Single-stream Extractor Network with Contrastive Pre-training for Remote Sensing Change Captioning

Qing Zhou, Junyu Gao, *Member, IEEE*, Yuan Yuan, *Senior Member, IEEE* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Remote sensing (RS) image change captioning is a visual semantic understanding task that has received increasing attention. The change captioning methods are required to understand the visual information of the images and capture the most significant difference between them, then describe it in natural language. Most existing methods mainly focus on improving the difference feature encoder or language decoder, while ignoring the visual feature extractor. The current feature extractors suffer from several issues, including 1) domain gap between pre-training on single temporal natural images and downstream bi-temporal RS task, 2) limited difference feature modeling in the implicit single-stream network, and 3) high computational costs caused by extracting features for each temporal phase image under the dual-stream extractor. To address these issues, we propose a Single-stream Extractor Network (SEN). It consists of a single-stream extractor pre-trained on bi-temporal RS images using contrastive learning to mitigate the domain gap and high computational cost. Additionally, to improve feature modeling for difference information, we propose a shallow feature embedding (SFE) module and a cross attention guided difference (CAGD) module, which enhance the representation of temporal features and extract the difference features explicitly. Extensive experiments and visualizations demonstrate the effectiveness and advanced performance of SEN. The code and model weights are available at https://github.com/mrazhou/SEN.**

***Index Terms*—Change captioning, remote sensing images, contrastive pre-training, single-stream.**

I. Introduction

THE earth is constantly undergoing changes, which are closely related to human activities such as urbanization [1], [2] and natural events like natural disasters [3]. With the advancement of earth observation technologies, the availability of multi-temporal remote sensing (RS) images that capture these changes is increasing. Automated perception and understanding of these changes hold significant importance for human production and livelihood, including urban planning [1], [4], resource management [5], and disaster monitoring [6]. Compared to traditional pixel-level change detection methods [7]–[11], change captioning involves interpreting the changes between two images at a semantic level. These approaches can describe the change areas in the images as well as the attributes and relationships of objects within them using natural language, making the output more easily understandable for humans. Therefore, remote sensing image change captioning (RSICC) has gained attention in recent years [12]–[15].

Existing change captioning methods are almost based on the encoder-decoder paradigm. This paradigm couples visual feature extraction and difference feature encoding, blurring the boundaries between the two. Therefore, we divide the encoder-decoder paradigm into three parts: visual feature extractor, difference feature encoder and natural language caption decoder. As shown in Figure 1 (a), most existing methods employ a shared dual-stream backbone network as the visual feature extractor. The backbone network [16], [17] is initialized by parameters pre-trained on ImageNet [18] and kept frozen during downstream RSICC tasks to reduce the number of trainable parameters. To better extract discriminative difference features related to changes, several methods have focused on improving the difference feature encoder, such as DUBA [19], PSNet [13], and even unsupervised difference encoding [20], [21]. Additionally, some methods [12], [19], [22] have made improvements from the perspective of the natural language decoder, enhancing the interaction between textual and visual features, using techniques such as attention mechanisms, transformer decoders [23], and introducing reinforcement learning [24], [25]. However, these methods overlook the research on the visual feature extractor, which profoundly influences the performance of RSICC. As depicted in Figure 1(b), Hoxha *et al.* [15] constructed an implicit single-stream network based on image subtraction and explored the effects of single-stream and dual-stream network structures on RSICC performance. Nevertheless, existing methods still face the following issues: 1) **domain gap**. It includes two gaps between pre-training and downstream task. The first one is the data distribution gap between ImageNet's natural images and RS images. The second one is the input gap between using a single RGB image as input and using the difference of bi-temporal RGB images as input. Using the difference between two images as input would lose a lot of scene information about the image as a whole, especially for image pairs that change dramatically, as shown in Figure 1(b). 2) **Limited difference modeling**. The implicit single-stream network explored by Hoxha *et al.* [15] may limit the network's ability to effectively capture and utilize discriminative difference features, which results in suboptimal representation [26]–[28]. Furthermore, the absence of individual temporal phase features hinders explicit modeling and analysis of the differences between bi-temporal images [29]. 3) **High computational cost**. Although most existing methods use shared dual-stream networks to minimize parameter count, the computational overhead of the dual-stream structure is still nearly twice that of the single-

This work was supported by the National Natural Science Foundation of China under Grant U21B2041 and Natural Science Foundation of China under Grant 62306241.

All authors are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China (e-mail: chautsing@gmail.com, gjy3035@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com). (Corresponding author: Qi Wang.)

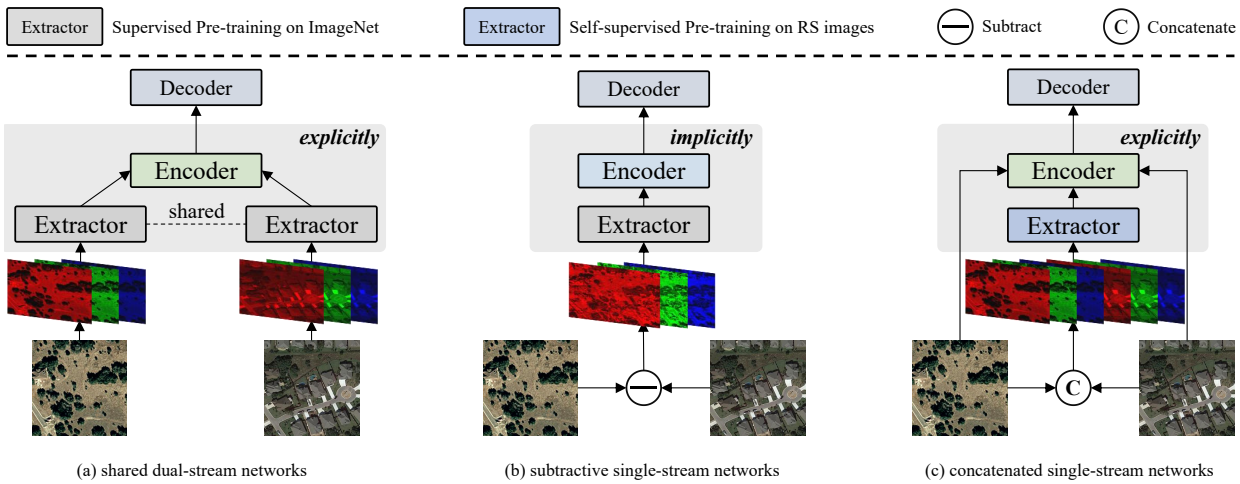


Fig. 1. Different architectures of RSICC models. (a) Shared dual-stream network with explicit difference modeling. (b) Difference-based single-stream network with implicit difference modeling. (c) Concatenation-based single-stream network with explicit difference modeling. The visual feature extractor of (a) and (b) is pre-trained on natural scenes images, and the visual feature extractor of (c) is pre-trained on RS images.

stream network due to the separate visual feature extraction for each image.

To handle these issues, we propose a Single-stream Extractor Network (SEN) based on bi-temporal contrastive pre-training, as shown in Figure 1(c). Compared to the shared dual-stream networks in Figure 1(a), SEN utilizes a single-stream feature extractor, which significantly reduces computational costs. In contrast to the implicit difference modeling networks in Figure 1(b), SEN's explicit encoder can better capture differential information. Importantly, the extractor in SEN is pre-trained based on self-supervised learning using bi-temporal concatenated RS images. This ensures input consistency and avoids the influence of domain gaps on feature extraction. Specifically, the overall framework of SEN is illustrated in Figure 2. Firstly, SEN employs contrastive learning to pre-train the single-stream visual feature extractor on numerous unlabeled multi-temporal RS image pairs. Secondly, to model and enhance the difference features at the high-level in the single-stream feature extraction network, we introduce a shallow feature embedding module (SFE) and a cross attention guided difference module (CAGD) to form an explicit difference feature encoder. The SFE module initially embeds the shallow features of each temporal phase using channel attention. This module enriches the low-level image information of the single temporal phase features, capturing visual details from each individual image. Then, the CAGD module enhances the feature expression of each temporal phase and models the difference information. By incorporating subtraction and self attention mechanisms, the CAGD module explicitly models the difference information. This facilitates the network in effectively capturing and representing the discriminative difference information between the bi-temporal images. The combination of the SFE module and the CAGD module enables the network to explicitly model and utilize the difference information, enhancing the representation and performance of the single-stream visual feature extractor for RSICC tasks.

In summary, our contributions can be outlined as follows:

1) We propose a single-stream extractor for RSICC, which is pre-trained on concatenated RS images in temporal using contrastive learning. The extractor has a lower computational cost than the dual-stream extractor, and a smaller domain gap between pre-training and downstream task.
2) We propose SFE and CAGD modules to form an explicit difference feature encoder to better model the difference information in the single-stream extractor.
3) We conducted thorough experiments on the RSICC datasets to validate the efficacy of the proposed SEN method and attain advanced performance.

II. Related Works

In this section, we offer a concise review of change captioning in both natural images and RS images, and pre-training on large-scale RS images.

A. Change Captioning

Change captioning aims to generate a natural language description of the changes observed between two images. The generated captions should be accurate and concise, and consistent with human perception. As a high-level semantic understanding task, change captioning has attracted increasing attention and achieved promising development in recent years. Jhamtani *et al.* [30] first proposed the concept of change captioning and constructed a dataset derived from video surveillance footage. Park *et al.* [19] collected a CLEVR-Change dataset to explore diverse change types and model robustness when presented with distracting information. Qiu *et al.* [31] proposed a dataset depicting multiple changes and a transformer-based captioning model to describe and locate multiple changes. To reduce the interference caused by illumination, viewpoint and other factors, and improve focus on significant changes, many methods focus on the difference feature encoder. DUBA [19] utilized a dual dynamic attention model to learn meaningful semantic difference features from changes involving illumination and viewpoint distractions, and

generated accurate captions by a dynamic speaker component, adaptively focusing attention on visual features. Shi *et al.* [22] focused on handling viewpoint difference interference using a viewpoint-adapted matching encoder that computes the similarity between image pairs in the feature spaces. Tu *et al.* [32] proposed a semantic difference representation learning network that captures the tiny difference ignored by DUBA via self-semantic relation embeddings of object features.

In order to cope with limitations in RS change detection, where specialists may still be required to interpret the binary map, Chouaf *et al.* [14] first constructed a private RSICC dataset based on a building change detection dataset and proposed a model that generates captions by RNN conditioned on the concatenated feature extracted by a shared pre-trained VGG16. Hoxha *et al.* [15] further proposed two small RSICC datasets. They also designed two strategies to encode discriminative features, one based on image subtraction and the other based on feature concatenation, and the latter achieved better performance. Based on discriminative features, two decoders RNN and support vector machine (SVM), were used to generate captions. Recently, Liu *et al.* [12] released a larger RSICC dataset and proposed a dual-stream transformer-based architecture to effectively generate captions. To adequately extract and exploit multi-scale difference information, they [13] proposed a progressive difference perception (PDP) layer and a scale-aware reinforcement (SR) module. Furthermore, they also proposed decoupling paradigm [26], multi-scale aggregation methods [33] and pseudo-label learning [34] to improve caption generation performance.

However, most of these methods focus on improving the difference feature encoder so that it can deal with the interference of irrelevant pseudo changes more robustly and pay attention to meaningful changes. Less attention is given to the single-stream extractor. Despite Hoxha *et al.* [15] studied the single-stream extractor, the potential of the single-stream extractor has not been fully explored, due to the domain gap and implicit difference modeling. The domain gap between natural images and RS images is also a challenge faced by other methods that are based on a two-stream extractor. In section B, we will further introduce this problem and the existing solutions.

B. Pre-training on Remote Sensing Images

RS images exhibit a sizable domain gap with when compared natural images. Factors such as imaging distance, perspective, environmental conditions introduce variances that conventional models trained predominantly on general imagery struggle to overcome. These variances limit the ability of fine-tune the ImageNet pre-trained extractor on RS images [35]. Wang *et al.* [35] demonstrated that pre-training on RS images could mitigate the domain gap with an empirical study on a large-scale RS scene recognition dataset [36]. Before [35], Sumbul *et al.* [37] presented a large-scale multi-label classification datasets as a training source for pre-training on RS images. Although supervised pre-training based on classification RS datasets has made significant progress, it comes with the challenge of annotating numerous labels, which incur substantial labor and time costs. Self-supervised pre-training methods can leverage unlabeled datasets to eliminate the cost of manual annotation while promoting the performance of models on downstream tasks with similar data distributions. These methods include contrastive learning [38] using positive and negative sample pairs, teacher-student distillation learning [39], and contrasting cluster assignments approach [40]. Among them, contrastive learning with positive and negative sample pairs constructed through data augmentation is one of the popular and classic methods. SeCo [41], for instance, extensively explores and utilizes a large amount of unlabeled remote sensing data based on contrastive learning methods. They gathered a large-scale unlabeled multi-temporal RS dataset and pre-trained a ResNet on it with seasonal contrastive learning to learn generalizable representation for RS images. Due to the limitations of sparse geographical distribution, task-specific, etc., Wang *et al.* [42] shared SSL4EO, a comprehensive RS dataset spanning multiple modalities, time periods, and global coverage. They conducted extensive experiments on a set of typical self-supervised pre-training methods and downstream RS tasks to validate that pre-training on SSL4EO can yield more generalized and adaptive representations for RS images. Notably, the ResNet18 model pre-trained on SSL4EO using the MoCo-v2 method achieved excellent performance in RS change detection. Considering the similarities between RS change detection and RSICC, we further investigate the performance of a self-supervised pre-trained single-stream visual feature extractor on RSICC in this paper, and attempt to uncover and showcase the potential of the single-stream extractor.

III. Methodology

A. Overview

The proposed framework involves two steps and three components, as shown in Figure 2. The two steps involve bi-temporal image pre-training on large-scale multi-temporal RS images via contrastive learning, followed by transfer learning on the RSICC. The three components include a pre-trained single-stream visual extractor, a difference feature encoder composed of SFE and CAGD, and a transformer-based natural language caption decoder.

Specifically, the temporal contrastive pre-training randomly selects and concatenates two images from different seasons of the same location to form a bi-temporal image I. The bi-temporal image I undergoes image transformations to generate different views for contrastive learning, training a single-stream visual extractor f_q. The pre-trained visual extractor f_q is then applied to the RSICC task, deriving the concatenated bi-temporal image's visual characteristics F. The raw images' rich and complete visual information is fused with the extracted visual features F using the SFE module, comprising a convolution layer and a channel attention component, resulting in single-temporal features F_{bef} and F_{aft}. The CAGD module then extracts the difference feature from F_{bef} and F_{aft} through cross attention and self attention, and then generates the change captions S_p using a transformer-based decoder. SEN is optimized through minimization of the cross-entropy (CE) loss between the generated change captions S_p and the target text descriptions S_t.

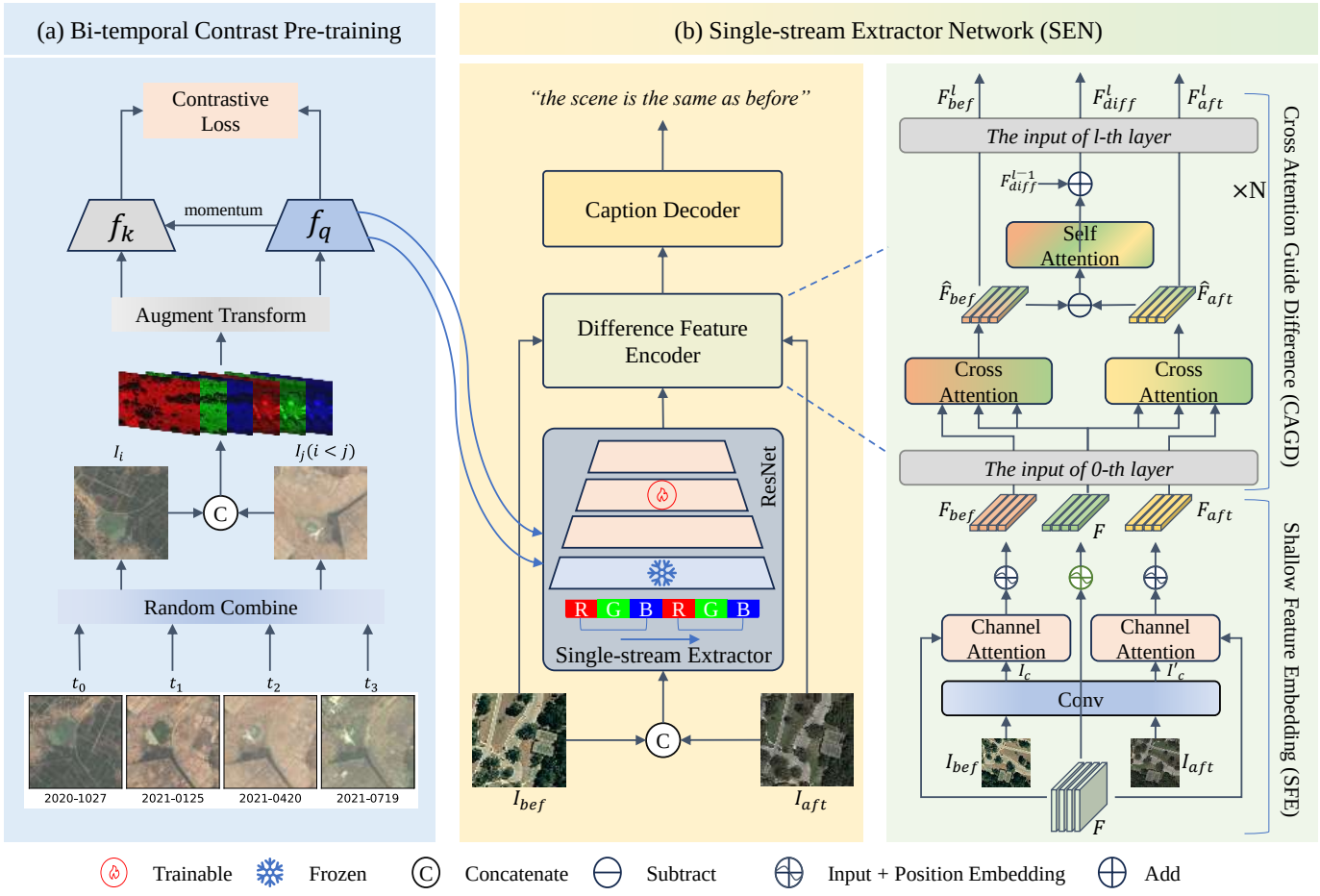


Fig. 2. Overall framework of our proposed SEN method. It mainly consists of two steps (a) bi-temporal contrastive pre-training and (b) transfer learning on the RSICC with SEN. The SEN includes a single-stream visual extractor, a difference feature encoder composed of SFE and CAGD, and a transformer-based natural language caption decoder.

B. Bi-temporal Contrastive Pre-training

The overall process of temporal contrastive pre-training is shown in Figure 2 (a). Specifically, given a set of RS images $I_{t_i} \in R^{H\times W\times 3}$ captured at the same location during different seasons $t_i \in \{t_1, t_2, t_3, t_4\}$, we randomly select two images I_{t_i} and I_{t_j}, where $i, j = 1, 2, 3, 4$, and concatenate them in channel to obtain a six-channel bi-temporal image $I_{i,j} = Concat(I_{t_i}, I_{t_j}) \in R^{H\times W\times 6}$. Let Γ represent a set of commonly used image argument transformations [38], [43], such as random flipping, random brightness, and random contrast. We randomly select two six-channel bi-temporal images I_{i_q,j_q} and I_{i_k,j_k}, then apply random augmentation operations to generate query and key images, $q = \Gamma(I_{i_q,j_q})$ and $k = \Gamma(I_{i_k,j_k})$.

Positive samples are formed by pairing q with corresponding k^+, where both images are from the same location but captured at random seasons. Negative samples are formed by pairing q with k^- that do not correspond to the same location, and the seasons are chosen randomly. For positive and negative samples, they come from images with different locations and random combinations of seasons. So, the locations are always different, while the combinations of seasons could potentially be the same.

The single-stream visual feature extractor f_q (detailed in section C, same structure as single-stream extractor) is optimized via a contrastive loss, InfoNCE [44]. The loss is calculated as follows:

$$L_c = -log\frac{\exp(f_q(q)\cdot f_k(k^+)/\tau)}{\sum_k \exp(f_q(q)\cdot f_k(k)/\tau)}, \tag{1}$$

where τ is the temperature [45].

Regarding the relationship between i and j, we experiment with two strategies: ordered and unordered. The ordered strategy is defined as $0 \le i \le j \le 4$, where the older images are placed in the first three channels and newer images in the last three channels. This strategy aligns with the bi-temporal images used in downstream RSICC tasks. The unordered strategy is defined as $0 \le i, j \le 4$, where the temporal order is not considered during image concatenation. Since SSL4EO-S12 encompasses only a single year without forming a complete four-season cycle, the unordered strategy can simulate the cyclic nature of the four seasons in the real world (where autumn may occur before summer). This temporal contrastive pre-training enables f_q to learn generic visual features that are invariant to seasonal variations for the bi-temporal RS images I_{i_q,j_q} and reduce the impact of interference factors

such as brightness, contrast, color saturation and atmospheric conditions caused by seasonal changes.

C. Single-stream Extractor Network

1) Single-stream Extractor: The visual feature extractor is a single-stream ResNet network f_e initialized with the pre-trained f_q. Compared to the original ResNet architecture, the single-stream ResNet for RSICC has six input channels in the input convolutional layer to handle the concatenated bi-temporal RS images, as shown on the left side of Figure 2(b). Given a pair of change images (I_{bef}, I_{aft}), where I_{bef} represents the image before the change and I_{aft} represents the image after the change, the concatenated bi-temporal RS image $I_{bef,aft} = Concat(I_{aft}, I_{bef})$ is extracted by the single-stream extractor f_e and the extracted visual features $F \in R^{h\times w\times D}(h = \frac{H}{32}, w = \frac{W}{32})$ are formulated as follows:

$$I_{bef,aft} = Concat(I_{aft}, I_{bef}), \tag{2}$$

$$F = P(f_e(I_{bef,aft}; \theta_f, \theta_t)), \tag{3}$$

where θ_f and θ_t denote the frozen parameters and trainable parameters of f_e. P denotes a convolutional layer that projects the extracted features to a targeted embedding dimension D, adapting them for further processing by the difference feature encoder. For the parameters freezing, we conducted experiments by freezing the parameters in f_e according to the four stages of ResNet. Additionally, we also performed experiments to determine the optimal embedding dimension.

2) Difference Feature Encoder: To explicitly model the content changes between two images conditioned on the single-stream feature extractor, we designed two modules: SFE (Shallow Feature Encoder) and CAGD (Cross Attention Guided Difference) as illustrated in Figure 2(b) on the right.

The SFE module contains a convolution layer and two channel attention blocks. The convolution layer serves to derive preliminary visual characteristics from the raw images, shared across both temporal instances to extract common low-level features. The two channel attention [46] blocks are separately applied to merge each temporal shallow features with the high-level bi-temporal visual features. Due to the visual features F are extracted from the concatenated bi-temporal images, the channel attention enables us to focus on the channel features that are most correlated with each temporal image. This allows us to obtain the single-temporal features $F_{bef}, F_{aft} \in R^{h\times w\times D}$. The formal calculation process is as follows:

$$F_m = CA(Conv_s(I_m, F; \theta_s); \theta_{c\_m}), m \in \{bef, aft\}, \tag{4}$$

where $Conv_s$ is a shared convolutional layer, θ_s and θ_{c_m} denote the parameters of convolution layer and channel attention blocks. CA represents channel attention as follows:

$$\begin{aligned} CA(I_c, F; \theta) =& \sigma(MLP(AP(I_c)) + \\ & MLP(MP(I_c))) \times F, \end{aligned} \tag{5}$$

where AP and MP are average and max pooling layers, respectively. σ is the sigmoid activation function, and MLP is a multi-layer perceptron. I_c is the output of the shared convolutional layer. Then F_{bef}, F_{aft} are reshaped as two-dimensional matrices $R^{h\times w\times D}$ and added the position embedding for difference feature encoding.

The CAGD module is responsible for encoding the differences between the two single-temporal features, capturing the semantic visual features related to the changes. It comprises N_e layers of cross attention (multi-head attention and feed-forward network) and self attention. In cross attention, the single-temporal feature F_m acts as the query, interacting with the bi-temporal features F. This allows the single-temporal features F_m to merge high-level semantic concepts, enhancing the query's perception of visual patterns and relationships across time. The process is formalized as follows:

$$\begin{aligned} \hat{F}_m &= CrossAtt(F_m, F) \\ &= LN(FFN(F_{mha}) + F_{mha}), \end{aligned} \tag{6}$$

where $\hat{F}_m$ is the output of cross attention, FFN is a fully connected layer and LN is layer normalization. F_{mha} is calculated as follows:

$$F_{mha} = LN(MHA(F_m, F, F) + F_m), \tag{7}$$

$$MHA(Q, K, V) = Concat(h_1, ..., h_n)W^O, \tag{8}$$

$$h_i = softmax(A_i)VW^{V_i}, \tag{9}$$

where W^O and W^{V_i} are trainable weights, and W^{V_i} representing the weight allotted to the i-th head when numerous perspectives are taken into account simultaneously. Q, K and V are query, key and value matrices, respectively. A is the attention matrix, calculated by dot product as follows:

$$A_i = \frac{QW^{Q_i}(KW^{K_i})^T}{\sqrt{d_k}}, \tag{10}$$

where W^{Q_i} and W^{K_i} are trainable weights of the i-th head, and d_k is the dimension of the key. T is the transpose operation.

After obtaining the single-temporal features $\hat{F}_{bef}, \hat{F}_{aft}$ enhanced by cross attention, the difference feature $\hat{F}_{diff} = \hat{F}_{aft} - \hat{F}_{bef}$ is computed. Then, self attention is applied to highlight the salient elements of the distinguishing characteristic, and can be formalized as such:

$$\begin{aligned} F_{diff} &= SelfAtt(\hat{F}_{diff}) \\ &= LN(MHA(\hat{F}_{diff}, \hat{F}_{diff}, \hat{F}_{diff}) + \hat{F}_{diff}), \end{aligned} \tag{11}$$

Therefore, the complete calculation of the l-th layer of CAGD is outlined as follows:

$$\begin{aligned} CAGD^l = & SelfAtt^l(CrossAtt^l(F^l_{aft}, F^l_{diff}) \\ & - CrossAtt^l(F^l_{bef}, F^l_{diff})) + F^l_{diff}, \end{aligned} \tag{12}$$

$$F^l_m = \begin{cases} F_m, & \text{if } l = 1 \\ CrossAtt^l(F^{l-1}_m, F^{l-1}_{diff}), & \text{if } l \geq 1 \end{cases}, m \in \{bef, aft\}, \tag{13}$$

$$F^l_{diff} = \begin{cases} F, & \text{if } l = 1 \\ CAGD^l(F^{l-1}_{bef}, F^{l-1}_{aft}, F^{l-1}_{diff}), & \text{if } l \geq 1 \end{cases}. \tag{14}$$

3) Caption Decoder: To generate natural language descriptions accurately conditioned on the difference feature $F_{diff}^{N_e}$, Transformer [23] is utilized as the caption decoder. The decoder consists of N_d layers, each layer comprises masked self attention and cross attention. Given the difference feature $F_{diff}^{N_e}$, the target description S_t, we apply the word embedding and position embedding to S_t to obtain word embedding representation $S_e \in R^{n \times D}$. Then, the l-th layer of the decoder can be expressed as follows:

$$Decoder^l = CrossAtt^l(MaskAtt^l(S^{l-1}), F_{diff}^{N_e}), \quad (15)$$

$$S^l = \begin{cases} S_e, & \text{if } l = 1 \\ Decoder^l(S^{l-1}), & \text{if } l \geq 1 \end{cases}, l \in \{1, ..., N_d\}, \quad (16)$$

where $MaskAtt^l$ is similar to self attention, but the attention matrix A is multiplied by a mask to avoid leakage of the words at the next position [23].

The word representations S^{N_d} output from the final layer of the decoder are fed into a linear layer and softmax function to predict $S_p = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_n] \in R^{n \times d_{vocab}}$, the probability distribution of words in the vocabulary. n represents the length of predicted descriptions and d_{vocab} signifies the size of the vocabulary. The CE loss between the predicted description and the target description is used to optimize the model:

$$L = -\sum_{i=1}^{n} \sum_{j=1}^{d_{vocab}} s_{i,j} log(\hat{s}_{i,j}), \quad (17)$$

where $s_{i,j} \in \{0, 1\}$ denotes the one-hot encoded value of the target caption at time step i and word j. $\hat{s}_{i,j} \in (0, 1)$ denotes the predicted probability of the j-th word at time step i.

IV. Experiments

A. Datasets

1) SSL4EO: The SSL4EO dataset [42] provides a comprehensive and multi-faceted collection of unlabeled remote sensing data well-suited for self-supervised learning applications in earth observation, spanning multiple modalities, time periods, and global locations. This diverse dataset was created by accessing synthetic aperture radar (SAR) and optical satellite images. To provide global coverage and capture seasonal variations, it samples 251,079 locations from 10,000 most populated cities in the world, with each location collecting four images drawn from four seasons. To enhance the data modalities, it exploits ESA's Sentinel-1 and Sentinel-2 satellites to collect SAR and multi-spectral optical images, respectively. Through tapping this rich trove of satellite imagery, the dataset amassed over 3 million images from 250k locations across the globe, with two modalities and four seasons. Given the similarity between RSICC and RS change detection, we followed their experimental setup in the change detection task. Specifically, we utilized the RGB channels and employed the MoCo-v2 method to pre-trian the ResNet feature extractor.

2) LEVIR-CC: The LEVIR-CC dataset [12] comprises 10,077 pairs of bi-temporal RS images, with each image having a spatial size of 256×256 pixels and a resolution of 0.5 m/pixels. To each pair, five descriptive sentences were attached, resulting in a comprehensive corpus of 50,385 annotations the RSICC task. The image pairs originate from the LEVIR-CD dataset [47], which archives bi-temporal aerial views from 20 locales across Texas spanning 5 to 14 years in time. The captions in the LEVIR-CC dataset consist of two types: change descriptions and non-change descriptions. The non-change images consist of 5,039 pairs, and each pair is associated with five fixed sentences, such as "the two scenes seem identical." On the other hand, the change images consist of 5,038 pairs, and each pair has varied sentences to describe significant changes, such as "a large building replaces the woods." For data partitioning, the training set, validation set, and test set each consist of 6,815, 1,333, and 1,929 pairs of images, respectively. We follow the same data partitioning scheme for all experiments in this paper.

B. Experimental Setup

1) Implementation Details: Our implementation utilizes the PyTorch framework [48], training and evaluating on an NVIDIA RTX 3090 GPU. We utilize the Adam optimizer [49] with an initial learning rate of 2e-4. Training proceeds over a maximum of 40 epochs with a batch size of 128. After each epoch, the model's performance is assessed on the validation set. If the BLEU-4 score decreases for three consecutive epochs on the validation set, the decoder's learning rate is decayed by 0.7. The model achieving the highest BLEU-4 score on the validation data over the 40 epochs is selected as optimal. To ensure reproducibility, we set the random seed to 42. This ensures that the results can be replicated and compared across different runs or experiments.

2) Evaluation Metrics: To evaluate how closely our model's generated descriptions match the ground truths, we employ the assessment protocol commonly adopted for image captioning tasks [12], [50]–[52]. This protocol includes popular evaluation metrics such as BLEU-N ($N = 1, 2, 3, 4$) [53], METEOR [54], ROUGE-L [55], and CIDEr [56]. These scores primarily quantify the similarity between the generated descriptions and the ground-truth descriptions, serving as indicators of descriptive quality. By leveraging these metrics, we obtain automated performance assessments for our model, with higher scores signifying better ability. We also follow [26], [57] and report an average score S_m^* as an overall metric. S_m^* specifically represents the mean score of BLEU-4, METEOR, ROUGE-L, and CIDEr, aggregating results from these metrics into a single measure of the model's descriptive powers.

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

We assessed our proposed technique against several state-of-the-art approaches, including Capt-Rep-Diff [19], Capt-Att [19], Capt-Dual-Att [19], DUDA [19], MCCFormer-S [31], MCCFormer-D [31], and RSICCFormer$_c$ [12]. Capt-Rep-Diff, Capt-Att, Capt-Dual-Att, DUDA, MCCFormer-S, MCCFormer-D were originally designed for natural image

TABLE I
COMPARISONS WITH STATE-OF-THE-ART RSICC METHODS ON THE LEVIR-CC DATASET.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*	Params (M)	FPS
Capt-Rep-Diff [19]	72.90	61.98	53.62	47.41	34.47	65.64	110.57	64.52	-	-
Capt-Att [19]	77.64	67.40	59.24	53.15	36.58	69.73	121.22	70.17	-	-
Capt-Dual-Att [19]	79.51	70.57	63.23	57.46	36.56	70.69	124.42	72.28	-	-
DUDA [19]	81.44	72.22	64.24	57.79	37.15	71.04	124.32	72.58	-	-
MCCFormer-S [31]	79.90	70.26	62.68	56.68	36.17	69.46	120.39	70.68	69.0	12.9
MCCFormer-D [31]	80.42	70.87	62.86	56.38	37.29	70.32	124.44	72.11	69.0	12.4
RSICCformer$_c$ [12]	83.09	74.32	66.66	60.44	38.76	72.63	130.00	75.46	56.2	15.0
PSNet [13]	83.86	75.13	67.89	62.11	38.80	73.60	132.62	76.78	-	-
Δ	+1.24	+1.92	+2.12	+1.98	+0.79	+0.97	+3.40	+1.79	-16.3	+8.7
SEN (ours)	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**	**39.9**	**23.7**

* Note: Due to the lack of recorded Params and FPS for most of the comparison methods in their original papers, we only included the Params and FPS for the more recent methods RSICCformer, MCCFormers-S, and MCCFormers-D, which have publicly available code in this task.

TABLE II
ABLATION STUDY OF BI-TEMPORAL PRE-TRAINING.

Init	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
Random	82.42	73.81	66.56	60.83	37.56	72.17	128.18	74.68
ImageNet (3C)	84.17	75.77	68.61	62.82	39.17	73.78	132.92	77.17
SSL4EO-S12 (3C)	84.57	76.53	69.61	63.93	39.04	74.00	134.34	77.83
SSL4EO-S12 (6C)	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**

* Note: 3C denotes the pre-training on 3-channel RGB images, and 6C denotes the pre-training on 6-channel bi-temporal images.

change description and utilized the first three residual convolutional blocks (ResNet101-3) of ResNet101 as the extractor to extract visual features. ResNet101-3 (27.54M) and ResNet50 (24.57M) have comparable parameter counts, ensuring reasonable comparisons. Observing the results presented in Table I, our method SEN achieves advanced outcomes, significantly outperforming other methods on all evaluation metrics. Specifically, compared to PSNet, SEN improves by 1.24, 1.92, 2.12, and 1.98 percentage points on BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively. The comprehensive metric S_m^* increases by 1.79 percentage points. Furthermore, benefiting from the design of the single-stream network, SEN exhibits a **significantly faster inference speed**, surpassing RSICCformer by 36.71% in terms of FPS (Frames Per Second). Therefore, our method demonstrates advanced performance in both model accuracy and inference speed, indicating its effectiveness and advantage.

D. Ablation Study

1) Effects of Bi-temporal Pre-training: To validate the effectiveness of bi-temporal pre-training, we compare the performance of models initialized with random weights, ImageNet pre-trained weights, SSL4EO pre-trained weights, and temporal pre-trained weights on the RSICC task. ImageNet (3C) and SSL4EO (3C) are both pre-trained on 3-channel RGB images, while our bi-temporal pre-training (SSL4EO (6C)) is pre-trained on 6-channel bi-temporal images. As observed in Table II, SSL4EO (3C) helps diminish the divergence between pre-training and downstream data distributions relative to random or ImageNet initialization. Critically however, SSL4EO (6C) is able to further mitigate the disparity in input structure between the pre-training and fine-tuning phases, by leveraging both temporal views. This additional reduction in input domain shift afforded by bi-temporal pre-training appears to confer performance advantages over single-view pre-training, as evidenced by the improved evaluation metrics. Specifically, as shown in Table II, we observe that the model initialized with bi-temporal pre-training weights outperforms the other initialization methods across all metrics. For example, compared to ImageNet (3C), the BLEU-4 score of the bi-temporal pre-training model improves by 1.27 percentage points, the CIDEr score enhances 3.10 percentage points, and the S_m^* score rises 1.4 percentage points. This suggests that the model initialized with bi-temporal pre-training weights can better learn the inherent features of the data, thereby **reducing the domain gap** and improving the model's ability to generalize. Furthermore, as shown in Table III, our method also demonstrates advantages compared to other representative pre-training methods, such as MoCo-v2 and SeCo.

2) Effects of Single-stream Structure: To further demonstrate the advantages of the proposed method over traditional single-branch and dual-branch approaches, we compare SEN with subtractive single-stream structure and dual-stream structure. To ensure fairness, all structures are pretrained on RS images, and other settings are kept consistent. The results, as shown in Table IV, indicate that SEN outperforms other structures in all metrics. It is worth noting that the performance of the subtractive single-stream structure (Single-sub) is inferior to SEN (Single-cat). This suggests that using concatenated bi-temporal images as input is crucial as it preserves more image information, whereas subtraction loses significant image

TABLE III
COMPARISON OF DIFFERENT PRE-TRAINING MODELS.

Init	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
SeCo	84.51	76.07	69.00	63.32	39.35	73.84	133.89	77.60
MoCo-v2	84.57	76.53	69.61	63.93	39.04	74.00	134.34	77.83
SeCo_S	84.89	76.98	70.16	64.50	39.59	74.26	134.84	78.30
SEN (ours)	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**

* Note: SeCo is derived from the weights of the original paper, while the others are pretrained on SSL4EO-S12 to ensure fairness.

TABLE IV
ABLATION STUDY OF SINGLE-STREAM STRUCTURE.

Structure	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
Single-sub	83.21	74.65	67.85	62.59	38.00	72.55	130.18	75.83
Dual	84.22	76.21	69.34	63.83	38.56	73.56	131.59	76.88
Single-cat (ours)	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**

* Note: Single-sub denotes the subtractive single-stream structure, and Dual denotes the dual-stream structure.

TABLE V
ABLATION STUDY OF SEF MODULE DESIGNS.

SC	CA	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
		82.92	74.44	67.06	60.88	38.22	73.44	132.48	76.25
✓		83.39	75.37	68.51	63.03	38.38	73.08	130.47	76.24
	✓	84.00	75.99	69.23	63.68	39.04	73.76	134.36	77.71
✓	✓	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**

* Note: SC and CA denote the shared convolutional layer and channel attention respectively.

TABLE VI
ABLATION STUDY OF CAGD MODULE DESIGNS.

CrossAtt	SelfAtt	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
		84.00	75.88	68.68	62.74	38.70	73.13	130.40	76.24
✓		84.37	76.39	69.40	63.75	38.79	73.78	133.02	77.34
	✓	84.62	76.61	69.52	63.71	39.47	74.51	133.96	77.91
✓	✓	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**

information before inputting it into the network.

3) Effects of SFE Module: The SFE module consists of a shared convolutional layer and channel attention. We explore the effectiveness of the shared convolutional layer and channel attention in the SFE module through experiments. As shown in Table V, we observe that both the shared convolutional layer and channel attention in the SFE module can improve the descriptive ability of the model, with the shared convolutional layer achieving better results. This may be because the shared convolutional layer can better extract low-level features from the images, while the channel attention can better focus on the channels related to each bi-temporal image. Therefore, we choose to use the shared convolutional layer and channel attention in the SFE module.

4) Effects of CAGD Module: The CAGD module consists of cross attention and self attention. We explore the effectiveness of cross attention and self attention in the CAGD module through experiments. As shown in Table VI, the results demonstrate that incorporating either cross attention or self attention into the model yields performance gains over the baseline. This confirms the value of integrating attention mechanisms to help capture rich contextual relationships within the data. Cross attention interacts with the features of the single temporal image to enhance the feature representation of the single temporal image. Then, self attention models the difference between the enhanced features of the single temporal image feature before and after the change, thereby better learning the change information. Therefore, we select the cross attention and self attention mechanisms in the CAGD module to facilitate **explicit modeling of visual differences** within the single-stream feature extraction network.

5) Different Temporal Concatenation and Season Augmentation: The temporal relationship composition of bi-temporal concatenated images $I_{i,j} = Concat(I_{t_i}, I_{t_j})$ has two types of temporal relations: ordered ($0 \leq i \leq j \leq 4$) and unordered ($0 \leq i, j \leq 4$). Ordered relation means that the images are temporally ordered, the same as the RSICC. Unordered relation means that the images are temporally unordered in a random way to simulate the cyclical nature of the seasons. We compare the performance of ordered and unordered temporal concatenation in Table VII. The performance of ordered temporal concatenation is better than that of unordered

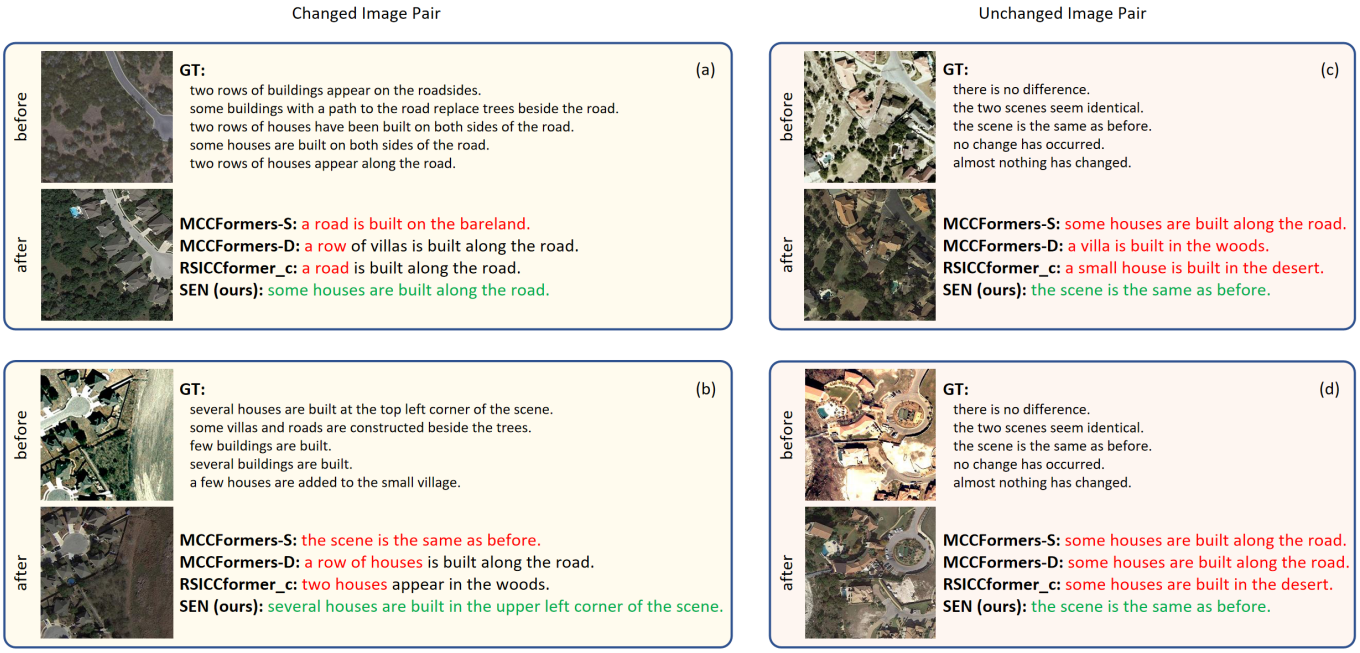


Fig. 3. Captioning results compared with different methods. The red font represents wrong predictions and the green font represents correct predictions.

TABLE VII
ABLATION STUDY OF DIFFERENT TEMPORAL RELATIONS AND SEASONS AUGMENTATION.

relations	season	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
unordered	random	84.42	76.41	69.52	63.82	38.42	73.22	130.96	76.61
unordered	augment	84.60	76.56	69.54	63.97	39.08	73.81	132.67	77.38
ordered	random	84.59	76.62	69.59	63.84	39.00	73.68	133.34	77.47
ordered	augment	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**

* Note: "relations" refers to the temporal relationship composition of bi-temporal concatenated images $I_{i,j}$, "ordered" denotes $0 \leq i \leq j \leq 4$ and "unordered" denotes $0 \leq i, j \leq 4$. Season of "augment" indicates that the random combinations of seasons for q and k are different (i.e., $q = \Gamma(I_{i_q,j_q})$ and $k = \Gamma(I_{i_k,j_k})$), and "random" indicates the same random combinations of seasons for q and k (i.e., $q = \Gamma(I_{i,j})$ and $k = \Gamma(I_{i,j})$).

temporal concatenation, which may be due to the fact that the importance of ordered change information is stronger than the periodicity of the four seasons, for example, the level of urbanization in a place is orderly, rather than maintaining periodic changes with the seasons. We also report the results of different random seasonal choices for the two positive views in Table VII. It is better to use different randomly selected seasons to form positive sample pairs than to randomly select the same seasons, which make the model insensitive to a wider range of seasonal variations and better adapt to seasonal changes. This is consistent with the results in [42].

E. Qualitative Visualization Analysis

1) Qualitative Comparisons of Different Methods: To more intuitive understanding of the effectiveness of our method compared to others, four representative image pairs are selected for qualitative comparison. It consists of two pairs of changed images and two pairs of unchanged images. We compared the caption results generated by different methods [12], [31] and conducted qualitative analysis. As shown in Figure 3, when presented with changed images, our method can better perceive the changed region and accurately describe the specific content of the change. Specifically, (a) and (b) both contain interference caused by ground color changes due to lighting variations, etc. In (a), MCCFormers-S and RSICCformer_c cannot handle these interferences, and both incorrectly identify the existing road as a new road. MCCFormers-D does not accurately describe the quantitative relationship, while our method can exclude these interferences through seasonal contrastive pre-training, accurately perceive the newly built houses on the side of the road, and correctly describe the quantitative relationship. This can be attributed to the benefits of contrastive pre-training, which not only helps alleviate the domain gap but also enables the learning of invariant features under different lighting conditions. In (b), due to the existence of old houses, the newly built houses in the upper left corner are not obvious enough, and other methods have failed to generate accurate descriptions, and even think that there is no change in the image pair (MCCFormers-S). However, our method utilizes the SFE and CAGD modules to explicitly model the differential regions and enhance the attention weights on the changed areas. This enables our model to capture these subtle changes that may be easily

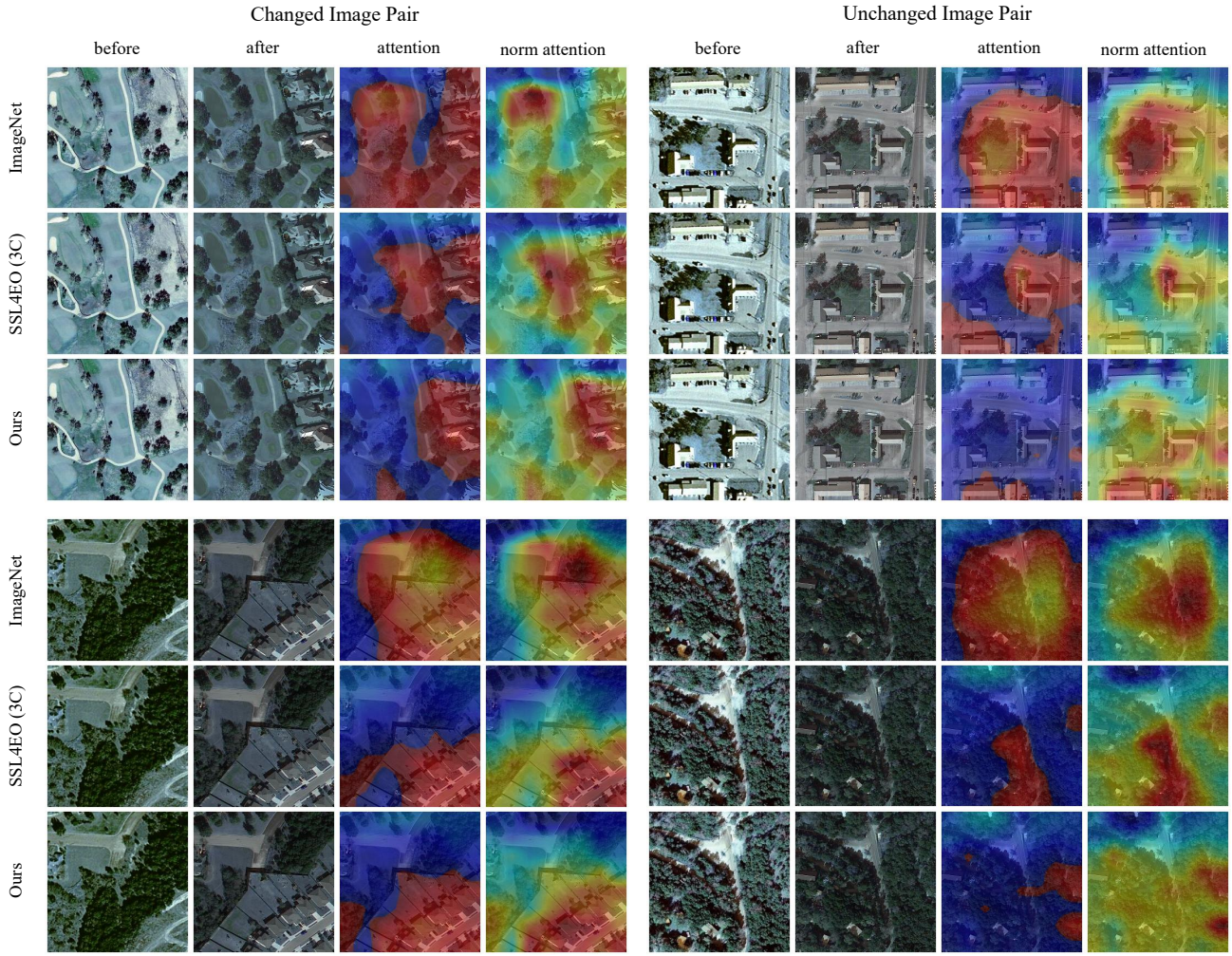


Fig. 4. Visualization of attention maps for various initialization methods of the feature extractor. These maps are derived from the difference feature attention weights of the encoder output. The smaller the change, the larger the attention weight, and vice versa. Smaller values correspond to cold tones (e.g., blue, cyan), larger values correspond to warm tones (e.g., red, pink). The intermediate values correspond to green, yellow, orange, etc. "norm attention" indicates the normalized attention weights. "attention" indicates the original unnormalized attention weights. "attention" mainly concerned with the fact that the attention weights should be small for the unchanged image pair, and normalization will break this relationship, so we retain the unnormalized weights to better observe the attention weights for the unchanged image pair.

overlooked in complex scenes. Additionally, our method is able to accurately predict the orientation of newly constructed houses. In the unchanged images (c) and (d), the pseudo-change interference caused by the change in illumination and the complex scene are challenging for the model. Under these challenges, other methods incorrectly generate some descriptions about the change (actually not), while our method can accurately respond to these pseudo-change interferences and determine that the image has not undergone meaningful changes such as object disappearance or addition. In summary, by incorporating contrastive pre-training and utilizing the SFE and CAGD modules, our method effectively addresses pseudo-changes and captures subtle variations in complex scenes.

2) Visualization Analysis of Pre-training: To better understand the effectiveness of bi-temporal pre-training, we visualize the attention maps of the feature extractor initialized with different pre-training weights: ImageNet, SSL4EO (3C), ours (SSL4EO (6C)). These attention maps are derived from the attention weights of the encoder outputs. It can be observed in Figure 4 that for the changed image pair, the attention weights of the model pre-trained with our method are more concentrated in the changed region, while the unchanged region is also excluded. Specifically, as seen upon comparing the three images rows in the upper left corner, the model pre-trained with ImageNet focuses on the unchanged region, and the model pre-trained with SSL4EO (3C) weakens this erroneous focus, but the focus on the changed region is still insufficient. Our method can correctly focus on the changed region and suppress the interference of irrelevant changes. From the comparison of the three rows of images in the lower right corner, it is apparent that the model pre-trained with ImageNet still does not put more attention on the most significant changes, and the model pre-trained with SSL4EO (3C) focuses more on the significant change region, but does not fully focus on all the change regions. Our method can focus on all the change regions more completely. This suggests that bi-temporal pre-training can better learn features that are more suitable for the RSICC task, reduce the data distribution

TABLE VIII
THE PERFORMANCE OF THE MODEL IN DIFFERENT PRE-TRAINING EPOCHS.

Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
100	83.64	75.72	68.80	63.19	38.69	73.16	131.68	76.68
200	84.30	76.38	69.62	64.09	39.21	74.20	134.94	78.11
300	**85.10**	**77.05**	**70.01**	64.09	**39.59**	**74.57**	**136.02**	**78.57**
400	84.68	76.28	69.08	63.22	39.41	73.90	134.27	77.70
500	84.81	76.81	69.82	**64.17**	39.56	74.27	133.99	78.00

gap and input gap, and strengthen the model's capacity for generalization. To the unchanged image pair, the entire image is an unchanged region, so the attention weights should be relatively low as a whole. As shown in the right column of the figure, whether it is the normalized attention weights or the unnormalized attention weights, the high attention weight region output by our model is significantly smaller than the high attention weight region output by the ImageNet and SSL4EO (3C) models. This is consistent with the feature requirements of the unchanged image pair, that is, the entire image is an unchanged region, so the attention weights should be relatively low as a whole. This further confirms the effectiveness of our method.

3) Visualization Analysis of SEF and CAGD Modules: To better understand the effectiveness of the SFE and CAGD modules, the attention maps generated by different modules are visualized. As shown in Figure 5, it can be seen that the SFE module first extracts the rough change region attention weights, and then the CAGD module further enhances the attention weights of the change region and concentrates on the region missed by the SFE. Specifically, the SFE module can focus on some regions with obvious changes, but it is not enough to focus on the changes in small regions. The CAGD module can further focus on the changes in small regions to cover the complete change region, and at the same time, it can further enhance the regions focused by the SFE.

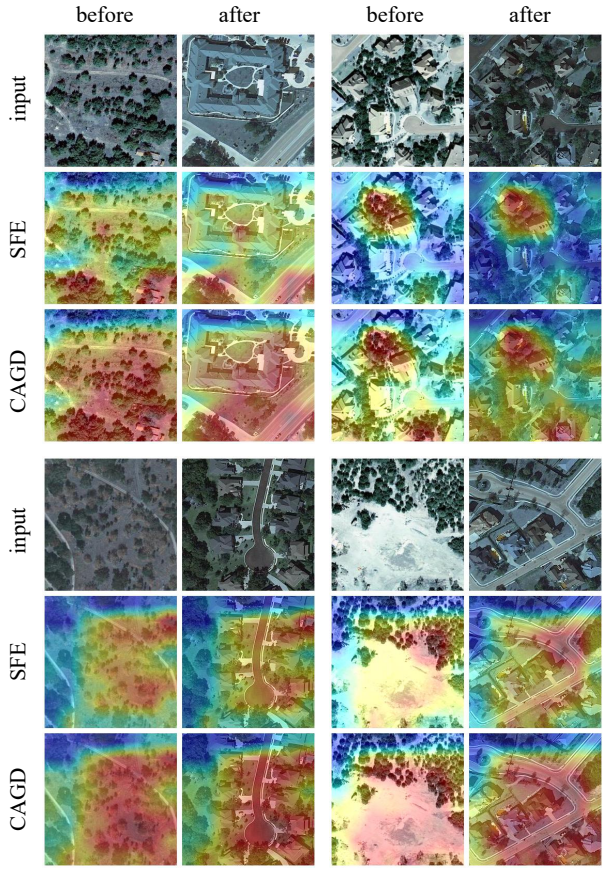


Fig. 5. Visualization of the attention maps for different modules of SEN.

F. Parametric Study

1) Number of Pre-training Epochs: The previous methods randomly select two images from the four seasonal images at a location to form sample pairs (total $4 \times 4 = 16$ combinations). However, our proposed method involves randomly selecting two images, I_i and I_j, and concatenating them to create a six-channel bi-temporal images $I_{i,j}$ (total of $4 + 3 + 2 + 1 = 10$ combinations). Then, sample pairs are generated by augmenting these bi-temporal images (total of $10 \times 10 = 100$ combinations). This significantly increases the variety of samples used in the training process. Therefore, we examine the effectiveness resulting from different pre-training epochs. The Table VIII presents the outcomes of different pre-training epochs. The results demonstrate that pre-training 300 epochs is sufficient. Using excessively high or low numbers of epochs may have a negative impact on the model's performance. This could be due to overfitting, where the model becomes too specialized to the pre-training data and the learned features are less applicable to downstream tasks, or underfitting, where the model has not learned enough generalizable features from the pre-training data.

2) Number of Fine-tuning Layers: For the pre-trained model weights f_q, the shallow layers typically learn low-level generic features, which tend to have strong generalization capability. It is common to freeze the weights of the shallow layers to prevent the model from learning low-level features, and to fine-tune the weights of the deeper layers to adapt to the downstream tasks. We examined how the performance of the model is affected by fine-tuning varying numbers of layers. The results are presented in Table IX. It can be observed that fine-tuning the layers starting from conv2_x achieves the best performance for the model. By fine-tuning the higher layers, the model can learn task-specific features and further improve its performance on the RSICC tasks. Furthermore, it can be observed that when all weights of f_q are frozen, the performance of the model on RSICC is significantly poor. The results imply that when all the pre-trained weights are frozen, it hampers the model's capability to adapt to the unique

TABLE IX
THE PERFORMANCE OF THE MODEL IN DIFFERENT FINE-TUNING LAYERS.

Layer	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
conv1	84.84	76.91	**70.09**	**64.61**	39.20	74.10	134.96	78.22
conv2_x	**85.10**	**77.05**	70.01	64.09	**39.59**	**74.57**	**136.02**	**78.57**
conv3_x	84.14	75.67	68.37	62.37	38.58	73.84	132.26	76.76
conv4_x	82.67	74.33	67.47	62.05	38.01	72.17	128.28	75.13
conv5_x	82.38	74.24	67.74	62.63	37.98	72.19	129.37	75.54
frozen all	76.87	68.21	61.87	57.45	33.53	67.24	113.23	67.86

* Note: conv1, conv2_x, conv3_x, conv4_x, conv5_x denote the convolutional layers corresponding to ResNet [16]. The "Layers" column represents the starting layer for the fine-tuning, and it includes all subsequent layers until the last layer of the pre-trained model f_q.

TABLE X
THE PERFORMANCE OF THE MODEL IN DIFFERENT DEPTHS OF CAGD AND CAPTION DECODER

N_e	N_d	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	S_m^*
1	1	83.94	75.31	68.03	62.11	38.67	73.91	133.17	76.97
1	2	84.29	75.68	68.26	62.23	39.03	73.78	132.42	76.87
1	3	82.97	74.15	66.79	60.87	38.20	72.81	128.98	75.22
2	1	**85.10**	**77.05**	**70.01**	**64.09**	**39.59**	**74.57**	**136.02**	**78.57**
2	2	84.45	76.14	69.04	63.45	39.08	73.72	132.76	77.25
2	3	83.96	75.54	68.45	62.65	39.01	73.86	132.65	77.04
3	1	84.15	76.18	69.13	63.36	38.60	73.49	131.80	76.81
3	2	84.50	75.93	68.73	62.98	38.79	73.11	131.23	76.53
3	3	83.54	74.97	67.84	62.23	38.66	73.56	131.06	76.38

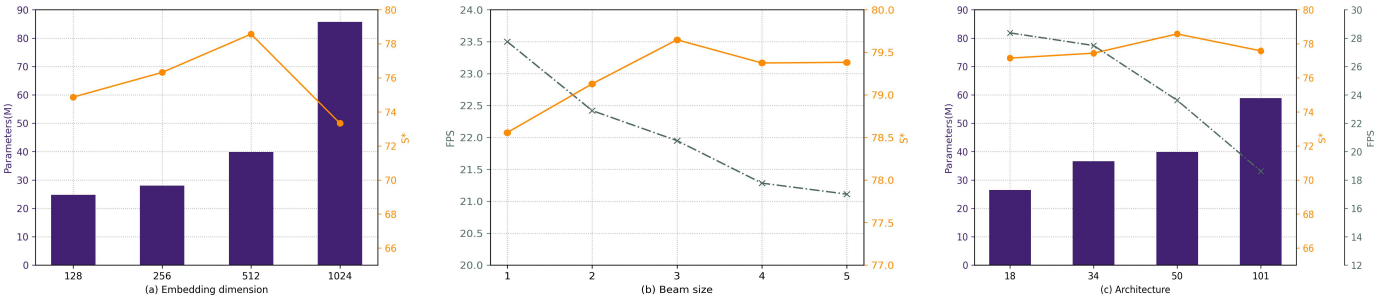


Fig. 6. (a) Embedding dimension: The impact of different embedding dimensions D on the model performance S_m^* and parameter size. (b) Beam size: The impact of different beam sizes on the model performance S_m^* and inference speed FPS during inference. (c) Architecture: The impact of different ResNet architectures on the model performance S_m^*, parameter size, and inference speed FPS during inference.

characteristics and intricacies of the RSICC dataset.

3) Depth of Network: The depth of the CAGD and caption decoder, denoted as N_e and N_d respectively, plays a crucial role in the performance of the model. Both excessively deep and shallow can have a negative impact on the model's ability to accurately generate descriptions. We explored different combinations of N_e and N_d, ranging from 1 to 3. From the Table X, we observe that the model performs best when $N_e = 2$ and $N_d = 1$, as indicated by the highest values of all metrics. We finally choose $N_e = 2$ and $N_d = 1$ as the optimal depth for our study.

4) Embedding dimension, beam size and architecture: As shown in Figure 6, We also conduct experiments on different embedding dimensions, beam sizes during inference, and feature extractor architectures. From Figure 6 (a), We observe that as the embedding dimension increases, the model's parameter size and performance gradually improve, but when the embedding dimension reaches 512, the performance of the model begins to decline. This may be because a large embedding dimension will cause the model to overfit. Therefore, we choose the embedding dimension to be 512. The beam size during inference has an impact on the model performance S_m^* and inference speed FPS. As shown in Figure 6(b), as the beam size increases, the model's performance gradually improve until the beam size reaches 3, after which they begin to decrease. At the same time, as the beam size increases, the model's inference speed continues to decrease. Described in Figure 6(c), as the number of layers of ResNet increases, the size of the model's parameters continues to increase, resulting in a gradual decrease in inference speed. However, the performance of the model reaches its peak at ResNet-50. Therefore, we choose ResNet-50 as the feature extractor.

V. CONCLUSION

In this work, we investigate the problem of insufficient potential mining of single-stream visual feature extractors from

the aspects of domain gap and implicit difference modeling, and propose the SEN model. In terms of domain gap, we propose the bi-temporal pre-training method, which learns features that are more suitable for the RSICC task through self-supervised learning on a large-scale bi-temporal RS image dataset, reduces the data distribution gap and input gap, and improves the model's generalization ability. In terms of implicit difference modeling, we propose the SEF module and CAGD module to better learn the change information. SEN achieves cutting-edge performance on the RSICC task, and also has faster inference speed.

REFERENCES

[1] H. Taubenböck, M. Wegmann, A. Roth, H. Mehl, and S. Dech, "Urbanization in india–spatiotemporal analysis using remote sensing data," *Computers, environment and urban systems*, vol. 33, no. 3, pp. 179–188, 2009.

[2] M. Imhoff, C. Tucker, W. Lawrence, and D. Stutzer, "The use of multisource satellite and geospatial data to study the effect of urbanization on primary productivity in the united states," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 6, pp. 2549–2556, 2000.

[3] J.-Y. Rau, L.-C. Chen, J.-K. Liu, and T.-H. Wu, "Dynamics monitoring and disaster assessment for watershed management using time-series satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1641–1649, 2007.

[4] M. Zhang, Q. Li, Y. Miao, Y. Yuan, and Q. Wang, "Difference-guided aggregation network with multi-image pixel contrast for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[5] C. Giardino, M. Bresciani, P. Villa, and A. Martinelli, "Application of remote sensing in water resource management: the case study of lake trasimeno, italy," *Water resources management*, vol. 24, pp. 3885–3899, 2010.

[6] O. M. Bello and Y. A. Aina, "Satellite remote sensing as a tool in disaster management and sustainable development: towards a synergistic approach," *Procedia-Social and Behavioral Sciences*, vol. 120, pp. 365–373, 2014.

[7] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Edge neighborhood contrastive learning for building change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2022.

[8] M. Zhang, Q. Li, Y. Miao, Y. Yuan, and Q. Wang, "Difference-guided aggregation network with multiimage pixel contrast for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[9] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[10] S. Chang, M. Kopp, and P. Ghamisi, "Sketched multiview subspace learning for hyperspectral anomalous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[11] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[12] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[13] C. Liu, J. Yang, Z. Qi, Z. Zou, and Z. Shi, "Progressive scale-aware network for remote sensing image change captioning," *arXiv preprint arXiv:2010.11929*, 2020.

[14] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, "Captioning changes in bi-temporal remote sensing images," in *IEEE International Geoscience and Remote Sensing Symposium*, 2021, pp. 2891–2894.

[15] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, "Change captioning: A new paradigm for multitemporal remote sensing image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[19] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4624–4633.

[20] L. Hu, Q. Liu, J. Liu, and L. Xiao, "Prbcd-net: Predict-refining-involved bidirectional contrastive difference network for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[21] L. Hu, R. Meng, Q. Liu, J. Liu, and L. Xiao, "Novel contrastive regularized bipartite network for unsupervised change detection," in *Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing*, 2023, pp. 1–5.

[22] X. Shi, X. Yang, J. Gu, S. Joty, and J. Cai, "Finding it at another side: A viewpoint-adapted matching encoder for change captioning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 574–590.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *International Conference on Learning Representations*, 2016, pp. 1–16.

[25] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[26] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[27] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "SAAN: similarity-aware attention flow network for change detection with VHR remote sensing images," *arXiv preprint arXiv:2308.14570*, 2023.

[28] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, and L. V. Gool, "Efficient and explicit modelling of image hierarchies for image restoration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 278–18 289.

[29] D. Zheng, Z. Wu, J. Liu, Y. Xu, C. Hung, and Z. Wei, "Explicit change-relation learning for change detection in VHR remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 21, pp. 1–5, 2024.

[30] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," *arXiv preprint arXiv:1808.10584*, 2018.

[31] Y. Qiu, S. Yamamoto, K. Nakashima, R. Suzuki, K. Iwata, H. Kataoka, and Y. Satoh, "Describing and localizing multiple changes with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1971–1980.

[32] Y. Tu, T. Yao, L. Li, J. Lou, S. Gao, Z. Yu, and C. Yan, "Semantic relation-aware difference representation learning for change captioning," in *Findings of the Association for Computational Linguistics*, 2021, pp. 63–73.

[33] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[34] C. Liu, K. Chen, Z. Qi, H. Zhang, Z. Zou, and Z. Shi, "Pixel-level change detection pseudo-label learning for remote sensing change captioning," *arXiv preprint arXiv:2211.07044*, 2023.

[35] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[36] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 14, pp. 4205–4230, 2021.

[37] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[38] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[39] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, 2020, pp. 21 271–21 284.

[40] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*, 2020, pp. 9912–9924.

[41] O. Mañas, A. Lacoste, X. Giro-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[42] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation," *arXiv preprint arXiv:2211.07044*, 2022.

[43] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv:2003.04297*, 2020.

[44] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[45] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[47] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[51] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[52] M. Hosseinzadeh and Y. Wang, "Image change captioning by learning from an auxiliary task," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2725–2734.

[53] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the Association for Computational Linguistics*, 2002, pp. 311–318.

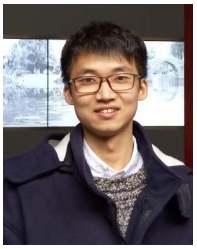
[54] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[55] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[56] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[57] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

Qing Zhou is currently pursuing the Ph.D degree in computer science and technology with the school of Artificial Intelligence, Optics and Electronics (iOPEN). His research interests include computer vision and pattern recognition.



Junyu Gao (Member, IEEE) received the B.E. and Ph.D. degrees in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2015 and 2021, respectively. He is curently a Researcher with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include computer vision and pattern recognition.

Yuan Yuan (Senior Member, IEEE) is currently a Full Professor with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC and ICIP. Her current research interests include visual information rocessing and image/video content analysis.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.