ESAN: An Efficient Semantic Attention Network for Remote Sensing Image Change Captioning

Anonymous ACL submission

Abstract

With the continuous progress of remote sensing technology, an increasing number of remote sensing images containing rich geographical and environmental information is obtained. Unlike natural images, remote sensing images usually cover a large area and have complex spatial distribution, making it a challenge to accurately extract and describe changes from images. In order to effectively mine and utilize the rich semantic information contained in the image to guide the decoder to generate high-quality change descriptions, we propose an efficient semantic attention network (ESAN). Specifically, we first perform global efficient semantic representation (GESR) on the obtained remote sensing feature map to promote the understanding of complex scenes in remote sensing images. Then we further propose a cross-semantic feature enhancement module (CSFE) to effectively distinguish semantic changes from irrelevant changes. Finally, we input the obtained image features into the adaptive multi-layer Transformer decoder to guide the generation of change description. Extensive experiments on two representative remote sensing datasets, Dubai-CC and LEVIR-CC, demonstrate the superiority of the proposed model over many advanced technologies.

1 Introduction

With the rapid development of remote sensing technology, a large amount of high-resolution remote sensing image data has been acquired. Remote sensing images are not only used for scientific research, but also widely used in damage assessment (Xu et al., 2019), urban planning (Chen and Shi, 2020), environmental monitoring (de Bem et al., 2020) and other fields. Accurate and semantically rich descriptions of these image changes not only help to improve the image interpretation capability, but also make remote sensing images easier to be understood by non-specialized users. In addition, the accurate change description also provides a powerful tool for decision-making, planning management and disaster response.

The remote sensing image change description task aims to describe the change content in a remote sensing image pair in natural language. It involves two remote sensing images, usually corresponding to different points in time in the same area. The model needs to understand the differences between these two images, including changes in features, new or disappeared elements, etc., and generate text descriptions that can clearly express these changes. Change descriptions have recently gained attention in geoscience and remote sensing due to their ability to extract high-level semantic information about land cover changes.

In recent years, several methods have been proposed to improve the performance of image change description models.

Early pioneer work (Jhamtani and Berg-Kirkpatrick, 2018) proposed a task to describe the difference between similar image pairs through object-level difference description. Subsequent research focused on the relationship between semantic changes and interference factors, and proposed a series of models, including dual dynamic attention model (DUDA) (Park et al., 2019), viewpoint adaptive matching encoding (Shi et al., 2020), multi-change caption transformer (MCCFormers) (Qiu et al., 2021), etc., to cope with the challenges in the actual scene. At the same time, some methods emphasize the importance of tasks, such as new training schemes (Hosseinzadeh and Wang, 2021) and multimodal end-to-end siamesed difference captioning model (SDCM) (Ariyo et al., 2019a). Recent work has further explored the relationship-aware attention mechanism (Tu et al., 2023b, 2021b), distance-sensitive self-attention (DSA) (Ji et al., 2023), cyclic consistency (VACC) (Kim et al., 2021), etc., to improve the model 's perception of complex changes. Methods such as the new modeling framework (Yao et al., 2022) and the progressive scale-aware network (PSNet) (Liu et al., 2023a) aim to optimize the overall performance of the model. The studies work together to overcome the challenges of semantic understanding, viewpoint change and multi-scale information utilization, and provide rich exploration and innovation for the task of remote sensing image change description. However, although significant progress has been made in the task of image change description, there are still some deficiencies in semantics.

At present, the change description model for remote sensing images lacks fine-grained semantic understanding, which often needs to rely on global context information to obtain a more accurate interpretation. For example, a single pixel change may only have a clear meaning in the global context. In order to provide scene background for fine-grained changes and make the model better understand the semantics of local changes, we proposes an Efficient Semantic Attention Network (ESAN), which uses different semantic relationship modules and adaptive decoder based on Transformer to generate remote sensing change descriptions. Through a large number of experiments, we prove that ESAN can produce a more accurate and realistic description of the changes between remote sensing image pairs, and achieve the best performance compared with the existing change description methods.

The contributions of this paper are summarized as follows:

(1) GESR module is designed to enhance the feature extraction of global semantics, which operates at the perceptual level, deeply mines internal feature associations, grasps global association information, and provides scenarios for fine-grained semantic understanding.

(2) CSFE module is designed to facilitate the accurate identification and description of fine-grained changes. It carefully checks and compares the information between the image 's own features and the common difference features, especially pays attention to the difference representation, and obtains the actual semantic changes based on the global features.

(3) In order to improve the adaptive ability of the model, a multi-stage adaptive Transformer model is formed as the decoder to translate the obtained change features into natural language sentences. Extensive experiments show that ESAN outperforms other state-of-the-art methods on the Dubai-CC and LEVIR-CC datasets.

2 Related Work

2.1 Image Captioning

Describing image content in natural language has been an active area of artificial intelligence research. A variety of image description methods dedicated to improving the state of the art of image description have been proposed. In order to fully exploit the short-term spatial semantic relations, (Li et al., 2022) introduced the long-short-term relational converter (LSRT). On the other hand, the paper (Tu et al., 2022) proposed an internal and relational embedding transformer (I^2Transformer) to effectively understand caption semantics and the relationship between them. (Yu et al., 2022) applied the dual attention mechanism to the pyramid feature map, fully considering the context information. Although the self-attention (SA) network has achieved great success in image captioning, the existing SA network has the problems of distance insensitivity and low-rank bottleneck. To this end, (Ji et al., 2023) introduced distance-sensitive self-attention (DSA) and multi-branch self-attention (MSA). The traditional attention mechanism usually only considers the one-way flow from vision to linguistics, resulting in that the visual features of attention are usually irrelevant to the state of the target word. (Tu et al., 2023b) improved the traditional attention mechanism and proposed a relationship-aware attention mechanism, namely, visual-to-visual homogeneity graph (HMG) and linguistic-to-visual heterogeneity graph (HTG), respectively. These studies have made in-depth explorations of image caption generation tasks at different levels. Although some achievements have been made in semantic understanding, there is still room for improvement.

2.2 Change Captioning

In recent years, the task of image change description has attracted wide attention, and researchers have proposed a series of innovative methods to solve this task. (Jhamtani and Berg-Kirkpatrick, 2018) made a pioneering contribution to this field, proposing for the first time the task of describing the difference between similar image pairs. Subsequently, (Park et al., 2019) introduced the Double Dynamic Attention Model (DUDA), which distinguishes the interference factors and semantic changes. In order to solve the viewpoint change problem, (Shi et al., 2020) proposed viewpoint adaptive matching coding. Different from other methods, (Hosseinzadeh and Wang, 2021) explored a new image change description training scheme. (Qiu et al., 2021) introduced the multi-change caption transformer (MCCFormers). (Tan et al., 2019) elaborated on the editing transformation between two images, providing a theoretical basis for subsequent research. Further, (Ariyo et al., 2019b) proposed a fully convolutional CaptionNet (FCC). Through the multi-modal end-to-end connected difference caption model (SDCM), (Ariyo et al., 2019a) captured, aligned, and calculated the differences between the two image features. (Chang and Ghamisi, 2023) proposed an attention change caption network, focusing on generating accurate captions. In order to improve the model 's ability to perceive various changes, a neighborhood contrast transformer is designed in (Tu et al., 2023a). In addition, (Yue et al., 2023) proposed the internal and internal representation interaction network (I3N), which focuses on learning fine differential representation. (Kim et al., 2021) proposed a view-independent changing subtitle network with cyclic consistency (VACC). Facing the challenges, (Yao et al., 2022) proposed a new modeling framework to learn stronger visual and linguistic associations. Then, (Liu et al., 2023a) introduced a progressive scale-aware network (PSNet) to solve the weaknesses in multi-scale information extraction and utilization. Finally, (Huang et al., 2022) proposed an instance-level fine-grained differential captioning (IFDC) model, which focuses on the rich explicit features of the object. However, although the above research has made significant progress, there are still some shortcomings. First of all, the current method mainly focuses on the description of object-level differences, while fine-grained semantic changes still need to be further explored. Secondly, there is still a lack of comprehensive solutions for subtle semantic changes in specific scenarios and complex situations. In addition, the current research pays less attention to the rich explicit features of objects in the context, which may pose some challenges in accurately locating changing objects.

3 ESAN Model

The description task for remote sensing image change aims to generate semantic descriptions of remote sensing image changes through automated methods. Formally, given a pair of images (I_1, I_2), the model generates a caption describing what has been changed between I_1 and I_2: $f(I_1, I_2; \theta) \to \hat{C}$, where θ denotes the model parameters of the change captioning network and $\hat{C}$ represents the generated caption.

As shown in Figure 1, the architecture of our method consists of three parts : (1) GESR module quickly captures the global semantic information of the image from two different directions; (2) CSFE module is responsible for the information flow interaction between different features, and learns the contrast information between them, so as to pay attention to the semantic information of actual changes; (3) The multi-stage adaptive Transformer decoder translates the learned change features into natural language sentences.

3.1 Global Efficient Semantic Representation

Given a dual-temporal image pair (I_1, I_2), we first use the pre-trained ResNet101 (He et al., 2016) model to extract image features and represent them as X_1, X_2, respectively, where, the feature map $X_i \in R^{C \times H \times W}$, C, H, W represent the number, height, and width of channels, respectively.

However, the features extracted by the ResNet network are relatively sparse and independent. It is difficult to distinguish fine-grained changes from a large number of unrelated object regions by using these features alone. In fact, there is a semantic relationship between these original object features (Wu et al., 2019; Huang et al., 2020; Yin et al., 2020). In image understanding, capturing the semantic relationship between objects is crucial for a comprehensive understanding of the image.

Global context information can provide the relationship between objects in the image, scene structure and deeper semantic understanding (Huang et al., 2019). Remote sensing images involve complex scenes. Therefore, global context information is of great significance for the task of remote sensing image caption generation, which is helpful to improve the comprehensive performance of image understanding. For remote sensing images, high-resolution feature maps are often generated, while non-local neural networks need to generate huge attention maps to measure the relationship between each pixel pair, resulting in high computational complexity and occupying a large amount of CPU memory.

We first implicitly model the global semantic relationship in each image. Then, we use a self-attention block to dynamically learn the relationship between different positions according to the

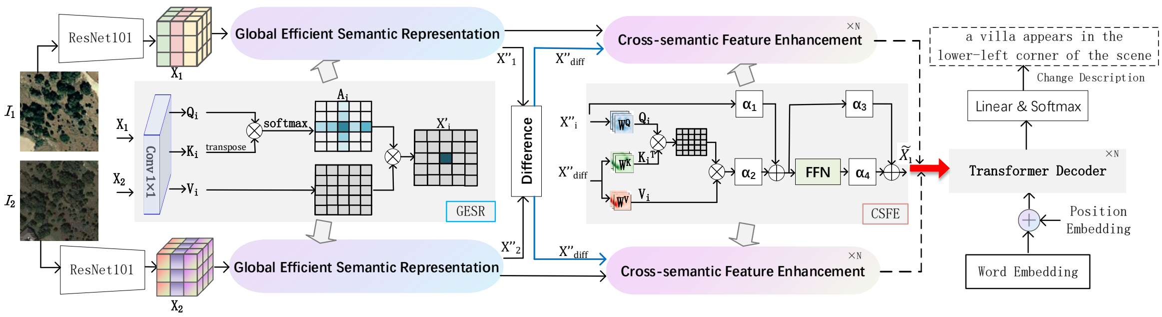

Figure 1: Overall architecture of our ESAN model.

semantic information of each position in the input sequence. For remote sensing images with a wide range of coverage, we believe that it is very important to capture the effective semantic information of each position in the sequence.

We first use two 1×1 convolution layers on the feature map $X_i \in R^{C\times H\times W}$ to generate two feature maps Q and K, where $\{Q, K\} \in R^{C'\times H\times W}$, C' is the number of channels after dimensionality reduction, and the value is less than C. At each position p in the Q-space dimension, the vector $Q_p \in R^{C'}$ can be obtained. At the same time, by extracting features from K, the feature vector set $\Omega p \in R^{(H+W-1)\times C'}$ is obtained, which is located in the same row or column as the position p. Then the attention map $A \in R^{(H+W-1)\times(H\times W)}$ is calculated by Equation 1, where $i = [1, \ldots, H + W - 1]$.

$$A_{i,p} = softmax(Q_p \Omega_{i,p}^T) \tag{1}$$

At the same time, another 1×1 convolution layer is used to generate the feature $V \in R^{C\times H\times W}$ on $X_i \in R^{C\times H\times W}$. On each position p in the V space dimension, the vector $V_p \in R^C$ and a set $\phi_p \in R^{(H+W-1)\times C}$ are obtained,ϕ_p is the set of eigenvectors in V that are in the same row or column as the position p. Finally, we can obtain the global context information as Equation 2:

$$X'_p = \sum_{i=0}^{H+W-1} A_{i,p}\phi_{i,p} + X_p \tag{2}$$

Where, X_p is the eigenvector of position p in $X' \in R^{C\times H\times W}$. After that, we transform the existing feature map $X'_i \in R^{C\times H\times W}$ into $X'_i \in R^{C\times N}$, where $N = H \times W, i \in (1, 2)$. Then, Q, K, V are embedded into the same-dimensional embedding. The process can be denoted as Equation 3:

$$X''_i = Softmax\left(\frac{(X'_i W_i^Q)(X'_i W_i^K)^T}{\sqrt{d_k}}\right)(X'_i W_i^V) \tag{3}$$

Where W_i^Q, W_i^K, W_i^V are learnable parameter matrices, $i \in (1, 2)$. d_k is the dimension of the vector. Softmax is the activation function.

After adding the global context information to the local feature X, the feature has a wide context view, which can better capture the global semantic information of the image. When the model can deeply grasp the comprehensive information in the image, it can better distinguish between semantic changes and irrelevant changes. That is to say, the results of the global efficient semantic representation module are used as the input of the cross-semantic feature enhancement stage, which effectively constructs the relationship between image sequence features, which is the basis for obtaining reliable difference representation in the cross-semantic feature enhancement stage.

3.2 Cross-Semantic Feature Enhancement

In order to enable the model to effectively locate semantic changes without being affected by irrelevant changes, we designed a cross-semantic feature enhancement module to effectively reveal the change features. Through feature interaction, the complementary relationship between different time-phase features is retrieved, supplementary information is learned, and the model 's ability to compare and locate different time-phase features is improved.

After GESR module, we get X''_1 and X''_2 as the input of CSFE module, and then we capture the semantic difference X''_{diff} in object features and relationships through $X''_2 - X''_1$. Due to the existence of interference information, the difference feature X''_{diff} contains irrelevant information. Through the semantic information flow interaction between X''_{diff} and X''_1, and between X''_{diff} and X''_2, we can distinguish semantic changes from unrelated changes (such as seasonal changes).

Firstly, the tokens of T_i are projected to one separate matrix $Q_i \in R^{HW\times C}$ to compute a set of queries. And then, the tokens of T_{diff} are projected to the other two separate matrices K_{diff} , $V_{diff} \in R^{HW \times C}$ to compute a set of keys and values (Equation 4).

$$Q_i = T_i W^Q, K_{diff} = T_{diff} W^K, V_{diff} = T_{diff} W^V \quad (4)$$

Where W^Q, W^K, W^V are learnable parameter matrices and $i \in (1, 2)$.

Secondly, the matrix is built via dot-product operation, followed by a softmax function normalizes the scores. After that, the feature vectors $\widetilde{X_1}$ and $\widetilde{X_2}$ are obtained by multiplying the matrix with V_{diff} (Equation 5), which refines the features X_1'' and X_2'' by leveraging the similarity across semantics. That is to say, we can establish the characteristic relationship between the corresponding positions between X_1'' and X_{diff}'', and between X_2'' and X_{diff}''. Where d_k is the dimension of the vector and $i \in (1, 2)$.

$$\widetilde{X_i} = softmax\left(\frac{Q_i\, K_{diff}{}^T}{\sqrt{d_k}}\right) V_{diff}, \quad (5)$$

Thirdly, the vectors $\widetilde{X_1}$ and $\widetilde{X_2}$ are added to the original input sequence through a residual connection (Dosovitskiy et al., 2021) (Equation 6), where W^O denotes the output weight matrix before FFN layer and $i \in (1, 2)$.

$$\widetilde{X_i'} = \partial_1 \widetilde{X_i} W^O + \partial_2 X_i'' \quad (6)$$

Finally, the feed-forward network (FFN) as that in the standard Transformer is applied to further improve the robustness and accuracy of the model and output the enhanced features $\hat{X}_1$ and $\hat{X}_2$ (Equation 7). Where ∂_1, ∂_2, ∂_3, ∂_4 are the learnable parameters.

$$\hat{X}_i = \partial_3 \widetilde{X_i'} + \partial_4 FFN\left(\widetilde{X_i'}\right) \quad (7)$$

3.3 Description Generation

In the image description task, the Transformer decoder (Vaswani et al., 2017) has multiple advantages over the traditional LSTM decoder. For example, Transformer captures long-distance dependencies through parallel computing and self-attention mechanisms, and provides spatial information through position coding. Therefore, we use the decoder shown in Fig 2 to generate the change description.

Specifically, each decoder consists of N stacked Transformer decoding blocks. Each block consists of a masked multi-head attention layer, an Encoder-Decoder cross-attention layer and a feed forward layer. Now we represent the visual sequence obtained from the visual encoder as $\widetilde{V_I}$. We cannot directly import descriptive sentences into the model, so each word in the sentence is represented as a one-hot vector w_i. The description decoder takes w_i as input, and the masked multi-head attention mechanism embeds the word through Equation 8. And the embedding feature $\hat{E}[W]$ is calculated. Then, through Encoder-Decoder cross-attention, $\hat{E}[W]$ is used to query the most relevant hidden layer feature $\hat{H}$ from the visual feature $\widetilde{V_I}$. After that, $\hat{H}$ learns the enhanced representation $\widetilde{H}$ through the forward propagation network.

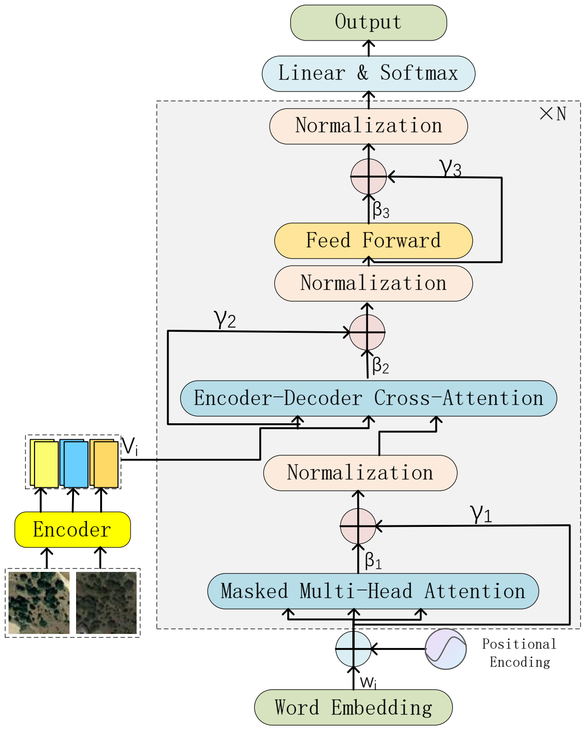

Figure 2: Visualization of the description generator.

$$E[W] = \{E[w_1], \ldots; E[w_m]\} \quad (8)$$

We apply learnable coefficients on each branch of the residual connection, such as β_1 , β_2 , β_3 , γ_1, γ_2, γ_3, so that each layer can be adaptively adjusted according to the characteristics of the upper and lower layers, thereby increasing the adaptability of the model. By adjusting these parameters, the model can better control the information interaction between different levels and realize the dynamic adjustment of different levels of features.

After stacking N Transformer decoding blocks, the hidden layer state output of the last block h^N

E.D	D.D	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1	1	84.36	77.06	69.73	63.56	38.82	73.86	131.07
2	1	**86.03**	**78.14**	**70.87**	**64.86**	**40.10**	74.82	135.60
3	1	84.99	76.42	68.62	62.06	39.24	74.76	135.57
4	1	82.50	73.45	65.96	59.92	38.20	73.10	130.17
1	2	84.87	76.10	68.86	62.93	39.58	74.19	134.66
2	2	84.90	76.59	69.25	63.15	39.65	74.40	134.94
3	2	85.21	76.38	69.17	63.34	39.70	74.41	135.07
4	2	85.12	77.09	69.80	63.75	39.00	73.83	132.59
1	3	85.80	77.32	69.80	63.32	39.57	74.42	134.89
2	3	85.78	77.06	69.42	63.23	40.04	**74.84**	**136.47**
3	3	83.54	74.62	67.41	61.70	39.14	73.77	132.45
4	3	84.98	76.77	69.08	62.88	39.17	73.98	132.62
1	4	84.71	76.24	69.02	63.25	39.34	74.10	133.66
2	4	85.34	77.30	70.08	64.01	39.91	74.95	135.66
3	4	85.21	77.04	69.78	63.69	39.33	73.90	133.72
4	4	85.00	76.58	68.91	62.44	39.04	73.18	130.56

Table 1: Performance of ESAN model at different depths on the LEVIR-CC dataset.

is used to predict the probability of each output word, which is expressed as Equation 9. Where W^T is the weight matrix, b_i is the bias term, h_i^N is the hidden layer state vector representation (the attention output of the i-th position), and p_i is the probability of the i-th word.

$$p_i = softmax\left(W^T h_i^N + b_i\right) \tag{9}$$

4 Experiments and Results

4.1 Datasets

We use LEVIR-CC and Dubai-CC datasets. The former provided by Liu et al. (Liu et al., 2022), which focuses on multiple changing scenes and objects. And the latter dataset, introduced by Hoxha et al. (Hoxha et al., 2022), offers a comprehensive description of urban transformation within the Dubai region. See Appendix A.1.1 for a detailed introduction.

4.2 Evaluation Metrics

Following the most advanced change description methods (Ji et al., 2023; Yu et al., 2022; Qiu et al., 2020; Tu et al., 2021a; Ak et al., 2023), we use four common indicators to evaluate the accuracy of all methods, namely BLEU-N (where N = 1,2,3,4) (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDEr-D (Vedantam et al., 2015). By comparing the consistency between the model output and the real ground

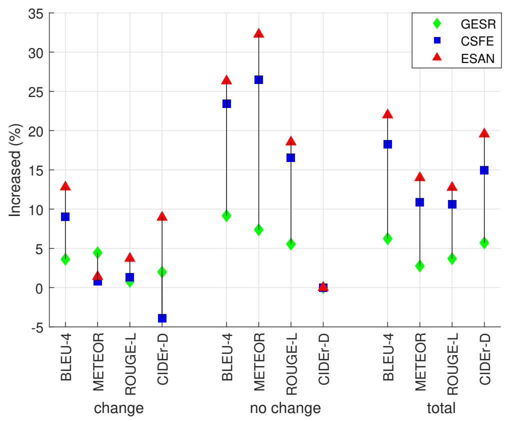

Figure 3: Ablation studies on LEVIR-CC.

reference data, these indicators provide a comprehensive assessment of the effect of the change description model. The higher the measurement score, the higher the similarity between the generated sentence and the reference sentence, that is, the higher the accuracy of the change description.

4.3 Experimental Details

The method based on the PyTorch framework is trained and evaluated on the NVIDIA A100 or V100. We use ResNet-101 (He et al., 2016) pretrained to extract image features. The dimension of the image features and the hidden state used in DG module is set to 1024. During training, we use the Adam optimizer (Kingma and Ba, 2015) with the learning rate of 0.0001. At the same time, the

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
LEVIR-CC							
DUDA (2019)	81.44	72.22	64.24	57.79	37.15	71.04	124.32
MCCFormer-S (2021)	79.90	70.26	62.68	56.68	36.17	69.46	120.39
MCCFormer-D (2021)	80.42	70.87	62.86	56.38	37.29	70.32	124.44
PSNet (2023a)	83.86	75.13	67.89	62.11	38.80	73.60	132.62
Chg2Cap (2023)	**86.14**	78.08	70.66	64.39	40.03	**75.12**	**136.61**
RSICCformer (2022)	84.72	76.27	68.87	62.77	39.61	74.12	134.12
Prompt-CC (2023b)	83.66	75.73	69.10	63.54	38.82	73.72	136.44
ESAN(Ours)	86.03	**78.14**	**70.87**	**64.86**	**40.10**	70.82	135.60
Average	↑ 3.88%	↑ 5.63%	↑ 6.61%	↑ 7.47%	↑ 4.92%	—	↑ 4.66%
Dubai-CC							
DUDA (2019)	58.82	43.59	33.63	25.39	22.05	48.34	62.78
MCCFormer-S (2021)	52.97	37.02	27.62	22.57	18.64	43.29	53.81
MCCFormer-D (2021)	64.65	50.45	39.36	29.48	25.09	51.27	66.51
RSICCformer (2022)	67.92	53.61	41.37	31.28	25.41	51.96	66.54
Chg2Cap (2023)	72.04	60.18	50.84	41.70	28.92	58.66	92.49
ESAN(Ours)	**73.56**	**61.62**	**52.44**	**42.89**	**30.02**	**60.72**	**99.84**
Average	↑ 17.62%	↑ 29.46%	↑ 41.79%	↑ 48.88%	↑ 27.76%	↑ 20.93%	↑50.54%

Table 2: Comparison with the state of the art.

training batch size is set to 32. After each epoch, the model is evaluated on the validation set, and the best performance model is selected according to the highest BLEU-4 score to evaluate the test set. We evaluate the performance of the model on the test set from the following three aspects: 1) the whole data set; 2) the data set only containing the image pairs with changes; 3) the data set only containing the image pairs without changes. For the data set only containing the image pairs with changes, the recognition accuracy and the sensitivity of the model to the changed area are reflected. For the data set only containing the image pairs without changes, there are some changes only in the interference factors. It is used to verify whether the model can correctly identify the interference factors in the image and provide meaningful description.

4.4 Ablation Studies

In order to clarify the contribution of each module of the network, we verify the overall performance of each block of the method by simultaneously testing the model performance under the changed image pairs and the unchanged image pairs. Baseline is without any module. The experimental results on LEVIR-CC are shown in Fig 3. In the overall data set performance, using GESR, the model has improved in all indicators, such as BLEU-4 increased

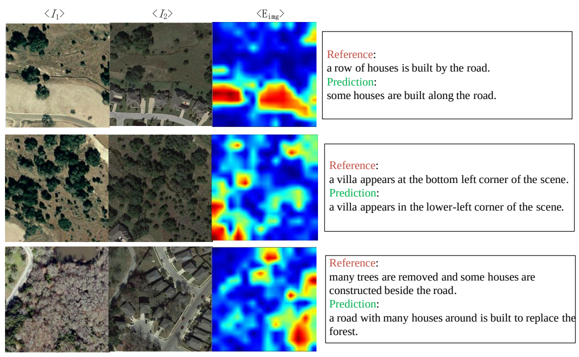

Figure 4: Case studies of our model on the LEVIR-CC dataset.

by 6.24% and CIDEr-D increased by 5.71%. Compared with the base model, after adding CSFE, BLEU-4, METEOR, ROUGE-L and CIDEr-D increased by 18.2%, 10.89%, 10.16% and 14.94%, respectively. Using GESR, CSFE, and the combination of the two are applicable. The results show that it is very effective to rely on GESR to obtain the global semantic information and use CSFE to capture the difference representation. The results of the same settings on Dubai-CC dataset are shown in Appendix A.1.2

4.5 Parameter Analysis

In order to evaluate the performance of the model at different depths, a series of experiments in Table 1 were performed. E.D represents the depth of the encoder, and D.D represents the depth of the decoder. When E.D = 2 and D.D = 1, the model exhibits outstanding performance. See appendix A.1.3 for other similar experiments.

4.6 Performance Comparison

In order to evaluate the relative advantages and disadvantages of our method in the remote sensing image change description task, the performance with other advanced change description methods is compared and the results are shown in Table 2.

The results show that ESAN performs better than other advanced methods in key indicators such as BLEU-1, BLEU-2, BLEU-3, BLEU-4 and METEOR, with an average increase of 3.88%, 5.63%, 6.61%, 7.47% and 4.92%, respectively on LEVIR-CC. Compared with Prompt-CC advanced method, the model shows superior performance, and the indicators of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 are improved 2.83%, 3.18%, 2.56% and 2.08%, respectively. And the key indicators of BLEU-2, BLEU-3, BLEU-4 and METEOR are higher than the recently excellent Chg2Cap, and the model shows competitive results. In general, ESAN performs better than other methods. The results on Dubai-CC dataset show that ESAN has achieved the best results on all indicators, with an average increase of 17.62%, 29.46%, 41.79%, 48.88%, 27.76%, 20.93% and 50.54%, respectively. BLEU-4 increased to 42.89, METEOR increased to 30.02, ROUGE-L increased to 60.72, and CIDEr-D increased to 99.84. Compared with the recently outstanding Chg2Cap, EASN is 2.85%, 3.80%, 3.51% and 7.95% higher on BLEU-4, METEOR, ROUGE-L and CIDEr-D, respectively. It fully demonstrates that our network can use the semantic relationship to generate a description closer to the reference sentence.

4.7 Qualitative evaluation

In order to evaluate the quality of the change descriptions generated by our model, we visualize the image embedding and the predicted change description generated by the description decoder, as shown in Fig 4 and Fig 5, where I_1 and I_2 represent the images captured at time 1 and time 2, respectively. E_{img} is the image embedding and E_{diff} is the difference image embedding extracted by the semantic relation embedding encoder.

As shown in Fig 4 and Fig 5, we can see that the difference captions generated by ESAN can accurately locate the change area and highlight it. At the same time, in the case of image pairs invariant, the network focuses on identifying invariant objects. Taking the last pair of images in Fig 5 as an example, we can see that the scene interference is very large. Compared with the first standard description, our model not only successfully describes the changing target, namely "residence", but also describes a more advanced scene concept, namely "desert". This is because ESAN uses the global semantic information to more fully understand and describe objects in the entire image and their relationships in the scene. It demonstrates the ability of our model to accurately locate and describe the differences from noisy real world environments.

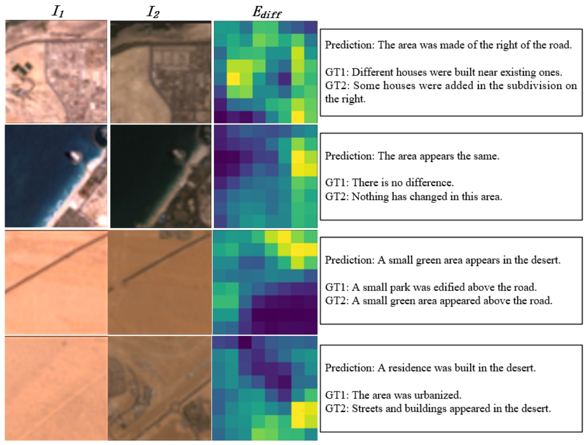

Figure 5: Case studies of our model on the Dubai-CC dataset.

5 Conclusion

In this paper, we propose an efficient semantic attention network (ESAN). The network has significant advantages in fully understanding the internal semantic information of the image by efficiently obtaining the semantic relationship between image features. In addition, the network can effectively identify and ignore interference factors. Therefore, it is good at accurately representing image changes and generating descriptions with rich semantics.

Limitations

We propose a new remote sensing image change description method, ESAN. Although it has been verified the performance on the general datasets, through the observation of relevant visualization cases and the analysis of the generated change description statements, it is found that the change description statements are not perfect in some logical expressions and still need to be further optimized.

In addition, with the increase of the sample size of the experimental data set, how to further optimize the model for large-scale remote sensing image data is also the direction of our future research.

References

Kenan E. Ak, Ying Sun, and Joo Hwee Lim. 2023. Learning by imagination: A joint framework for text-based image manipulation and change captioning. *IEEE Trans. Multim.*, 25:3006–3016.

Oluwasanmi Ariyo, Muhammad Umar Aftab, Eatedal Alabdulkreem, Bulbula Kumeda, Edward Yellakuor Baagyere, and Zhiquang Qin. 2019a. Captionnet: Automatic end-to-end siamese difference captioning model with attention. *IEEE Access*, 7:106773–106783.

Oluwasanmi Ariyo, Enoch Frimpong, Muhammad Umar Aftab, Edward Yellakuor Baagyere, Zhiguang Qin, and Kifayat Ullah. 2019b. Fully convolutional captionnet: Siamese difference captioning attention model. *IEEE Access*, 7:175929–175939.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Shizhen Chang and Pedram Ghamisi. 2023. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Trans. Image Process.*, 32:6047–6060.

Hao Chen and Zhenwei Shi. 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote. Sens.*, 12(10):1662.

Pablo Pozzobon de Bem, Osmar Abílio de Carvalho Júnior, Renato Fontes Guimarães, and Roberto Arnaldo Trancoso Gomes. 2020. Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks. *Remote. Sens.*, 12(6):901.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Mehrdad Hosseinzadeh and Yang Wang. 2021. Image change captioning by learning from an auxiliary task. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2725–2734. Computer Vision Foundation / IEEE.

Genc Hoxha, Seloua Chouaf, Farid Melgani, and Youcef Smara. 2022. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–14.

Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. 2022. Image difference captioning with instance-level fine-grained feature representation. *IEEE Trans. Multim.*, 24:2004–2017.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7166–7176. Association for Computational Linguistics.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 603–612. IEEE.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4024–4034. Association for Computational Linguistics.

Jiayi Ji, Xiaoyang Huang, Xiaoshuai Sun, Yiyi Zhou, Gen Luo, Liujuan Cao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. 2023. Multi-branch distance-sensitive self-attention network for image captioning. *IEEE Trans. Multim.*, 25:3962–3974.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021. Viewpoint-agnostic change captioning with cycle consistency. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2075–2084. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Trans. Image Process.*, 31:2726–2738.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chenyang Liu, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2023a. Progressive scale-aware network for remote sensing image change captioning. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2023, Pasadena, CA, USA, July 16-21, 2023*, pages 6668–6671. IEEE.

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. 2022. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–20.

Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2023b. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–18.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4623–4632. IEEE.

Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. 2020. 3d-aware scene change captioning from multiview images. *IEEE Robotics Autom. Lett.*, 5(3):4743–4750.

Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. 2021. Describing and localizing multiple changes with transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1951–1960. IEEE.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq R. Joty, and Jianfei Cai. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 574–590. Springer.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1873–1883. Association for Computational Linguistics.

Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. 2022. I^2transformer: Intra- and inter-relation embedding transformer for TV show captioning. *IEEE Trans. Image Process.*, 31:3565–3577.

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. 2023a. Neighborhood contrastive transformer for change captioning. *IEEE Trans. Multim.*, 25:9518–9529.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021a. R^3net: Relation-embedded representation reconstruction network for change captioning. *CoRR*, abs/2110.10328.

Yunbin Tu, Tingting Yao, Liang Li, Jiedong Lou, Shengxiang Gao, Zhengtao Yu, and Chenggang Yan. 2021b. Semantic relation-aware difference representation learning for change captioning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 63–73. Association for Computational Linguistics.

Yunbin Tu, Chang Zhou, Junjun Guo, Huafeng Li, Shengxiang Gao, and Zhengtao Yu. 2023b. Relation-aware attention for video captioning via graph learning. *Pattern Recognit.*, 136:109204.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Aming Wu, Linchao Zhu, Yahong Han, and Yi Yang. 2019. Connective cognition network for directional visual commonsense reasoning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5670–5680.

Joseph Z. Xu, Wenhan Lu, Zebo Li, Pranav Khaitan, and Valeriya Zaytseva. 2019. Building damage detection in satellite imagery using convolutional neural networks. *CoRR*, abs/1910.06444.

Linli Yao, Weiying Wang, and Qin Jin. 2022. Image difference captioning with pre-training and contrastive learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3108–3116. AAAI Press.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3025–3035. Association for Computational Linguistics.

Litao Yu, Jian Zhang, and Qiang Wu. 2022. Dual attention on pyramid feature maps for image captioning. *IEEE Trans. Multim.*, 24:1775–1786.

Shengbin Yue, Yunbin Tu, Liang Li, Ying Yang, Shengxiang Gao, and Zhengtao Yu. 2023. I3N: intra- and inter-representation interaction network for change captioning. *IEEE Trans. Multim.*, 25:8828–8841.

A Appendix

A.1 Additional Experimental Setup

A.1.1 Datasets

LEVIR-CC is composed of 10,077 small bitemporal tiles with a size of 256 × 256 pixels, and each tile is annotated as containing changes or not containing changes. Among them, there are 5038 image pairs with changes and 5039 image pairs without changes. Each image pair is composed of five different sentence descriptions, and the length of most sentences is between 5 and 15 words. In the experiment, the data set is divided into training set, validation set and test set, including 6815, 1333 and 1929 image pairs respectively. The original images in Dubai-CC dataset have been trimmed into 500 tiles of sizes 50 × 50, with five change descriptions annotated for each small bitemporal tile. In the course of the experiments, the dataset has been divided into three parts: training, validation, and testing sets, comprising 300, 50, and 150 bitemporal tiles, respectively. The images were enlarged to dimensions of 256 × 256 pixels prior to being processed by the network.

A.1.2 Ablation Experiment

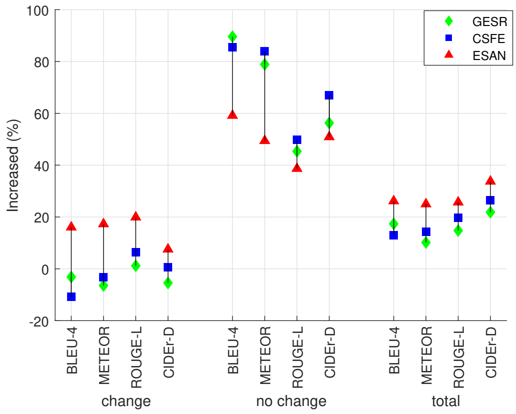

Figure 6: Ablation studies on Dubai-CC.

A.1.3 Model Parameter Comparison Experiment

E.D	**D.D**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr-D**
1	1	72.22	59.9	51.24	42.99	30.57	**62.35**	**102.09**
2	1	**73.97**	**62.10**	**53.26**	**45.37**	**30.83**	60.04	97.59
3	1	66.44	53.28	43.72	36.46	25.67	54.23	80.49
4	1	71.51	57.90	48.85	41.05	29.05	57.59	88.39
1	2	60.34	50.90	43.54	36.86	24.41	51.09	76.62
2	2	63.32	50.11	42.24	37.80	23.63	53.24	81.45
3	2	70.64	59.03	49.55	41.32	29.14	58.08	91.22
4	2	65.80	54.45	46.24	40.11	26.81	55.63	86.65
1	3	69.57	57.66	47.90	39.69	28.38	56.17	79.96
2	3	59.73	51.09	44.96	39.68	25.19	52.33	89.56
3	3	64.33	52.45	43.59	36.21	23.82	53.47	78.71
4	3	64.03	50.85	42.27	33.91	26.62	50.92	68.48
1	4	69.80	54.35	44.04	35.55	26.30	53.65	70.66
2	4	62.86	54.08	47.37	39.47	26.28	53.98	85.17
3	4	60.53	45.74	37.70	33.40	22.30	50.59	74.18
4	4	64.12	52.77	44.79	35.73	24.81	55.16	74.29

Table 3: Performance of ESAN model at different depths on the Dubai-CC dataset.