remote sensing

MDPI

Article

Remote Sensing Image Change Captioning Using Multi-Attentive Network with Diffusion Model

Yue Yang [1,*], Tie Liu [1], Ying Pu [1], Liangchen Liu [1], Qijun Zhao [1] and Qun Wan [2]

[1] College of Computer Science, Sichuan University, Chengdu 610025, China; 2022326045009@stu.scu.edu.cn (T.L.); 2024223045042@stu.scu.edu.cn (Y.P.); 2024223045059@stu.scu.edu.cn (L.L.); qjzhao@scu.edu.cn (Q.Z.)

[2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

* Correspondence: yueyang7@scu.edu.cn

Abstract: Remote sensing image change captioning (RSICC) has received considerable research interest due to its ability of automatically providing meaningful sentences describing the changes in remote sensing (RS) images. Existing RSICC methods mainly utilize pre-trained networks on natural image datasets to extract feature representations. This degrades performance since aerial images possess distinctive characteristics compared to natural images. In addition, it is challenging to capture the data distribution and perceive contextual information between samples, resulting in limited robustness and generalization of the feature representations. Furthermore, their focus on inherent most change-aware discriminative information is insufficient by directly aggregating all features. To deal with these problems, a novel framework entitled Multi-Attentive network with Diffusion model for RSICC (MADiffCC) is proposed in this work. Specifically, we introduce a diffusion feature extractor based on RS image dataset pre-trained diffusion model to capture the multi-level and multi-time-step feature representations of bitemporal RS images. The diffusion model is able to learn the training data distribution and contextual information of RS objects from which more robust and generalized representations could be extracted for the downstream application of change captioning. Furthermore, a time-channel-spatial attention (TCSA) mechanism based difference encoder is designed to utilize the extracted diffusion features to obtain the discriminative information. A gated multi-head cross-attention (GMCA)-guided change captioning decoder is then proposed to select and fuse crucial hierarchical features for more precise change description generation. Experimental results on the publicly available LEVIR-CC, LEVIRCCD, and DUBAI-CC datasets verify that the developed approach could realize state-of-the-art (SOTA) performance.

Keywords: change captioning (CC); image captioning; remote sensing (RS); diffusion model; attention mechanism

Citation: 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. https://doi.org/10.3390/rs16214083

Academic Editor: Salah Bourennane

Received: 30 July 2024
Revised: 25 October 2024
Accepted: 27 October 2024
Published: 1 November 2024

1. Introduction

With the increasing availability of multi-temporal remote sensing (RS) data, the RS image change captioning (RSICC) technologies have become a research trend; they can generate accurate descriptions of changes between RS images [1,2]. This ability enables the RSICC techniques to be extensively applied in various applications, including environmental monitoring, landscape damage examination, and urban planning, to name just a few [3,4].

Compared to conventional RS change detection methods [5–8], RSICC is more flexible and complex since it requires fully comprehending visual contents and describing them in natural language automatically. RSICC is essentially a multi-disciplinary interaction task, where both computer vision (CV) and natural language processing (NLP) techniques are involved. The encoder–decoder pattern enabled by deep neural network has attracted much research attention for the RSICC task. Herein, the CV technique serves as an encoder to extract key features and calculates the differences between bitemporal images, while the NLP approach works as a decoder to recover the encoded visual pattern into word sequence to represent the differences [9–12]. For instance, Hoxha et al. [1,2] first used a convolutional neural network (CNN) as an encoder to extract the temporal changes in scenes and utilized a recurrent neural network (RNN) to decode the former and generate sentences of the changes. However, these models mainly concentrate on feature extraction, while the long-range contextual information within the spatial and temporal scope is overlooked. To address this constraint, Transformer architectures have been introduced into RSICC to gain a greater receptive field, thereby facilitating change caption generation [13–16]. In [13–16], the authors introduced a Transformer encoder combined with a CNN feature extractor, followed by a Transformer-based decoder for change descriptions generation. Even though these methods have achieved favorable outcomes, they still suffer from three main limitations: first, ignorance of the distinctive characteristics between natural images and RS images, such as differences in observation view, color, texture, layout and objects, which consequently leads to capture ineffective changes and generate inferior change caption; second, inability to mine the data distribution and fully representat the data generation process, thus having limited robustness and generalization; third, difficulty in recognizing inherent most change-aware discriminative information because of directly aggregating all features.

Recently, the diffusion model has gained great interest in the field of image generation due to the strong denoising generation ability [17,18]. Compared with the classical networks such as generative adversarial network (GAN) and Transformer, the diffusion model shows higher performance in many fields, e.g., super-resolution [19], segmentation [20], inpainting [21], and conditional image generation [22]. Besides its impressive capability in image generation, the denoising process in the diffusion model could well describe the semantic correlations of temporal–spatial features within images, rendering it possible to recover the semantically relevant features in the original image [23–29]. Specifically, the diffusion model utilizes probabilistic modeling to capture the diversity and complexity of input data and establishes an intricate data distribution model. This enables the model to improve its semantic extraction capability, making it suitable for many scenarios and image features like the shape, size, and texture of targets, which is instrumental in acquiring accurate change captions.

On the basis of the above analysis, in this paper, a RSICC framework based on Multi-Attentive network with Diffusion model, termed as MADiffCC, is proposed. The devised framework is composed of three parts: a diffusion feature extractor, a time-channel-spatial attention (TCSA)-based difference encoder, and a gated multi-head cross-attention (GMCA)-guided change captioning decoder. First, the diffusion feature extractor pre-trained on an RS image dataset is introduced to retrieve multi-scale feature representations from the original bitemporal RS images within the reverse diffusion process. Since the pre-training is achieved by a denoising diffusion probabilistic model (DDPM) with noisy images, it yields more robust and generalized representations to Gaussian noise. Furthermore, the difference encoder in conjunction with the TCSA mechanism is proposed to compute the difference representations of the diffusion features, while the change captioning decoder combined with the GMCA mechanism is designed to select and fuse the key hierarchical difference information for generating change descriptions. By doing so, the model's capabilities of obtaining change-aware discriminative features can be enhanced, which helps improving the change captioning performance.

The main contributions of this paper include:

(1) A hierarchical feature extractor based on an RS image dataset pre-trained diffusion model is proposed for RSICC. It learns the data distribution with noisy images and fully considers the semantic correlations of temporal–spatial features, thereby strengthening the robustness and generalization of feature representations. To the best of our knowledge, this is the first trial to apply a diffusion model to RSICC.

(2) A TCSA-based difference encoder is designed to compute the change-aware discriminative information to the fullest extent and filter out the irrelevant changes through the TCSA mechanism and difference encoding.
(3) We introduce a GMCA-guided change captioning decoder which enables the network to learn and fuse essential multi-scale features, facilitating better ability of change captioning.

The remainder of this paper is organized as follows. In Section 2, the related work of this article is summarized. Section 3 describes the developed method for RSICC in detail. In Section 4, experiments and analyses are given, followed by the conclusion remarks in Section 5.

2. Related Work

In this section, the prior research associated with the developed approach is reviewed, consisting of the introduction of existing techniques for RSICC and the overview of the diffusion model.

2.1. Remote Sensing Image Change Captioning

The RSICC task has received great attention during recent years thanks to its ability of illustrating the differences between bitemporal RS images via natural language. With the development of CNN architectures such as VGG [30] and ResNet [31], many researchers have studied their pre-trained models on large-scale natural image datasets like ImageNet [32] and ADE20k [33]. On this basis, RS image analysis employing transfer learning from natural images to RS images has emerged. Hoxha et al. [1] first explored a VGG-16 pre-trained on ImageNet as the encoder for feature extraction from bitemporal RS images and utilized an RNN as the decoder for change description generation. Then, the authors in [2] designed two encoders on the basis of an image- or feature-based fusion framework and utilized two caption decoders, i.e., an RNN and an SVM to generate change descriptions. However, these works solely concentrate on feature extraction based on CNN while overlooking the long-term dependencies in sequential data which is critical in identifying the change of interest in RS images for change captioning.

Accordingly, the latest efforts have been focused on increasing the model's reception field through applying attention mechanisms [34,35]. For instance, the dual dynamic attention (DUDA) model [36] was proposed by employing a spatial dual attention module to identify changes and a semantic attention module to highlight difference feature representations for captioning. Nevertheless, since the attention is simply applied by reweighting the fused bitemporal features/images in the channel or spatial dimension, these methods are still struggling to relate long-range concepts in space-time.

Recently, the emerging Transformer networks [37] with the idea of multi-head attention (MHA) have been successfully applied in image analysis and have shown good performance in image change captioning [13–15,38–40]. For example, in [39], MCCFormers-S and MCCFormers-D were proposed for the image change captioning task, where the MHA mechanism in Transformer is able to obtain long-range context and relationships between different positions. In particular, MCCFormers-S directly inputs the bitemporal image features extracted by CNNs into a Transformer encoder for dense feature interaction, while MCCFormers-D uses Siamese Transformer encoders with the coattention mechanism to capture relevance between local regions of two images. Inspired by these approaches, RSICCFormer [13] was developed via employing a dual-branch Transformer encoder with a CNN feature extractor to improve the feature discrimination capacity for RSICC. Later, the authors of [14] improved the method in [13] and developed the Chg2Cap model, a Siamese attentive Transformer encoder that consists of a hierarchical self-attention module and a residual module which has achieved great success. Furthermore, for enhancing the capability of changed object perception with different sizes, the authors in [15] developed a progressive scale-aware network for caption decoding to extract multi-scale discriminative features. In [38], Liu et al. proposed to utilize prompt learning in a decoupling paradigm and adopted the large language model (LLM) trained for the RSICC task in which multiple prompts including change classes, language representation, and visual features are cascaded with a frozen LLM for change captioning.

While existing methods have shown satisfactory performance, they still face challenges. Most of the pre-trained models for RSICC are operating on latent feature representations obtained from ImageNet pre-trained ResNet. The RSICC performance is seriously limited when using these pre-trained models, since aerial images possess distinctive characteristics compared to natural images, such as the differences in observation view, image color, texture, layout and objects. Fortunately, with the emergence of large-scale aerial scene datasets, the authors in [23] pre-trained a DDPM using readily available RS image datasets. Motivated by this, in our paper, we utilize the pre-trained DDPM to produce feature representations for the RSICC task. DDPM is able to capture data distribution and enable the extraction of highly informative and compressed feature representations of a given image. Because DDPM is trained using noisy images, it is conceivable that more robust and generalized representations can be acquired.

2.2. Diffusion Model

Compared to CNNs and Transformers, the diffusion model offers several advantages that enable it to infer the inherent patterns of a dataset and to capture the complex data distribution. Currently, the diffusion model can be divided into three main categories, i.e., the score-based generative model [41], stochastic differential equations (SDE) [42], and DDPM [17]. The multi-scale feature representations obtained via the diffusion model have been proven to be very powerful in various applications, like classifying images, detecting objects, and so on [23–29]. By approximating the distribution of the input dataset, the diffusion model has potential to represent the inherent structures and relationships among images. However, although the diffusion model has been successfully applied in natural image analysis and behaves well in many tasks, there is little research employing the diffusion model in RS applications. Recently, the authors in [23] pre-trained a DDPM using a large amount of unlabeled RS images and utilized it as a pre-trained feature extractor. Extensive experimental results in [23] demonstrated the pivotal role of a pre-trained DDPM as a feature extractor for downstream applications. On this basis, a diffusion model-based multi-attentive framework for RSICC is developed in this article. The proposed framework is able to analyze the RS data distribution with high dimensionality and high redundancy and to effectively exploit the multi-scale feature representations extracted from pre-trained DDPM, thereby enhancing the performance of RSICC.

3. Methodology

This section offers a detailed description of the devised MADiffCC model as summarized in Figure 1. It is a combination of a diffusion feature extractor, a TCSA-based difference encoder, and a GMCA-guided change captioning decoder. In the beginning, the RS image pairs are fed into the RS image dataset pre-trained DDPM, such that multi-scale diffusion features could be extracted. Then, the TCSA-based difference encoder is utilized to enhance the feature discrimination capability. Finally, the GMCA-guided change captioning decoder is adopted to translate the encoded multi-scale difference features into language descriptions. By doing so, the developed framework is able to illustrate the relationship between RS scenes, focus on changes of interest, and provide accurate change captioning.

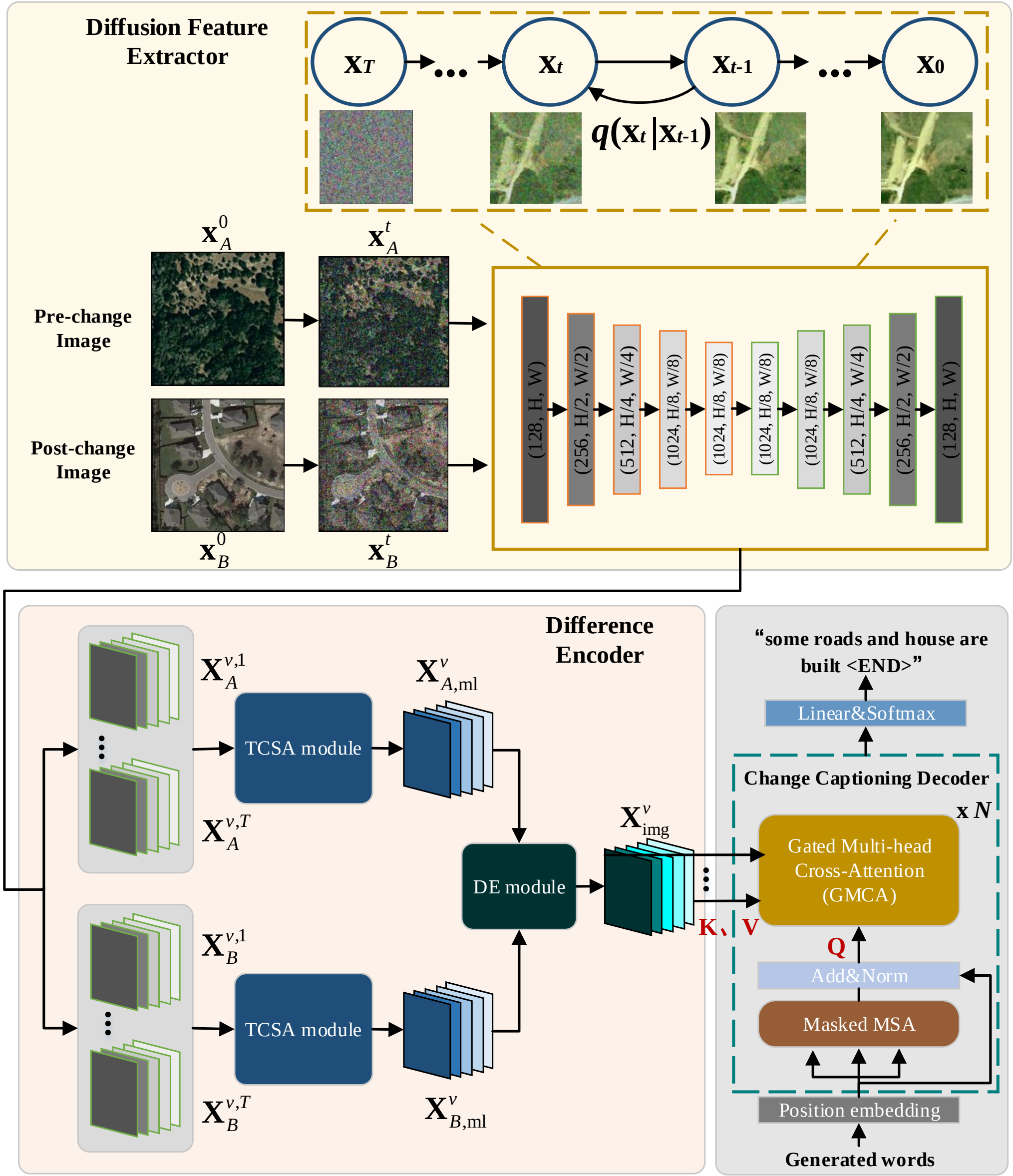


Figure 1. Overall framework of the proposed MADiffCC. It includes three main components: a diffusion feature extractor based on RS image dataset pre-trained DDPM to retrieve semantically related multi-level and multi-time-step feature representations, a TCSA-based difference encoder to effectively obtain most change-aware difference representations, and a GMCA-guided change captioning decoder to learn critical difference information for generating change descriptions.

3.1. Diffusion Feature Extractor

DDPM is endowed with the ability to imitate the distribution $p(\mathbf{x}_0)$ of training data $\mathbf{x}_0$ by virtue of variational inference on the basis of a Markov chain process with T time steps. For the sake of obtaining the joint latent structure of images, one could add Gaussian noises to the input data during the forward process while mitigating the noise by training a denoising network during the reverse process. In the forward diffusion procedure, the clean image $\mathbf{x}_0$ is gradually contaminated with Gaussian noise until the image is destroyed completely and one obtains an isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Mathematically, at the tth time step, the image instance contaminiated with noise is formulated as

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \quad (1)$$

where $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, respectively, represent the image instances at time steps $t-1$ and t, and α_t is the noise schedule. Under the Markovian assumptions, given a clean data sample $\mathbf{x}_0$, a noisy sample $\mathbf{x}_t$ could be derived by sampling Gaussian vectors with distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as

$$\mathbf{x}_t = \sqrt{\gamma_t}\mathbf{x}_0 + \sqrt{1-\gamma_t}\boldsymbol{\epsilon} \tag{2}$$

where $\gamma_t = \prod_{i=1}^{t} \alpha_i$.

For the training of the denoising network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, a similar architecture to U-ViT [43] for $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is generated to predict $\boldsymbol{\epsilon}$ by minimizing the following training objective function:

$$\mathbb{L}_{\mathbf{x}_t,t,\boldsymbol{\epsilon}} = \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\|_1 \tag{3}$$

In the reverse diffusion stage, the noisy image $\mathbf{x}_t$ is denoised to approximate the instance corresponding to the previous step (i.e., $\mathbf{x}_{t-1}$), using the trained model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$. This procedure is formulated mathematically as

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)) + \gamma_t \mathbf{g} \tag{4}$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A sequence of denoising operations leads to the recovery of the original image $\mathbf{x}_0$.

In this paper, the pre-trained DDPM trained by many unlabeled RS images [23] is utilized to extract the change feature representations between bitemporal RS images for RSICC. More specifically, a great number (approximately 500 k) of the RS images of Sentinel-2 [44] patches (with earth resolution of approximately 2.65 × 2.65 km) are served as the training dataset to pre-train the SOTA unconditional diffusion model.

As depicted in Figure 1, we employ the pre-trained DDPM to extract feature representations for bitemporal RS images $\mathbf{x}_A^0$ and $\mathbf{x}_B^0$. It is noted that the weights of pre-trained DDPMs are being frozen during the feature extraction. First, the noisy RS images $\mathbf{x}_A^t$ and $\mathbf{x}_B^t$ of the given original images $\mathbf{x}_A^0$ and $\mathbf{x}_B^0$ are acquired by Formula (2). Then, we input the noisy RS images $\mathbf{x}_A^t$ and $\mathbf{x}_B^t$ into the pre-trained DDPM to obtain the multi-scale feature representations $\mathbf{X}_A^{v,t} \in \mathbb{R}^{A_\mathrm{h} \times A_\mathrm{w} \times C}$ and $\mathbf{X}_B^{v,t} \in \mathbb{R}^{A_\mathrm{h} \times A_\mathrm{w} \times C}$, with A_h, A_w and C, respectively, being the height, width, and channel dimension of the feature map.

In particular, the multi-scale feature representation includes two components, i.e., multi-level $v \in \{1, ..., V\}$ and multi-time-step $t \in \{1, ..., T\}$. Multi-level $v \in \{1, ..., V\}$ feature representations can be expressed as

$$\left\{\mathbf{X}_A^{v=1,t}, ..., \mathbf{X}_A^{v=V,t}\right\} = \mathcal{F}_{\mathrm{DDPM}}(\mathbf{x}_A^t) \tag{5}$$

$$\left\{\mathbf{X}_B^{v=1,t}, ..., \mathbf{X}_B^{v=V,t}\right\} = \mathcal{F}_{\mathrm{DDPM}}(\mathbf{x}_B^t) \tag{6}$$

with $v \in \{1, ..., V\}$ representing level of the feature extracted from pre-trained DDPM $\mathcal{F}_{\mathrm{DDPM}}(\cdot)$. Usually, the smallest level is 1 and its related highest spatial resolution is $(H/2^{(v-1)}, W/2^{(v-1)}) = (H, W)$, while the largest level is 5 and it corresponds to the lowest spatial resolution $(H/2^{(v-1)}, W/2^{(v-1)}) = (H/16, W/16)$.

Furthermore, the multi-time-step $t \in \{1, ..., T\}$ feature representations corresponding to multiple noisy versions of the bitemporal RS images are also employed to obtain RSICC performance gains, expressed as

$$\left\{\mathbf{X}_A^{v=1,t}, ..., \mathbf{X}_A^{v=V,t}\right\}_{t=1}^{t=T} = \left\{\mathcal{F}_{\mathrm{DDPM}}(\mathbf{x}_A^t)\right\}_{t=1}^{t=T} \tag{7}$$

$$\left\{\mathbf{X}_B^{v=1,t}, ..., \mathbf{X}_B^{v=V,t}\right\}_{t=1}^{t=T} = \left\{\mathcal{F}_{\mathrm{DDPM}}(\mathbf{x}_B^t)\right\}_{t=1}^{t=T} \tag{8}$$

The multi-level and multi-time-step feature representations are able to provide objects with rich information. It is noticed that the features extracted from larger level relate to simpler patterns like edges and gradients, while those from smaller level stand for more abstract features, demonstrating the abilities of DDPM in extracting hierarchical features (i.e., semantics) from input images.

3.2. TCSA-Based Difference Encoder

It is critical for RSICC to obtain the difference information reflecting change regions. The TCSA-based difference encoder is designed to obtain the highly discriminative features between diffusion feature pairs. As shown in Figure 2, it is composed of two components: a time-channel-spatial attention (TCSA) module and a difference encoding (DE) module.

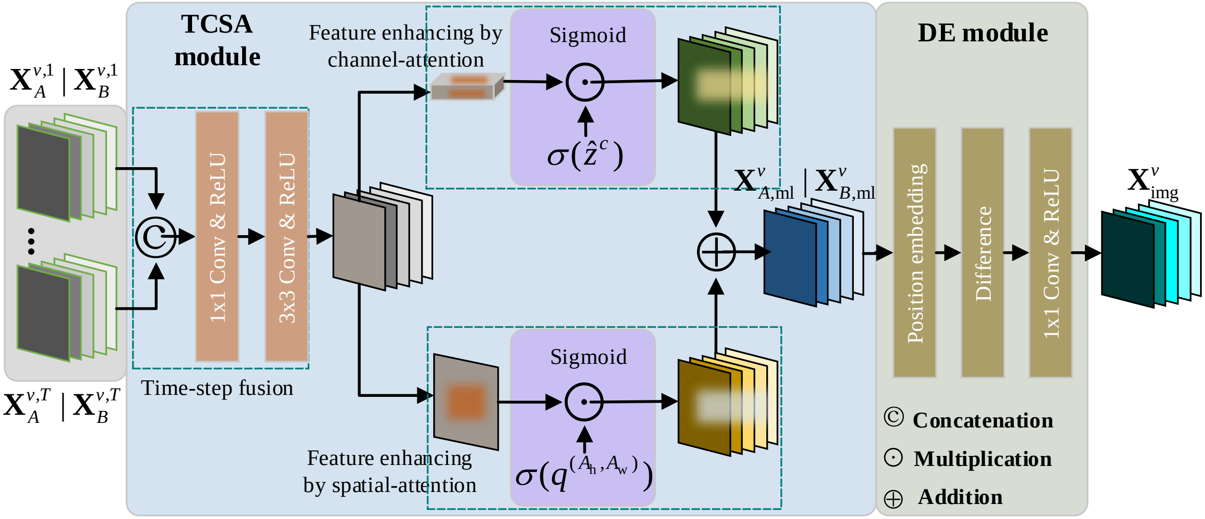

TCSA module
Feature enhancing by channel-attention
Sigmoid
$\sigma(\hat{z}^c)$
$\mathbf{X}_A^{v,1} \mid \mathbf{X}_B^{v,1}$
$\mathbf{X}_A^{v,T} \mid \mathbf{X}_B^{v,T}$
1x1 Conv & ReLU
3x3 Conv & ReLU
Time-step fusion
$\mathbf{X}_{A,\mathrm{ml}}^v \mid \mathbf{X}_{B,\mathrm{ml}}^v$
Sigmoid
$\sigma(q^{(A_h,A_w)})$
Feature enhancing by spatial-attention
DE module
Position embedding
Difference
1x1 Conv & ReLU
$\mathbf{X}_{\mathrm{img}}^v$
© Concatenation
⊙ Multiplication
⊕ Addition

Figure 2. Visualization of the TCSA-based difference encoder.

3.2.1. TCSA Module

In order to focus on the changes of interest and suppress the redundant information, we introduce the TCSA module which computes the semantic information along different time-step features and performs attention on multi-level concatenated features. For simplicity, here, we omit subscripts A and B of the multi-scale feature representations $\mathbf{X}_A^{v,t}$ and $\mathbf{X}_B^{v,t}$. First, the diffusion feature representations $\mathbf{X}^{v,t}$ of time-step t and level v are processed as follows:

$$\mathbf{X}_{\mathrm{mt}}^{v} = \mathrm{ReLU}(\mathrm{Conv}_3(\mathrm{ReLU}(\mathrm{Conv}_1(\mathrm{Concat}(\mathbf{X}^{v,1}, .., \mathbf{X}^{v,T}))))) \tag{9}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation along channel dimension, $\mathrm{ReLU}(\cdot)$ is the Rectified Linear Unit (ReLU) activation function, and Conv_n denotes the 2D convolutional layer with kernel size of $n \times n$.

The feature maps $\mathbf{X}_{\mathrm{mt}}^{v}$ are further fine-grained via the channel-spatial attention operation [45], respectively, to concurrently capture changes of interest while ignoring irrelevant changes. On one hand, considering $\mathbf{X}_{\mathrm{mt}}^{v} = [\mathbf{x}_{\mathrm{mt}}^{v,1}, ..., \mathbf{x}_{\mathrm{mt}}^{v,C}]$ as a combination of channels, channel attention is achieved via a global average pooling layer, yielding vector $\mathbf{z} \in \mathbb{R}^{1\times1\times C}$. By doing so, the global spatial information is embedded in vector $\mathbf{z}$. Then, we transform this vector for encoding the channel-wise dependencies, that is, $\hat{\mathbf{z}} = \mathbf{W}_1(\mathrm{ReLU}(\mathbf{W}_2\mathbf{z}))$, with $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively, denoting weights of fully connected layers and the ReLU operator. After that, sending $\hat{\mathbf{z}}$ into a sigmoid layer, we obtain

$$\mathbf{X}_{\mathrm{ml-c}}^{v} = \left[\sigma(\hat{z}^1)\mathbf{x}_{\mathrm{mt}}^{v,1}, ..., \sigma(\hat{z}^C)\mathbf{x}_{\mathrm{mt}}^{v,C}\right] \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The significance of the cth channel is indicated by the activation $\sigma(\hat{z}^c)$. Within the training process, the activations can be tuned automatically to put more importance on important channels while ignoring less contributing ones.

On the other hand, the spatial attention operation is achieved through a convolution of $\mathbf{X}^v_{\mathrm{mt}}$ to produce a projection tensor $\mathbf{q} \in \mathbb{R}^{A_\mathrm{h} \times A_\mathrm{w}}$, which stands for the linear combination of all channels for spatial locations. This projection is processed by a sigmoid layer, i.e.,

$$\mathbf{X}^v_{\mathrm{ml-s}} = \left[\sigma(q^{(1,1)})\mathbf{x}^{v,(1,1)}_{\mathrm{mt}}, ..., \sigma(q^{(A_\mathrm{h},A_\mathrm{w})})\mathbf{x}^{v,(A_\mathrm{h},A_\mathrm{w})}_{\mathrm{mt}}\right] \tag{11}$$

Formula (11) focuses on relevant spatial positions and neglects irrelevant ones.

Finally, the concurrent channel and spatial features are obtained by element-wise addition as

$$\mathbf{X}^v_{\mathrm{ml}} = \mathbf{X}^v_{\mathrm{ml-c}} + \mathbf{X}^v_{\mathrm{ml-s}} \tag{12}$$

3.2.2. DE Module

In the DE module, we aim to construct accurate feature difference embedding between the feature pairs $\{\mathbf{X}^v_{A,\mathrm{ml}}, \mathbf{X}^v_{B,\mathrm{ml}}\}$ obtained from the previous TCSA module. To begin with, the DE module takes feature maps with position embeddings as inputs. Then, the difference between two feature maps is calculated and further sent to a 2D convolutional layer with a ReLU activation unit. Mathematically, the processing procedures within the DE module are

$$\mathbf{X}^v_{A,\mathrm{pos}} = \mathbf{X}^v_{A,\mathrm{ml}} + \mathbf{X}_{\mathrm{pos-0}} \tag{13}$$

$$\mathbf{X}^v_{B,\mathrm{pos}} = \mathbf{X}^v_{B,\mathrm{ml}} + \mathbf{X}_{\mathrm{pos-0}} \tag{14}$$

$$\mathbf{D}^v = \mathbf{X}^v_{A,\mathrm{pos}} - \mathbf{X}^v_{B,\mathrm{pos}} \tag{15}$$

$$\mathbf{X}^v_{\mathrm{img}} = \mathrm{ReLU}(\mathrm{Conv}_1(\mathbf{D}^v)) \tag{16}$$

in which $\mathbf{X}_{\mathrm{pos-0}} \in \mathbb{R}^{A_\mathrm{h} \times A_\mathrm{w} \times C}$ denotes a tunable 2D position embedding. It combines the spatial position information into bitemporal feature sequences.

The multi-scale feature difference maps $\mathbf{X}^v_{\mathrm{img}}, v \in \{1, ..., V\}$ with resolution of $A^v_\mathrm{h} \times A^v_\mathrm{w} \times C^v$ are processed through a downsampling operation to unify the resolution as $A^1_\mathrm{h} \times A^1_\mathrm{w} \times C^1$. Towards this end, we utilize a 2D convolutional layer with 3×3 kernel size, stride $S = 1$ and padding $P = 1$ for the initial downsampling, and kernel size of 4×4, stride $S = 2$ and padding $P = 1$ for the rest.

3.3. GMCA-Guided Change Captioning Decoder

As illustrated in Figure 3, the GMCA-guided change captioning decoder utilizes the previously obtained change features of multi-scale image from the difference encoder. To obtain a satisfactory result in the change description process, an improved Transformer decoder integrated with a GMCA mechanism is proposed. Not only can the GMCA module allow us to take advantage of all the multi-scale representations, but it can also enable the captioning decoder to concentrate on essential crucial features by virtue of the gated structure.

At the training stage for the sentence generation, we first transform the original text tokens m into the word embeddings $\mathbf{X}_{\mathrm{txt}} \in \mathbb{R}^{M \times d_m}$, which serve as inputs of the decoder, with M denoting the sentence length and d_m being the dimension of words embedding. Then, the masked multihead self-attention (MMSA) with the head number of H is formulated as

$$\begin{aligned}\mathbf{S}_{\mathrm{txt}} &= \mathrm{MMSA}(\mathbf{X}_{\mathrm{txt}}) \\ &= \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_H)\mathbf{W}^{\mathrm{O}}\end{aligned} \tag{17}$$

where the single-head attention is given by

$$\mathrm{head}_h = \mathrm{Att}(\mathbf{X}_{\mathrm{txt}}\mathbf{W}_h^{\mathrm{Q}}, \mathbf{X}_{\mathrm{txt}}\mathbf{W}_h^{\mathrm{K}}, \mathbf{X}_{\mathrm{txt}}\mathbf{W}_h^{\mathrm{V}}) \tag{18}$$

and it is a triple (query **Q**, key **K**, value **V**) whose formulation is

$$\mathrm{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_m}})\mathbf{V} \tag{19}$$

In the above formulas, $\mathbf{W}^{\mathrm{O}} \in \mathbb{R}^{Hd_{\mathrm{m}} \times C}$ is the linear projection matrix to aggregate the feature channel dimension. $\mathbf{W}_h^{\mathrm{Q}}, \mathbf{W}_h^{\mathrm{K}}, \mathbf{W}_h^{\mathrm{V}} \in \mathbb{R}^{C \times d_{\mathrm{m}}}$ are the learnable weight matrices for query, key, and value of the word embedding of the hth head. $\mathrm{Softmax}(\cdot)$ denotes a softmax activation function. Different from MSA, the advantage of MMSA is that each predicted word relies only on the generated words with the aid of mask operation.

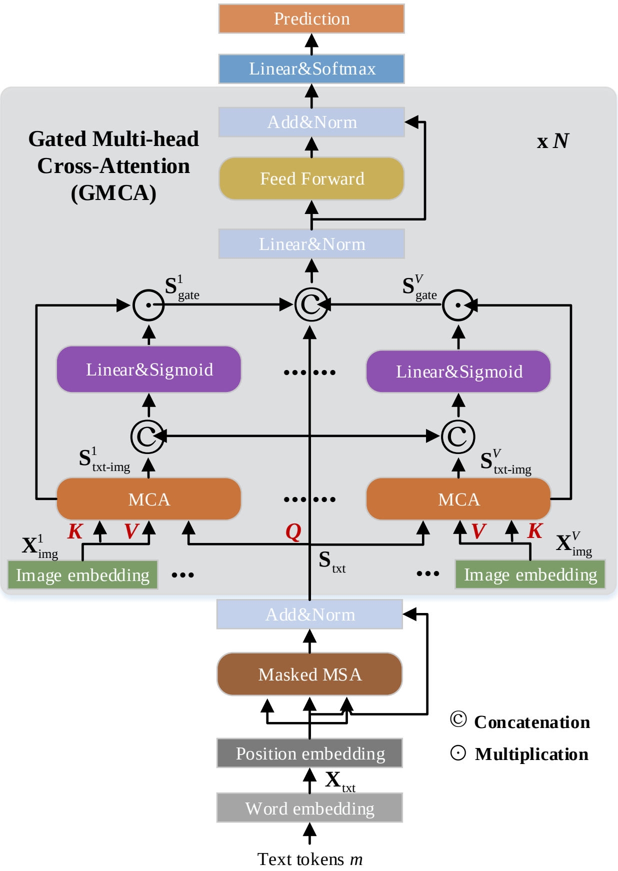

Prediction
Linear&Softmax
Add&Norm
Gated Multi-head Cross-Attention (GMCA)
x N
Feed Forward
Linear&Norm
$\mathbf{S}^1_{\mathrm{gate}}$
$\mathbf{S}^V_{\mathrm{gate}}$
Linear&Sigmoid
......
Linear&Sigmoid
$\mathbf{S}^1_{\text{txt-img}}$
$\mathbf{S}^V_{\text{txt-img}}$
MCA
......
MCA
$\mathbf{X}^1_{\mathrm{img}}$
K
V
Q
V
K
$\mathbf{X}^V_{\mathrm{img}}$
$\mathbf{S}_{\mathrm{txt}}$
Image embedding
...
...
Image embedding
Add&Norm
Masked MSA
© Concatenation
⊙ Multiplication
Position embedding
$\mathbf{X}_{\mathrm{txt}}$
Word embedding
Text tokens m

Figure 3. Visualization of the GMCA-guided change captioning decoder.

For the GMCA mechanism of the captioning decoder, it connects the generated word feature sequences $\mathbf{S}_{\mathrm{txt}}$ from the previous MMSA with multi-scale change-aware feature maps $\mathbf{X}^v_{\mathrm{img}}$ outputted by the difference encoder. The procedure can be expressed as follows:

$$\begin{aligned} \mathbf{S}^v_{\mathrm{txt-img}} &= \mathrm{MCA}(\mathbf{S}_{\mathrm{txt}}, \mathbf{X}^v_{\mathrm{img}}) \\ &= \mathrm{Concat}(\mathrm{head}^v_1, \ldots, \mathrm{head}^v_H)\mathbf{W}^{\mathrm{O}} \end{aligned} \tag{20}$$

where

$$\mathrm{head}^v_h = \mathrm{Att}(\mathbf{S}_{\mathrm{txt}}\mathbf{W}_h^{\mathrm{Q}}, \mathbf{X}^v_{\mathrm{img}}\mathbf{W}_h^{\mathrm{K}}, \mathbf{X}^v_{\mathrm{img}}\mathbf{W}_h^{\mathrm{V}}) \tag{21}$$

Then, the gated attention is utilized to highlight relevant changes of interest among $\mathbf{S}^{v}_{\text{txt}-\text{img}}$ for change description generation, which is expressed as

$$\mathbf{S}^{v}_{\text{gate}} = \sigma(\text{Linear}(\text{Concat}(\mathbf{S}_{\text{txt}}, \mathbf{S}^{v}_{\text{txt}-\text{img}}))) \odot \mathbf{S}^{v}_{\text{txt}-\text{img}} \tag{22}$$

where $\odot$ is the element-wise multiplication, and $\text{Linear}(\cdot)$ denotes the linear projection layer. The sentence features $\mathbf{S}^{v}_{\text{gate}}$ can be fused by

$$\mathbf{Y} = \text{LN}(\text{Linear}(\text{Concat}(\mathbf{S}^{1}_{\text{gate}}, ..., \mathbf{S}^{V}_{\text{gate}}, \mathbf{S}_{\text{txt}}))) \tag{23}$$

in which $\text{LN}(\cdot)$ stands for layer normalization (LN).

To obtain the word probabilities $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_M] \in \mathbb{R}^{M \times K}$ where K denotes the vocabulary size, one feeds the output of the captioning decoder, $\mathbf{Y}$, to a linear layer with softmax activation. The word probability vector $\mathbf{p}_m = [p_m^{(1)}, ..., p_m^{(K)}]$ is then used to infer the most probable word at position m in the generated sentence. In formulation, we have

$$\mathbf{p}_m = \text{Softmax}(\mathbf{y}_m) = \frac{\exp(\mathbf{y}_m)}{\sum_{k=1}^{K} \exp(y_m^{(k)})} \tag{24}$$

where $\mathbf{y}_m = [y_m^{(1)}, ..., y_m^{(K)}]$ is obtained from the linear layer.

The training criteria of our model is to minimize the cross-entropy loss between the word probability predictions and the reference sentences. Mathematically, it is formulated as

$$\mathcal{L} = -\sum_{m=1}^{M} \log\left(\sum_{k=1}^{K} \bar{p}_m^{(k)} p_m^{(k)}\right) \tag{25}$$

where $\bar{\mathbf{p}}_m = [\bar{p}_m^{1}, ..., \bar{p}_m^{(K)}]$ is used to represent the word vector at position m within the reference caption.

4. Experimental Results and Analysis

4.1. Datasets

4.1.1. LEVIR-CC Dataset

The experiments in this section are carried out using the publicly available large-scale The LEVIR-CC dataset [13], which consists of both RS image pairs and the associated change captions. LEVIR-CC dataset contains a total of 10,077 bitemporal RS image pairs, in which there are 5038 image pairs with changed objects and 5039 image pairs with no changes. These images are fixed as 256 × 256 pixels. The numbers of image pairs used for the training set, the validation set, and the testing set are, respectively, 6815, 1333, and 1929. Each image pair is labeled with five sentences to provide the descriptions of changes occurred between them. The number of the corresponding ground truth change descriptions is 50,385, where 25,190 sentences are utilized to express changed image pairs and the remaining 25,195 sentences are adopted to describe image pairs without changed regions.

4.1.2. LEVIRCCD Dataset

The experiemnts are further carried out by using the LEVIRCCD dataset [2] to highlight the effectiveness of the proposed scheme. The size of each image in the LEVIRCCD dataset is 256 × 256 pixels, and each image is annotated with 5 change descriptive sentences. The total number of images is 500 pairs, where 60% of the image pairs with their change desciprtions are used for training, 10% pairs are used for validation, and the others are used for testing.

4.1.3. DUBAI-CC Dataset

The original images in the DUBAI-CC dataset [1] are cropped into 500 tiles of sizes 50 × 50, with five change descriptions annotated for each small bitemporal tile, resulting in a total number of descriptions being 2500. The training, validation, and testing datasets are set, respectively, as 300, 50, and 150 bitemporal tiles. The cropped images are upsampled to a size of 256 × 256 in the experiments.

4.2. Experimental Setup

4.2.1. Implementation Details

The deep-learning framework PyTorch is utilized to implement the models. All models are trained and assessed on a desktop with an NVIDIA GTX 4060Ti GPU. The Adam optimizer [46] is adopted for training, where the learning rate is initialized as 0.0001. We consider setting the learning rate to decrease by a factor of 0.7 as the training step increases by three epochs. We set the maximum epoch as 40. The beam search size is fixed as 3 when generating change captions.

4.2.2. Evaluation Metrics

The evaluation indicators of the image sentence-description generation tasks depend on whether the generated descriptive sentences are in accordance with human judgments (i.e., similar to the labeled sentences). Here, we adopt four metrics to measure the change captioning performance, i.e., BLEU-N (N = 1, 2, 3, 4) [47], ROUGE-L [48], METEOR [49], and CIDEr-D [50]. These metrics are very popular in RSICC tasks. Higher scores indicate better quality of the produced sentences.

4.3. Comparison to SOTA Methods

Here, the LEVIR-CC, LEVIRCCD, and DUBAI-CC datasets are adopted to compare the performance of the proposed method with several SOTA methods, including Capt-Dual-Att [36], DUDA [36], MCCFormer-S [39], MCCFormer-D [39], RSICCFormer [13], Chg2Cap [14]. A concise introduction to these benchmark approaches is given below.

- Capt-Dual-Att [36]: In this method, the bitemporal spatial attention maps are learned using two convolutional layers. The obtained feature maps are then differentiated and fed into an LSTM to generate descriptions of changes.
- DUDA [36]: DUDA studies a dual dynamic attention model based on two LSTMs in which an LSTM utilizes visual and textual features to generate attention weights and then input the feature representations into another LSTM for captioning.
- MCCFormer-S [39]: It is a Transformer-based change captioning approach. After linear transforming, position encoding and flattening of the extracted feature maps, one concatenates the outputs to serve as the sole input of the Transformer encoder and decoder for captioning.
- MCCFormer-D [39]: It utilizes a Siamese Transformer framework where the cross-attention mechanism is employed to capture differences between the bitemporal feature maps; then, they are sent to the Transformer decoder to produce the change captions.
- RSICCFormer [13]: RSICCFormer develops a dual-branch Transformer framework, in which multiple Transformer layers equiped with cross-attention are cascaded to gradually extract and utilize the differences in images to recognize changes.
- Chg2Cap [14]: Chg2Cap makes use of an attentive encoder including two blocks. One is a hierarchical self-attention block utilized to represent change-related features, and the other is a residual block adopted to produce the image embedding. The change description generator is based on Transformer.

Table 1 presents the performance of the developed MADiffCC model compared with SOTA methods based on the LEVIR-CC dataset. It is easy to observe that our developed method exhibits better scores, resulting in an improved performance of 3.6% on BLEU-4 and 4.1% on CIDEr-D compared with the best-performing Chg2Cap method. In addition, Tables 2 and 3, respectively, list the results of the compared methods based on LEVIR-CCD and DUBAI-CC datasets. Note that for simplicity, here, we choose the methods which achieve SOTA performance on the LEVIR-CC dataset for comparison. It is seen in Tables 2 and 3 that, compared with the benchmark methods, our developed approach leads to better performance. The substantial performance improvement in these datasets could be attributed to the ability of MADiffCC, which utilizes more robust and generalized feature representations extracted from RS image dataset pre-trained DDPM, and finds out the change-aware discriminative information while ignoring irrelevant ones by the multi-attentive network.

Table 1. Comparison of the developed method and other SOTA methods on the LEVIR-CC dataset. Higher score indicates better change captioning performance. The best scores are in bold.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Capt-Dual-Att [36]	79.51	70.57	63.23	57.46	36.56	70.69	124.42
DUDA [36]	81.44	72.22	64.24	57.79	37.15	71.04	124.32
MCCFormers-S [39]	82.16	72.95	65.42	59.68	38.17	72.46	128.39
MCCFormers-D [39]	80.42	71.87	63.86	57.38	38.29	71.32	126.44
RSICCFormer [13]	84.72	75.27	68.87	62.77	39.61	74.12	133.12
Chg2Cap [14]	86.14	76.08	70.66	63.39	40.03	75.12	134.55
ine Ours	**86.28**	**77.50**	**71.09**	**66.93**	**40.16**	**75.37**	**138.61**

Table 2. Comparison of the developed method and other SOTA methods on the LEVIRCCD dataset. Higher score indicates better change captioning performance. The best scores are in bold.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
MCCFormers-D [39]	66.81	56.89	48.57	41.53	26.16	54.63	78.58
RSICCFormer [13]	69.02	59.78	52.42	46.39	28.18	56.81	80.08
Chg2Cap [14]	72.33	61.45	54.19	47.23	29.87	58.02	83.21
Ours	**72.99**	**63.11**	**56.05**	**49.78**	**30.80**	**58.55**	**85.99**

Table 3. Comparison of the developed method and other SOTA methods on the DUBAI-CC dataset. Higher score indicates better change captioning performance. The best scores are in bold.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
ine MCCFormers-D [39]	64.25	50.31	39.56	29.43	25.22	51.38	66.23
RSICCFormer [13]	67.88	53.06	41.38	30.99	25.50	51.48	66.29
Chg2Cap [14]	71.17	59.26	49.87	41.11	28.51	57.98	91.76
Ours	**73.18**	**61.36**	**52.25**	**45.41**	**30.85**	**60.56**	**96.47**

4.4. Ablation Studies

The numerical results of our ablation study are given here with the aim of evaluating the effectiveness of the modules developed in the MADiffCC model: the diffusion feature extraction module (abbreviated as DFE here), the TCSA module, and the GMCA module. The experiments are carried out on the LEVIR-CC dataset.

First, we investigate the contributions of different sampling locations (i.e., time-step $t \in [0, T]$) in the noise-schedule of the developed diffusion feature extractor to look for the optimal feature representations for the overall performance of the MADiffCC model. The results of our experiments are presented in Table 4, where the change captioning performance varies with the diffusion features extracted at t = 50, 100, (50 and 100), (50, 100 and 400). Specifically, t = (50 and 100) represents the time steps 50 and 100, while t = (50, 100, and 400) stands for the time steps 50, 100, and 400. It is observed that when the feature representations are sampled with t = (50 and 100), we can obtain the best change captioning performance. As a result, the feature representations sampled with t = (50 and 100) are utilized as input to the difference encoder in the MADiffCC model.

Table 4. Ablation studies on the time-step t utilized to extract features from the diffusion feature extractor. Higher score indicates better change captioning performance. The best scores are in bold.

Time-Step t	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
50	85.83	76.35	70.32	65.80	39.41	74.89	136.87
100	85.96	76.42	70.55	66.06	39.57	75.04	137.13
(50, 100)	**86.28**	**77.50**	**71.09**	**66.93**	**40.16**	**75.37**	**138.61**
(50, 100, 400)	86.04	77.31	71.02	66.71	40.01	75.22	138.22

Then, the following methods are designed (optionally including or excluding the three modules) to compare with our proposed model.

(1) Baseline: The image features are extracted by the commonly natural images pre-trained in Resnet-101 [31], the difference map is obtained using a common difference encoding, and the caption decoder uses a plain transformer to generate captions.

(2) MADiffCC†: MADiffCC† employs the RS images pre-trained in DDPM to extract multi-scale features of the input images, which includes multi-level and multi-time-step diffusion feature representations. MADiffCC† has the same difference encoding and the caption decoder as Baseline.

(3) MADiffCC††: MADiffCC†† utilizes the DFE and TCSA modules simultaneously for the RSICC task. A plain transformer is adopted as caption decoder to generate change descriptions.

(4) MADiffCC: Our proposed MADiffCC model.

In addition, we design the experiments by training the model on the same training set and then evaluating its performance on the test set from the following three cases. Case 1: samples only without changes; Case 2: samples only with changes; Case 3: the entire set. It is noted that in the case of the no-change test, the CIDEr-D metric is meaningless since the descriptions of no-change cases are relatively monotonous.

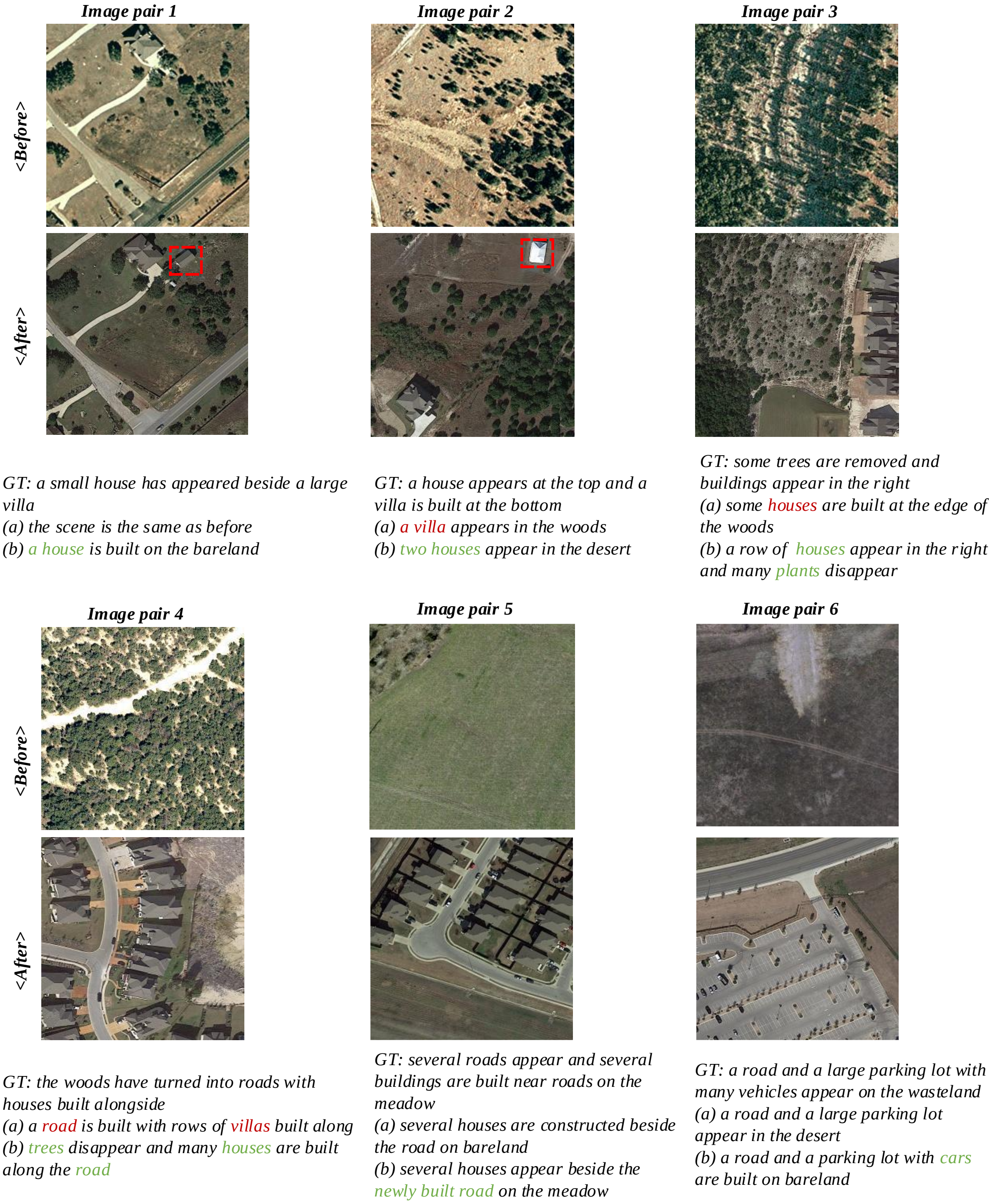
Table 5 reports the quantitative ablation results of different components in the proposed MADiffCC model for three settings. To begin with, we analyze the benefits of the DFE module which extracts feature representations from a DDPM pre-trained on RS images. It is noted that the time-step is set as (50 and 100) here, since its performance is the best as illustrated in Table 4. By observing the scores obtained by Baseline and MADiffCC†, it is obvious that the results of MADiffCC† highlight the advantages of using diffusion features as input over Baseline in all three settings. This shows that the diffusion feature extractor effectively perceives the contextual information between RS images. Furthermore, from the results of MADiffCC††, it is observed that with the inclusion of the TCSA module which applies time-step features fusion layer and channel-spatial attention layers on the extracted diffusion features, higher evaluation performance can be achieved in all three settings. It proves that the TCSA module can find out the change-aware features precisely and filter out the irrelevant changes in bitemporal RS images. In addition, significant enhancement of model performance could be witnessed from the incorporation of the GMCA module, since it makes full use of multi-scale difference feature relationships and assists in selecting essential information for better change captioning. It is concluded from Table 5 that obvious performance gains can be achieved by the developed model in terms of both the change discrimination ability and caption generation performance.

Table 5. Ablation studies on the DFE, TCSA, and GMCA modules using samples only without changes, samples only with changes and the entire test set. × and ✓ respectively indicate that the corresponding module is excluded/included in the model. Higher score indicates better change captioning performance. The best scores are in bold.

Cases	Methods	DFE	TCSA	GMCA	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Case 1	Baseline	×	×	×	91.44	89.61	88.86	88.14	66.20	91.35	-
	MADiffCC†	✓	×	×	93.93	93.11	92.75	92.51	69.23	92.46	-
	MADiffCC††	✓	✓	×	95.81	95.25	95.04	93.81	73.19	96.16	-
	MADiffCC	✓	✓	✓	**97.21**	**97.72**	**96.77**	**96.23**	**76.12**	**97.53**	**-**
Case 2	Baseline	×	×	×	66.62	52.24	38.88	28.67	21.04	45.26	39.05
	MADiffCC†	✓	×	×	72.51	56.92	45.50	37.11	23.84	49.36	58.67
	MADiffCC††	✓	✓	×	75.40	60.91	47.98	38.79	26.13	53.11	62.06
	MADiffCC	✓	✓	✓	**77.08**	**62.63**	**49.85**	**41.53**	**26.61**	**53.62**	**62.33**
Case 3	Baseline	×	×	×	78.22	68.04	61.86	55.33	35.02	70.35	124.14
	MADiffCC†	✓	×	×	81.66	72.29	66.81	60.74	37.15	73.03	130.63
	MADiffCC††	✓	✓	×	84.14	75.87	70.12	65.23	38.64	74.35	134.45
	MADiffCC	✓	✓	✓	**86.28**	**77.50**	**71.09**	**66.93**	**40.16**	**75.37**	**138.61**

4.5. Qualitative Results

To verify the availability of the change captions from our developed MADiffCC framework, a qualitative assessment is conducted via choosing several representative cases in the LEVIR-CC dataset. The change captioning results are as shown in Figure 4, where the ground truth (GT) caption and the descriptions generated by (a) Chg2Cap [14] and (b) the proposed MADiffCC model are presented. Words in green represent more precise and detailed predicted change objects by our method, while the red text indicates an inaccurate representation. It is obvious in Figure 4 that the results achieved by our developed approach are able to describe the RS image pairs more precisely and accurately when compared with the benchmark ones. For example, as shown in Image Pair 1, MADiffCC demonstrates sensitivity to the small object "house" on the bareland, whereas the benchmark method cannot predict it. In addition, as presented in Image Pair 2, our method is able to accurately recognize "two houses" in the desert (marked in green) rather than just "a villa" obtained by the benchmark method (highlighted in red), demonstrating its capability of simultaneously recognizing multiple changes of the objects. Furthermore, it is seen in Image Pair 3 that MADiffCC can provide more informative descriptions shown in (b) compared with the results of the benchmark method given in (a), where the description "many plants disappear" is given besides the "houses". The case in similar in Image Pairs 4, 5 and 6. The results presented in Figure 4 demonstrate that MADiffCC can generate more informative and accurate change descriptions in bitemporal RS images, which may contribute to its abilities of leveraging contextual information between samples captured from diffusion feature extractor, and discriminative multi-scale change information retrieved by multi-attention structures.



GT: a small house has appeared beside a large villa
(a) the scene is the same as before
(b) a house is built on the bareland

GT: a house appears at the top and a villa is built at the bottom
(a) a villa appears in the woods
(b) two houses appear in the desert

GT: some trees are removed and buildings appear in the right
(a) some houses are built at the edge of the woods
(b) a row of houses appear in the right and many plants disappear

GT: the woods have turned into roads with houses built alongside
(a) a road is built with rows of villas built along
(b) trees disappear and many houses are built along the road

GT: several roads appear and several buildings are built near roads on the meadow
(a) several houses are constructed beside the road on bareland
(b) several houses appear beside the newly built road on the meadow

GT: a road and a large parking lot with many vehicles appear on the wasteland
(a) a road and a large parking lot appear in the desert
(b) a road and a parking lot with cars are built on bareland

Figure 4. Qualitative comparison of change captioning results. GT: ground truth caption, (a): Chg2Cap [14], (b): the proposed MADiffCC model. Words marked in green stand for more precise and detailed predicted change objects by the proposed method, while the red text indicates inaccurate representations. The red boxes in image pairs 1 and 2 indicate the small object undetected by benchmark method.

5. Conclusions

In this article, a novel framework for RSICC by exploiting a multi-attentive network integrated with a diffusion model (MADiffCC) is proposed. The framework consists of a diffusion feature extractor, a time-channel-spatial attention (TCSA)-based difference encoder, and a gated multi-head cross-attention (GMCA)-guided linguistic decoder. The diffusion feature extractor based on a diffusion model pre-trained with an RS image dataset is able to recover the data distribution and contextual information of targets in RS images, leading to more robust and generalized feature representations with a hierarchical form. In addition, the multi-attention structures, i.e., TCSA and GMCA mechanisms used in the encoder–decoder pattern, can enhance the capability of obtaining inherent most change-aware discriminative information for more precise change description generation. Extensive experiments are carried out to corroborate the effectiveness of our devised RSICC model. Through the experimental results, it is validated that the developed approach is effective in generating more informative and accurate change descriptions in bitemporal RS images.

Author Contributions: Conceptualization, Y.Y.; Methodology, Y.Y.; Software, Y.Y., T.L., Y.P. and L.L.; Validation, Y.Y., T.L., Y.P. and L.L.; Formal analysis, Y.Y.; Investigation, Y.Y., T.L., Y.P. and L.L.; Writing—original draft, Y.Y.; Writing—review & editing, Y.Y., T.L., Y.P. and L.L.; Visualization, Y.Y.; Supervision, Q.Z. and Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: The authors thank the anonymous reviewers for providing valuable comments and suggestions that have helped improve the quality of the paper. This work is supported in part by the National Natural Science Foundation of China under Grant 62301464 and Sichuan Science and Technology Program under Grant 2024NSFSC1427.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Chouaf, S.; Hoxha, G.; Smara, Y.; Melgani, F. Captioning Changes in Bi-Temporal Remote Sensing Images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2891–2894.
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. Wang, Q.; Huang, W.; Zhang, X.; Li, X. GLCM: Global–Local Captioning Model for Remote Sensing Image Captioning. *IEEE Trans. Cybern.* **2023**, *53*, 6910–6922. [CrossRef] [PubMed]
4. Xu, Y.; Yu, W.; Ghamisi, P.; Kopp, M.; Hochreiter, S. Txt2Img-MHN: Remote Sensing Image Generation From Text Using Modern Hopfield Networks. *IEEE Trans. Image Process.* **2023**, *32*, 5737–5750. [CrossRef]
5. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
6. Feng, H.; Zhang, L.; Yang, X.; Liu, Z. Enhancing class-incremental object detection in remote sensing through instance-aware distillation. *Neurocomputing* **2024**, *583*, 127552. [CrossRef]
7. Pang, S.; Lan, J.; Zuo, Z.; Chen, J. SFGT-CD: Semantic Feature-Guided Building Change Detection From Bitemporal Remote-Sensing Images with Transformers. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2500405. [CrossRef]
8. Li, Z.; Cao, S.; Deng, J.; Wu, F.; Wang, R.; Luo, J.; Peng, Z. STADE-CDNet: Spatial–Temporal Attention with Difference Enhancement-Based Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611617. [CrossRef]
9. Chen, L.; Liu, C.; Chang, F.; Li, S.; Nie, Z. Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery. *Neurocomputing* **2021**, *451*, 67–80. [CrossRef]
10. Zhao, R.; Shi, Z.; Zou, Z. High-Resolution Remote Sensing Image Captioning Based on Structured Attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603814. [CrossRef]
11. Wang, Y.; Zhang, W.; Zhang, Z.; Gao, X.; Sun, X. Multi-scale Multi-interaction Network for Remote Sensing Image Captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2154–2165. [CrossRef]
12. Zhuang, S.; Wang, P.; Wang, G.; Wang, D.; Chen, J.; Gao, F. Improving Remote Sensing Image Captioning by Combining Grid Features and Transformer. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6504905. [CrossRef]
13. 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]
14. Chang, S.; Ghamisi, P. Changes to Captions: An Attentive Network for Remote Sensing Change Captioning. *IEEE Trans. Image Process.* **2023**, *32*, 6047–6060. [CrossRef] [PubMed]

15. Liu, C.; Yang, J.; Qi, Z.; Zou, Z.; Shi, Z. Progressive Scale-Aware Network for Remote Sensing Image Change Captioning. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6668–6671.
16. Cai, C.; Wang, Y.; Yap, K.H. Interactive change-aware transformer network for remote sensing image change captioning. *Remote Sens.* **2023**, *15*, 5611. [CrossRef]
17. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
18. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10674–10685.
19. Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **2022**, *479*, 47–59. [CrossRef]
20. Brempong, E.A.; Kornblith, S.; Chen, T.; Parmar, N.; Minderer, M.; Norouzi, M. Denoising Pretraining for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 21–24 June 2022; pp. 4174–4185.
21. Lei, J.; Tang, J.; Jia, K. RGBD2: Generative Scene Synthesis via Incremental View Inpainting Using RGBD Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 8422–8434.
22. Kim, M.; Liu, F.; Jain, A.; Liu, X. DCFace: Synthetic Face Generation with Dual Condition Diffusion Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12715–12725.
23. Bandara, W.G.C.; Nair, N.G.; Patel, V.M. DDPM-CD: Denoising Diffusion Probabilistic Models as Feature Extractors for Change Detection. *arXiv* **2024**, arXiv:2206.11892.
24. Zhang, X.; Tian, S.; Wang, G.; Zhou, H.; Jiao, L. DiffUCD: Unsupervised hyperspectral image change detection with semantic correlation diffusion model. *arXiv* **2023**, arXiv:2305.12410.
25. Wen, Y.; Ma, X.; Zhang, X.; Pun, M.O. GCD-DDPM: A generative change detection model based on difference-feature guided DDPM. *arXiv* **2023**, arXiv:2306.03424. [CrossRef]
26. Yang, X.; Wang, X. Diffusion model as representation learner. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18938–18949.
27. Yan, C.; Zhang, S.; Liu, Y. Feature prediction diffusion model for video anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 5527–5537.
28. Fuest, M.; Ma, P.; Gui, M. Diffusion models and representation learning: A survey. *arXiv* **2024**, arXiv:2407.00783.
29. 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 (ICCV), Paris, France, 2–3 October 2023; pp. 19773–19786.
30. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
31. 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, 26–30 June 2016; pp. 27–30.
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9. [CrossRef]
33. Zhou, B.; Zhao, H.; Fernandez, F.X.P.; Fidler, S.; Torralba, A. Scene parsing through ADE20K dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017.
34. 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]
35. 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]
36. 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; pp. 4623–4632.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, 30, 1–11.
38. Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; Shi, Z. A Decoupling Paradigm with Prompt Learning for Remote Sensing Image Change Captioning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5622018. [CrossRef]
39. 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, QC, Canada, 11–17 October 2021; pp. 1951–1960.
40. Guo, J.; Li, Z.; Song, B.; Chi, Y. TSFE: Two-Stage Feature Enhancement for Remote Sensing Image Captioning. *Remote Sens.* **2024**, *16*, 1843. [CrossRef]
41. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. In Proceedings of the Proc. NIPS, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 1–13.

42. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. In Proceedings of the Proc. ICLR, Virtual, 3–7 May 2021; pp. 1–36.
43. Bao, F.; Li, C.; Cao, Y.; Zhu, J. All are worth words: A vit backbone for score-based diffusion models. *arXiv* **2022**, arXiv:2209.12152.
44. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ. Interdiscip. J.* **2012**, *120*, 25–36. [CrossRef]
45. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 421–429.
46. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
47. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
48. Lin, C. 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.
49. 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.
50. Vedantam, R.; Zitnick, C.L.; 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; pp. 4566–4575.

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.