MSPN: Multiple Semantics Perception Network for Remote Sensing Change Captioning

Anonymous ACL submission

Abstract

Remote sensing images usually cover a large surface area, so the change information is usually difficult to be precisely localized. Especially, some changes are easy to be overlooked due to their inconspicuous locations and fuzzy shapes. In addition, unlike the natural image change description task, the remote sensing image change description task aims to capture the most significant changes without various influencing factors, such as light, seasonal influences and complex land cover. To address the above challenges, in this paper, we propose a multiple semantic perception network (MSPN) model to extract more accurate feature representations to guide the decoder in generating high-quality change descriptions. In the visual encoder stage, the global efficient semantic awareness module is designed for global feature embedding, the self-semantic awareness module digs deep into the internal connections between features, and the change semantic interaction module effectively distinguishes semantic changes from irrelevant ones. In the description generation phase, the Transformer-based decoder is designed to guide the change description generation. Extensive experiments on the LEVIR-CC dataset demonstrate the superiority of the MSPN model over many state-of-the-art techniques.

1 Introduction

With the rapid development of remote sensing technology, a large amount of high-resolution remote sensing image data has been acquired. The research on remote sensing images is 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 remote sensing image changes not only help to improve the image interpretation capability, but also make these images easier to be understood by non-specialized users, which provides a powerful tool to support decision-making, planning and 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. This task 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.

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. 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. 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) (Oluwasanmi et al., 2019a). Recent work has further explored the relationship-aware attention mechanism (Tu et al., 2023b), distance-sensitive self-attention (DSA) (Ji et al., 2022), 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., 2023) aim to optimize the overall performance of the model. These 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.

Currently, for the task of describing changes in remote sensing images, the key challenges are mainly in the following aspects: Firstly, the model lacks fine-grained semantic understanding because remote sensing images usually cover large surface areas, and change information is usually difficult to pinpoint. There are a number of changes that are usually easily overlooked due to their inconspicuous locations and ambiguous shapes. Secondly, the model still lacks resistance to confounding factors, and it is difficult to produce descriptions that involve only real semantic changes. This means that the model should be able to filter out noise, lighting variations, or other environmental factors that are not relevant to the change and focus on capturing the key semantic information in the image that reflects the actual surface or scene change. This immunity to perturbation makes the model more reliable for real-world applications.

To address the above challenges, we propose a Multi-Semantic Perceptual Network (MSPN), which utilizes different semantic relation modules and a transformer-based decoder for remote sensing change description generation. The contributions of this paper are summarized as follows:

(1) A multi-semantic perceptual network is proposed. Firstly, the global efficient semantic perception module operates at the perceptual level to grasp the global correlation information. Subsequently, the self-semantic awareness module digs deep into the internal feature association and enhances the understanding of subtle differences. On this basis, the change semantic interaction module carefully examines the comparative information between features, with particular attention to representing differences. Finally, the decoder translates the learned change features into natural language sentences.

(2) Comprehensively compare and analyze the effects of the encoder-extracted image feature representations of semantic relation embeddings in the description generation phase. By performing the analysis and evaluation of model parameters, we provide insights that may inspire researchers to design more effective models to fully utilize bi-chronological image features.

(3) Extensive experiments show that our method outperforms other state-of-the-art methods on the LEVIR-CC dataset.

2 Methodology

2.1 Overall Architecture of MSPN 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: $f(I_1, I_2; \theta) \rightarrow \hat{C}$, 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 four parts : (1) The global efficient semantic awareness module quickly captures the global semantic information of the image from two different directions ; (2) The self-semantic awareness module captures internal semantic information between all features of the same input ; (3) The change semantic interaction module is responsible for the information flow interaction between different scale features, and learns the contrast information between them, so as to pay attention to the semantic information of actual changes ; (4) The Transformer-based language decoder translates the learned change features into natural language sentences.

The proposed method follows encoder-decoder architecture for change description generation of remote sensing images. In the following, we give the details of visual feature extractor and description generation.

2.2 Semantic Relation Embedding

2.2.1 Global Efficient Semantic Awareness (GESA)

Given a dual-temporal image pair (I_1, I_2), we first use the pre-trained Resnet101 model to extract image features and represent them as X_1, X_2, respectively, where, $\mathrm{X}_i \in R^{C\times H\times W}, C, H, W$ 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, captur-

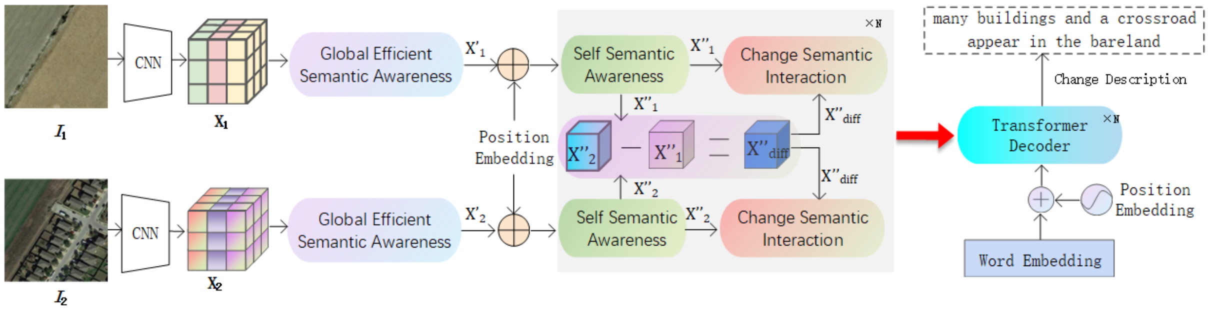

Figure 1: Overall architecture of our MSPN model.

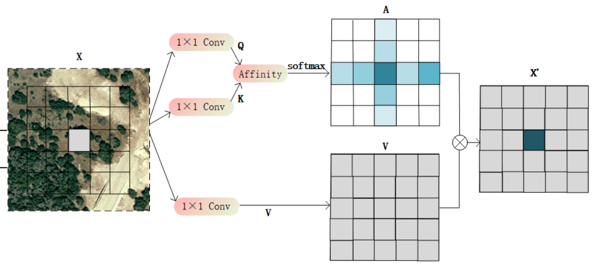

Figure 2: Detailed introduction of the global efficient semantic awareness module.

Algorithm 1. Procedure of Our MSPN

Input: (I_1, I_2) (a pair of bi-temporal images)
Output: Change Description
Define:
GESA: global efficient semantic awareness module
SSA: the self-semantic awareness module
CSI: the change semantic interaction module
DG: the description generation decoder
EM: the word embedding
1. // step1: Feature Extraction
2. for i in (I_1, I_2) do:
3. X_i = backbone (I_i)
4. end for
5. // step2: Semantic Relation Embedding
6. X_1', X_2' = GESA (X_1, X_2)
7. for n in (1 - N) do:
8. X_1'' = SSA (X_1', X_1', X_1')
9. X_2'' = SSA (X_2', X_2', X_2')
10. X_{diff}'' = X_2'' - X_1''
11. $\widetilde{X_1}$ = CSI $(X_1'', X_{diff}'', X_{diff}'')$
12. $\widetilde{X_2}$ = CSI $(X_2'', X_{diff}'', X_{diff}'')$
13. $\hat{X}_{diff}$ = $[\widetilde{X_1};\widetilde{X_2}]$
14. end for
15. // step3: Description Generation
16. Description = EM (“start”)
17. while w EM (“end”) do:
18. w = DG $(\hat{X}_{diff}$, Description)
19. Description [Description; w]
20. end while
21. return Description

Table 1: The processing procedure of our MSPN is shown in Algorithm 1.

ing 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. 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 memory. Inspired by the Criss-Cross attention used in semantic segmentation (Huang et al., 2019), the Global Efficient Semantic Awareness (GESA) module relies on it to implicitly model the global semantic relationships in each image.

As shown in Figure 2, 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 Eq. (1) and softmax layer.

$$d_{i,p} = Q_p \Omega_{i,p}^T \tag{1}$$

Where $d_{i,p} \in D$ is the degree of correlation between characteristic Q_p and $\Omega_{i,p}$, $i = [1, \ldots, H+W-1]$, $D \in R^{(H+W-1)\times(H\times W)}$.

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 follows:

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

Where, X_p is the eigenvector of position p in $X' \in R^{C\times H\times W}$.

After 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 to enhance the image feature representation.

2.2.2 Hierarchical Semantic Representation of Interaction

Through self-semantic representation, the model processes the input and deeply mines the feature representation of the image sequence, so as to understand the information in the input sequence more comprehensively. Further, the change semantic interaction is used to reveal the change characteristics, so that the model can effectively locate semantic changes without being affected by irrelevant changes. This special architecture design provides strong semantic coherence for the model, which enables it to accurately and comprehensively capture the key semantic information in the image sequence.

Self-Semantic Awareness (SSA) In order to better capture the semantic relationship between objects in image features, the self-semantic relationship perception module is used to explore the features between image pairs. We first apply two multi-head self-attention. Different from the previous work that directly uses the features obtained from the backbone deep neural network, the attention layer can establish an internal connection between all features of the same scale. In addition, since the features are down-sampled through the backbone network, the computational complexity of the model is relatively low.

Specifically, we first transform the existing feature $X'_i \in R^{C\times H\times W}$ into $X'_i \in R^{C\times N}$, where N=HW, $i \in (1,2)$. Then, Q, K, V are embedded into the same-dimensional embedding by Emb method. The SSA module is represented as follows:

$$(Q,K,V) = \left(X'_i W_i^Q, X'_i W_i^K, X'_i W_i^V\right) \tag{3}$$

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

$$SSA\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Where d_k is the dimension of the vector.

When the model can deeply grasp the comprehensive information in the image, it can better distinguish semantic changes from unrelated changes. That is to say, the result of the self-semantic awareness module is used as the input of the change semantic interaction stage, which effectively constructs the relationship between the image sequence features, which is the basis for obtaining reliable difference representation in the change semantic interaction stage.

Change Semantic Interaction (CSI) The self-semantic awareness module embeds the semantic relationship between all the features of the same input into the features X'_1 and X'_2, we get X_1'' and X_2'', 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). Inspired by multi-headed cross-attention (Vaswani et al., 2017), based on the feature representations X_1'', X_2'', X''_{diff} , the change semantic interaction module is defined as follows:

$$(Q, K, V) = \left(X_i''W_i^Q, X_{diff}''W_i^K, X_{diff}''W_i^V\right) \tag{5}$$

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

$$CSI\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Where d_k is the dimension of the vector.

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}'' as follows:

$$\widetilde{X_1} = CSI\left(X_1'', X_{diff}'', X_{diff}''\right) \tag{7}$$

$$\widetilde{X_2} = CSI\left(X_2'', X_{diff}'', X_{diff}''\right) \tag{8}$$

Then, in order to reduce the loss of image feature information, the difference information is accurately judged while highlighting the irrelevant information. By splicing and integrating them, the stable difference representation in each image pair is learned:

$$\hat{X}_{diff} = LN\left(\left[\widetilde{X}_1; \widetilde{X}_2\right]\right) \tag{9}$$

Where LN is the abbreviation of Layer Normalization (Ba et al., 2016).

2.3 Description Generation (DG)

In this part, we use the Transformer (Huang et al., 2019) decoder 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, a multi-head cross-attention layer and a forward propagation layer. Now we represent the visual sequence obtained from the visual encoder as $\widetilde{V_I}$.

Firstly, the description decoder takes each word as input, and the masked multi-head attention mechanism embeds the word through Eq. (10):

$$E\,[W] = \{E\,[w_1]\,, \ldots; E\,[w_m]\} \tag{10}$$

And the embedding feature $\hat{E}\,[W]$ is calculated. Then, through multi-head 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.

After stacking N Transformer decoding blocks, the hidden layer state output of the last block h^N is used to predict the probability of each output word, which is expressed as follows:

$$p_i = softmax\left(W^T h_i^N + b_i\right) \tag{11}$$

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.

3 Experiments and Results

3.1 Experimental Setup

3.1.1 Datasets

The data set used in the experiment is the LEVIR-CC data set provided by Liu et al. (Liu et al., 2022), which is tailored from the building change detection data set LEVIR-CD (Chen and Shi, 2020). Unlike LEVIR-CD, which only focuses on building-related changes, the LEVIR-CC dataset focuses on multiple changing scenes and objects. LEVIR-CC is composed of 10,077 small bi-temporal 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.

3.1.2 Evaluation Metrics

In this work, we followed the most advanced change description methods (Yu et al., 2021), (Ji et al., 2022), (Qiu et al., 2020), (Tu et al., 2021), (Ak et al., 2023) and used 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 (ROUGE, 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 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.

3.1.3 Experimental Details

In this paper, the proposed deep learning method based on the PyTorch framework is trained and evaluated on the NVIDIA A100 graphics processing unit. During training, on the LEVIR-CC dataset, we use the Adam optimizer (Kingma, 2015) to minimize the negative log-likelihood loss of the equation. At the same time, the initial learning rate is set to 0.0001, and the 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 train the model on the same training set, and then 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. It is used to verify the adaptability of the model to change detection and description generation. For the data set only containing the image pairs without changes, there are some changes only in the interference factors, such as seasonal changes and illumination changes. It is used to verify whether the model can correctly identify the interference factors in the image and provide meaningful description. The ability of the model to deal with irrelevant regions can be examined. In addition, we did not report the CIDEr-D measure of the test model in this case because the unchanged words are monotonous, CIDEr-D will approach 0. Therefore, in this case, CIDEr-D cannot measure the accuracy of sentences.

3.2 Ablation Studies

In order to clarify the contribution of each module of the proposed network, we conducted the following ablation studies on LEVIR-CC. We verify the overall performance of each block of the proposed method by simultaneously testing the model performance under the changed image pairs and the unchanged image pairs. And in the case of different test sets, the experimental results are all shown in Table 2.

It can be seen from Table 2:

1) Compared with Baseline model, GESA model has improved in all indicators. Among them, the BLEU-4 value increased by 6.24%, and the CIDEr-D value increased by 5.71%. It shows that after extracting image features in ResNet101 network, only relying on the global efficient semantic perception module to obtain the global semantic information between samples can improve the model, which proves the effectiveness of GESA module.

Method	B-4	M	R	C
Baseline	53.17	35.18	66.36	113.42
	35.35	24.99	51.72	57.35
	74.10	55.16	80.97	-
GESA	56.49	36.15	68.81	119.90
	36.63	25.20	52.14	58.49
	80.89	59.23	85.46	-
SSA+CSI	62.87	39.01	73.40	130.36
	38.55	25.34	52.40	55.12
	91.47	69.77	94.39	-
GESA+SSA+CSI	64.86	40.10	74.82	135.60
	39.88	26.10	53.64	62.48
	93.60	72.96	95.98	-

Table 2: Ablation studies on LEVIR-CC in terms of total performance, the change setting and the no-change setting, respectively. Where B-4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L, and CIDEr-D, respectively.

2) Using both SSA and CSI, the performance of the model is significantly improved. It shows that SSA can be well combined with CSI to improve the quality of model generation description. Compared with the baseline model, after adding hierarchical semantic interaction representation, B-4 increased by 18.24%, METEOR increased by 10.89%, ROUGE-L increased by 10.61%, and CIDEr-D increased by 14.94%.

3) After combining GESA with SSA + CSI, each evaluation index is improved again, and the CIDEr-D index representing the similarity between the generated description and the reference description reaches 135.60. It shows that the combination of the proposed modules can assist the model to generate a higher quality description, which also proves the superiority of each module.

According to the data of Table 2, we can come to the following conclusions:

1) This overall evaluation performance verifies the generalization ability of this method, that is, it can not only accurately determine whether there is a semantic change between image pairs, but also can ignore the interference factors to accurately describe the change.

2) It is very effective to capture the difference representation by SSA + CSI. Because it establishes an internal relationship between all the features of the same input, semantic changes and irrelevant changes can be effectively distinguished by semantic information flow interaction.

3) Capturing the semantic relationships between image features is very important, because these relationships can enrich the original object features and help to explore fine-grained changes.

3.3 Performance Comparison

In order to comprehensively and objectively evaluate the relative advantages and disadvantages of the proposed 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 3.

The experimental results in Table 3 show that our MSPN shows good performance compared with other methods. MSPN outperforms all other methods in all indicators. Among them, it increased the BLEU-4 to 64.86 and the CIDEr-D to 135.60. It fully shows that MSPN can make use of the semantic relationship between image features, so the model can obtain good performance and generalization ability.

3.4 Qualitative Evaluation

In order to evaluate the quality of the change descriptions generated by our proposed MSPN model, we conducted a qualitative evaluation by selecting several representative scenarios from the LEVIR-CC dataset. We visualize the image embedding and the predicted change description generated by the description decoder, as shown in Figure 3, where I_1 and I_2 represent the images captured at time 1 and time 2, respectively, and E_{img} is the visual image embedding extracted by the semantic relation embedding encoder.

By observing the visual image embedding, our network can accurately locate the change area and highlight it. At the same time, in the example of unchanged image pairs, the network focuses on identifying unchanged objects. It shows that our network can accurately highlight changes in high-resolution dual-time images, allowing the decoder to generate a more accurate description.

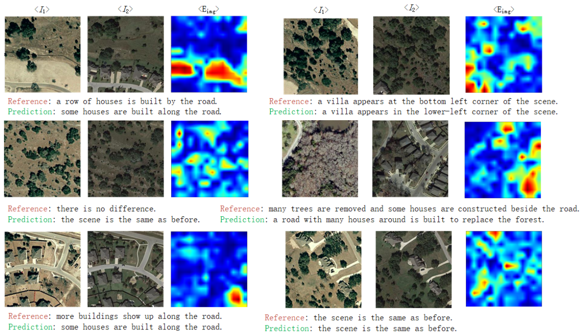

Figure 3: An example of visual image embedding and change description generated by MSPN in LEVIR-CC dataset.

4 Related Work

4.1 Image Captioning

Li et al. (2022) introduced the long-short-term relational converter (LSRT) to fully understand the relationship between objects. On the other hand, Tu et al. (2022) proposed an internal and relational embedding transformer (I^2Transformer), which makes full use of various modalities through the enhancement of cross-modal information. In the task of image caption generation, Yu et al. (2021) applied the dual attention mechanism to the pyramid feature map, so as to better locate the regions in the image. 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. (2022) introduced distance-sensitive self-attention (DSA). 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 with two kinds of graph learning.

4.2 Change Captioning

Jhamtani and Berg-Kirkpatrick (2018) made a pioneering contribution to this field. Subsequently, Park et al. (2019) introduced the Double Dynamic Attention Model (DUDA). In order to solve the common 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. Subsequently, Qiu et al. (2021) introduced the multi-change caption trans-

Method	B-1	B-2	B-3	B-4	M	R	C
DUDA (Park et al., 2019)	81.44	72.22	64.24	57.79	37.15	71.04	124.32
MCCFormer-S (Qiu et al., 2021)	79.90	70.26	62.68	56.68	36.17	69.46	120.39
MCCFormer-D (Qiu et al., 2021)	80.42	70.87	62.86	56.38	37.29	70.32	124.44
PSNet (Liu et al., 2023)	83.86	75.13	67.89	62.11	38.80	73.60	132.62
RSICCformer (Liu et al., 2022)	84.72	76.27	68.87	62.77	39.61	74.12	134.12
MSPN (Ours)	**86.03**	**78.14**	**70.87**	**64.86**	**40.10**	**74.82**	**135.60**

Table 3: Comparisons experiments on the LEVIR-CC dataset. Where B-1, B-2, B-3, B-4, M, R, and C are short for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr-D, respectively. The bold numbers are the best performance.

former (MCCFormers), and began to pay attention to the changes at the semantic level. Tan et al. (2019) elaborated on the editing transformation between two images, highlighting the differences in semantics. Further, Oluwasanmi et al. (2019b) proposed a fully convolutional CaptionNet (FCC). Through the multi-modal end-to-end connected difference caption model (SDCM), (Oluwasanmi et al., 2019a) captured, aligned, and calculated the differences between the two image features, which enhanced the understanding of the semantic level feature differences. 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 (I^3N), which focuses on learning fine differential representation. In order to make the changing caption model capture the actual changes, Kim et al. (2021) proposed a view-independent changing subtitle network with cyclic consistency (VACC). Facing the challenges in the Image Difference Captioning (IDC) task, Yao et al. (2022) proposed a new modeling framework to learn stronger visual and linguistic associations. Liu et al. (2023) introduced a progressive scale-aware network (PSNet). And Huang et al. (2021) proposed an instance-level fine-grained differential captioning (IFDC) model, which focuses on the rich explicit features of the object to solve the challenge of accurately locating the changing object in the context.

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.

5 Conclusion

In this paper, we propose a multi-semantic relationship perception network (MSPN). The network has significant advantages in fully understanding the internal semantic information of the images by obtaining a variety of semantic relationships. 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. Extensive experiments on the LEVIR-CC dataset show that the proposed method achieves state-of-the-art results.

6 Limitations

Although the proposed multiple semantic perception network can deeply understand various semantic relations in images, there are still some limitations for fine-grained semantic changes. Moreover, when dealing with interference factors in complex scenes, the method still has some limitations. In addition, the limitations of the experimental data set and the applicability and versatility of the method also need to be further verified and explored. Therefore, future research should focus on further improving the accuracy and robustness of semantic methods to more comprehensively analyze and describe changes in remote sensing images to meet the demands for higher standards of detail, diversity and complexity.

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 Transactions on Multimedia*, 25:3006–3016.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *stat*, 1050:21.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Shizhen Chang and Pedram Ghamisi. 2023. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 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 Sensing*, 12(10):1662.

Pablo Pozzobon De Bem, Osmar Abílio de Carvalho Junior, 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 Sensing*, 12(6):901.

Mehrdad Hosseinzadeh and Yang Wang. 2021. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2725–2734.

Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. 2021. Image difference captioning with instance-level fine-grained feature representation. *IEEE transactions on multimedia*, 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*, pages 7166–7176.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612.

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*, pages 4024–4034.

Jiayi Ji, Xiaoyang Huang, Xiaoshuai Sun, Yiyi Zhou, Gen Luo, Liujuan Cao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. 2022. Multi-branch distance-sensitive self-attention network for image captioning. *IEEE Transactions on Multimedia*.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2095–2104.

DP Kingma. 2015. Adam: a method for stochastic optimization. *arXiv: 1412.6980*.

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 Transactions on Image Processing*, 31:2726–2738.

Chenyang Liu, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2023. Progressive scale-aware network for remote sensing image change captioning. *arXiv e-prints*, pages arXiv–2303.

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 Transactions on Geoscience and Remote Sensing*, 60:1–20.

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

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

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*, pages 311–318.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633.

Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. 2020. 3d-aware scene change captioning from multiview images. *IEEE Robotics and Automation Letters*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, volume 5.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq 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 16*, 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 Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883.

Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. 2022. I2transformer: Intra- and inter-relation embedding transformer for tv show captioning. *IEEE Transactions on Image Processing*, 31:3565–3577.

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. 2023a. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021. R^3net: Relation-embedded representation reconstruction network for change captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9319–9329.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Aming Wu, Linchao Zhu, Yahong Han, and Yi Yang. 2019. Connective cognition network for directional visual commonsense reasoning. *Advances in Neural Information Processing Systems*, 32.

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

Linli Yao, Weiying Wang, and Qin Jin. 2022. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3108–3116.

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*, pages 3025–3035.

Litao Yu, Jian Zhang, and Qiang Wu. 2021. Dual attention on pyramid feature maps for image captioning. *IEEE Transactions on Multimedia*, 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 Transactions on Multimedia*.

A Parameter Analysis

The depth parameters of the network can have a significant impact on the accuracy of the generated image description. Usually, the optimal depth parameter is determined by experiments to obtain the best performance on specific tasks. In this section, in order to evaluate the performance of the proposed MSPN model at different depths on the LEVIR-CC dataset, a series of experiments in Table 4 were performed. In the quantization results of Table 4, E.D represents the depth of the encoder, and D.D represents the depth of the decoder. We observed that the model performed best when E.D = 2 and D.D = 1.

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 4: Performance of MSPN model at different depths on the LEVIR-CC dataset.