modules, which can effectively capture the change-aware features. The encoder avoids the need for additional encoders to enhance feature extraction complexity, which could impact performance. Fewer encoder layers will reduce MAF's ability to integrate multi-layer semantic information for captioning, resulting in an inferior performance. Similarly, CGA in the change caption decoder can assist the model in capturing essential multi-scale changes of interest in the image for better captioning results.

The utilization of the beam search strategy is a general approach to enhance the performance of image captioning methods, in which different beam sizes (e.g., 1, 3, 5, etc.) will affect the accuracy of the generated sentence. Table 7 demonstrates the effectiveness of using various beam sizes for caption generation. We can observe that the best performance was achieved when choosing a beam size equal to 3. A beam size that is smaller or larger than 3 can result in a lower performance. This also aligns with existing methods [3,4] that present the best results with a beam size of 3.

Table 6. Performance of the model in different layers on the LEVIR-CC dataset, where E.L and D.L denote the encoder layers and decoder layers, respectively. Bold numbers indicate the best result.

E.L.	D.L.	B-1	B-2	B-3	B-4	M	R	C
1	1	82.70	73.07	64.49	56.78	37.23	74.45	133.57
1	2	82.63	73.44	65.35	58.32	37.81	74.84	135.38
2	1	**86.06**	**78.12**	**71.45**	**66.12**	**40.51**	**75.21**	**138.36**
2	2	84.16	76.01	69.49	64.38	39.51	74.13	134.01
2	3	85.16	76.75	69.68	64.09	40.01	74.83	136.85
3	1	83.05	74.19	66.79	60.91	38.59	74.02	133.37

Table 7. Performance of the model when choosing different beam sizes during the inference stage. Bold numbers indicate the best result.

Beam Size	B-1	B-2	B-3	B-4	M	R	C
1	84.27	75.51	67.87	67.71	39.66	74.38	135.26
2	86.12	78.07	71.15	65.54	40.29	75.09	137.86
3	**86.06**	**78.12**	**71.45**	**66.12**	**40.51**	**75.21**	**138.36**
4	85.58	77.57	71.01	65.85	40.26	74.99	137.55
5	85.30	77.32	70.83	65.76	40.24	74.84	137.28
6	85.27	77.28	70.81	65.78	40.25	74.84	137.32

5. Conclusions

We introduced an Interactive Change-Aware Transformer Network (ICT-Net) to recognize changes in objects at various scales (e.g., small-scale objects) in remote sensing bitemporal images and generate a change caption to describe them accurately. We proposed the Interactive Change-Aware Encoder (ICE) to capture discrimination representations between each pair of multi-scale features and utilized a Multi-Layer Adaptive Fusion (MAF) module to aggregate relevant multi-layer change-aware features to generate better change captions. We proposed a novel Cross Gated-Attention (CGA) module to effectively utilize and select the multi-scale change-aware representations for better change captioning. We conducted extensive experiments that demonstrated the effectiveness of our proposed ICT-Net. ICT-Net significantly improves the performance of remote sensing image change captioning.

Author Contributions: Conceptualization, C.C.; methodology, C.C.; software, C.C.; validation, C.C.; formal analysis, C.C. and Y.W.; investigation, C.C.; writing—original draft preparation, C.C.; writing—review and editing, C.C. and Y.W.; visualization, C.C.; funding acquisition, Y.W.; supervision, K.-H.Y. We agreed that C.C. and Y.W. have an equal contribution. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Hong Kong Polytechnic University (PolyU) Start-up Fund for RAPs under the Strategic Hiring Scheme (P0047884).

Data Availability Statement: The LEVIR-CC and LEVIRCCD datasets can be obtained from https://github.com/Chen-Yang-Liu/RSICC (accessed on 14 November 2022) and https://disi.unitn.it/~melgani/datasets.html (accessed on 1 August 2022), respectively. Code of this work will be released at https://github.com/caicch/ICT-Net/tree/main.

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.

References

1. 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]
2. 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*, 5627414. . [CrossRef]
3. 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*, 5633520. [CrossRef]
4. Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; Shi, Z. A Decoupling Paradigm with Prompt Learning for Remote Sensing Image Change Captioning. *TechRxiv* **2023** . [CrossRef]
5. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
6. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 2048–2057.
7. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
9. Liu, C.; Yang, J.; Qi, Z.; Zou, Z.; Shi, Z. Progressive Scale-aware Network for Remote sensing Image Change Captioning. *arXiv* **2023**, arXiv:2303.00355.
10. Zhang, Z.; Zhang, W.; Yan, M.; Gao, X.; Fu, K.; Sun, X. Global visual feature and linguistic state guided attention for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615216. [CrossRef]
11. Huang, W.; Wang, Q.; Li, X. Denoising-based multiscale feature fusion for remote sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 436–440. [CrossRef]
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
13. Zhao, R.; Shi, Z.; Zou, Z. High-resolution remote sensing image captioning based on structured attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603814. [CrossRef]
14. Sumbul, G.; Nayak, S.; Demir, B. SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6922–6934. [CrossRef]
15. Wang, Y.; Zhang, W.; Zhang, Z.; Gao, X.; Sun, X. Multiscale Multiinteraction Network for Remote Sensing Image Captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2154–2165. [CrossRef]
16. Hoxha, G.; Melgani, F.; Demir, B. Toward remote sensing image retrieval under a deep image captioning perspective. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4462–4475. [CrossRef]
17. Chen, J.; Dai, X.; Guo, Y.; Zhu, J.; Mei, X.; Deng, M.; Sun, G. Urban Built Environment Assessment Based on Scene Understanding of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1436. . [CrossRef]
18. 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]
19. Zhou, H.; Du, X.; Xia, L.; Li, S. Self-Learning for Few-Shot Remote Sensing Image Captioning. *Remote Sens.* **2022**, *14*, 4606. [CrossRef]
20. Ren, Z.; Gou, S.; Guo, Z.; Mao, S.; Li, R. A Mask-Guided Transformer Network with Topic Token for Remote Sensing Image Captioning. *Remote Sens.* **2022**, *14*, 2939. [CrossRef]
21. Li, Y.; Fang, S.; Jiao, L.; Liu, R.; Shang, R. A Multi-Level Attention Model for Remote Sensing Image Captions. *Remote Sens.* **2020**, *12*, 939. [CrossRef]
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 x 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

23. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
24. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
25. Li, X.; Zhang, X.; Huang, W.; Wang, Q. Truncation Cross Entropy Loss for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5246–5257. [CrossRef]
26. Ma, X.; Zhao, R.; Shi, Z. Multiscale Methods for Optical Remote-Sensing Image Captioning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 2001–2005. [CrossRef]
27. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7296–7307. [CrossRef]
28. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
29. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure transformer network for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224713. [CrossRef]
30. Tong, L.; Wang, Z.; Jia, L.; Qin, Y.; Wei, Y.; Yang, H.; Geng, Y. Fully Decoupled Residual ConvNet for Real-Time Railway Scene Parsing of UAV Aerial Images. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 14806–14819. [CrossRef]
31. Cheng, G.; Wang, G.; Han, J. ISNet: Towards improving separability for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623811. [CrossRef]
32. Chen, G.; Zhao, Y.; Wang, Y.; Yap, K.H. SSN: Stockwell Scattering Network for SAR Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4001405. [CrossRef]
33. Bao, T.; Fu, C.; Fang, T.; Huo, H. PPCNET: A Combined Patch-Level and Pixel-Level End-to-End Deep Network for High-Resolution Remote Sensing Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1797–1801. [CrossRef]
34. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
35. Saha, S.; Bovolo, F.; Bruzzone, L. Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3677–3693. [CrossRef]
36. Tang, X.; Zhang, H.; Mou, L.; Liu, F.; Zhang, X.; Zhu, X.X.; Jiao, L. An Unsupervised Remote Sensing Change Detection Method Based on Multiscale Graph Convolutional Network and Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5609715. [CrossRef]
37. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
38. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
39. Gao, J.; Qing, L.; Li, L.; Cheng, Y.; Peng, Y. Multi-scale features based interpersonal relation recognition using higher-order graph neural network. *Neurocomputing* **2021**, *456*, 243–252. [CrossRef]
40. Wu, K.; Yang, Y.; Liu, Q.; Zhang, X.P. Focal Stack Image Compression Based on Basis-Quadtree Representation. *IEEE Trans. Multimed.* **2023**, *25*, 3975–3988. [CrossRef]
41. Wang, Y.; Hou, J.; Chau, L.P. Object counting in video surveillance using multi-scale density map regression. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2422–2426.
42. Zhou, Y.; Wang, Y.; Chau, L.P. Moving Towards Centers: Re-Ranking With Attention and Memory for Re-Identification. *IEEE Trans. Multimed.* **2023**, *25*, 3456–3468. [CrossRef]
43. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 19830–19843.
44. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
45. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
46. 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; pp. 4024–4034. [CrossRef]
47. Qiu, Y.; Satoh, Y.; Suzuki, R.; Iwata, K.; Kataoka, H. 3D-Aware Scene Change Captioning From Multiview Images. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4743–4750. [CrossRef]
48. Tu, Y.; Yao, T.; Li, L.; Lou, J.; Gao, S.; Yu, Z.; Yan, C. Semantic Relation-aware Difference Representation Learning for Change Captioning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 63–73. [CrossRef]
49. 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 (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 1971–1980.

50. 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 (ACL), Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318. [CrossRef]
51. 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.
52. 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; pp. 65–72.
53. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
55. Park, D.H.; Darrell, T.; Rohrbach, A. Robust Change Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.

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.