Scene Graph and Dependency Grammar Enhanced Remote Sensing Change Caption Network (SGD-RSCCN)

Qiaoli Sun
College of Computer Science
Inner Mongolia University
Hohhot 010021, China
32209089@mail.imu.edu.cn

Yan Wang[*]
College of Computer Science
Inner Mongolia University
Hohhot 010021, China
cswy@imu.edu.cn

Xiaoyu Song
Computer Engineering
Portland State University
Portland 97207, USA
songx@pdx.edu

Abstract

With the continuous advancement of remote sensing technology, it is easier to obtain high-resolution, multi-temporal and multi-spectral images. The images carry rich information of ground objects. However, how to effectively extract useful information from the complex image data and convert it into understandable semantic descriptions remains a challenge. To deal with the challenges, we propose a Scene Graph and Dependency Grammar Enhanced Remote Sensing Change Caption Network (SGD-RSCCN) to improve the accuracy and naturalness of extracting and describing change information from remote sensing images. By combining advanced visual analysis technology and natural language processing technology, the network not only optimizes the problem of insufficient understanding of complex scenes, but also enhances the ability to capture dynamic changes, thereby generating more accurate and smooth natural language description. In addition, we also proposes the decoder based on prior knowledge, which further improves the readability and comprehensibility of the description. Extensive experiments on LEVIR-CC and Dubai-CC datasets verify the advantages of the proposed method in generating accurate and true descriptions.

1 Introduction

Remote sensing images provide valuable data resources for surface monitoring and environmental analysis due to the unique perspective and coverage. With the rapid development of remote sensing technology, a large number of high-resolution remote sensing image data have been obtained. Remote sensing images are not only used in scientific research, but also widely used in disaster assessment (Xu et al., 2019), urban planning (Chen and Shi, 2020), environmental monitoring (de Bem et al., 2020) and other fields. The accurate and semantically rich description of the image changes not only helps to improve the ability of image interpretation, making remote sensing images easier to be understood by non-professional users, but also provides a powerful tool for supporting decision-making, planning and management, and disaster response.

The remote sensing image change description task aims to describe the change content in remote sensing image pairs in natural language. The task involves two remote sensing images, usually corresponding to different time points in the same area. The model needs to understand the differences between the two images, including feature changes, new or disappeared elements, and generate text descriptions that can clearly express these changes. Due to the ability to extract high-level semantic information about changes in ground objects, change description has recently received attention in the field of geosciences and remote sensing.

Automatic analysis and interpretation of remote sensing images has important application value in many fields. In recent years, a variety of methods have been proposed to improve the performance of image change description models.

(Jhamtani and Berg-Kirkpatrick, 2018) proposed the first task to describe the difference between similar image pairs. To deal with the fact that significant differences are usually described at the object level rather than the pixel level, visual analysis is first performed to expose different pixel groups as agents of object-level differences. To emphasize the importance of using natural language to identify and describe important scene changes in the presence of distractions, (Park et al., 2019) proposes a dual dynamic attention model (DUDA) to learn to distinguish distractions and semantic changes. Since there are usually perspective changes in practice, which may overwhelmingly describe the semantic differences to be described, (Shi et al., 2020)

[*] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 2121–2130
January 19–24, 2025. ©2025 Association for Computational Linguistics

proposes a perspective adaptive matching code to clearly distinguish the perspective changes and semantic changes in the change description task. Different from the latest methods that mainly focus on the image change description task of the new model architecture, (Hosseinzadeh and Wang, 2021) proposes a new training scheme for the image change description task. In order to describe multiple changes in complex scenes, (Qiu et al., 2021) proposes a multi-change title converter (MCCFormers), which identifies the change region by associating different regions in the image pair, and dynamically determines the change region associated with the words in the sentence.

Although the above work has made significant progress, most of the work focuses on learning accurate change representations to generate descriptions, and the grammatical structure of natural language is often ignored when visual and linguistic features are fused, resulting in the generated descriptions may appear unsmooth in grammar although they are semantically correct. In addition, the lack of understanding of complex scenes and the difficulty of capturing dynamic changes are also challenges for existing methods.

To cope with the challenges, we propose a Scene Graph and Dependency Grammar Enhanced Remote Sensing Change Caption Network (SGD-RSCCN), which combines visual features and linguistic features by making full use of attention mechanism and scene graph construction technology to generate more accurate and informative image change descriptions. It can not only improve the interpretability of remote sensing image data, but also provide support for decision-making in related fields, which has important research significance and application value. Through extensive experiments on LEVIR-CC and Dubai-CC datasets, we demonstrate that the proposed method can generate more accurate and realistic descriptions of changes between remote sensing image pairs, and achieve superior performance compared with existing change description methods.

The contributions of the paper are summarized as follows : (1) Effective sequence selection (ESS): Aiming at the problem of insufficient understanding of complex scenes, we propose a new scene graph construction method, which can effectively deal with various features and complex interrelationships in remote sensing images. In addition, to capture the dynamic process changing with time, the method is optimized in the capture and representation of dynamic changes, thus improving the accuracy and practicability of remote sensing image change description.

(2) Decoder with syntax knowledge (DSK): By introducing dependency grammar analysis, we enhance the application of grammar rules in the training process. The dependency grammar analysis reveals the dependency relationship between words in a sentence and guides the model to generate sentences that conform to grammatical norms. It not only makes the generated description more natural and smooth, but also improves the readability and comprehensibility of the description.

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

2 Related work

2.1 Image Caption

In recent years, image captioning in natural language has been an active field of artificial intelligence research. Various image description methods are proposed to improve the latest technology of image description. In this section, we briefly review the research progress of image captioning in the field of computer vision and remote sensing.

To capture short-term spatial semantic relations and long-term transformation dependencies, (Tu et al., 2022) proposed a long-short term relationship Transformer (LSRT) to fully mine the relationships between objects to generate the caption. To cope with the need to understand video content, caption semantics, and the relationship between them for effective caption modeling, (Yu et al., 2022) proposes an internal and relational embedding Transformer (I^2Transformer), which makes full use of various modalities and enhances them with cross-modal information during semantic interaction. (Ji et al., 2023) uses a dual attention mechanism when processing image captions, which is applied to pyramid feature maps. The method fully considers the context information provided by the hidden state, so as to locate the visually and semantically coherent regions in the image more effectively. At the same time, the context information helps to recalibrate the feature components and improve the discrimination ability of visual features. Although self-attention-based networks have achieved great success in image captioning, existing self-attention networks are still plagued by distance insensitivity and low-rank bottlenecks. (Tu

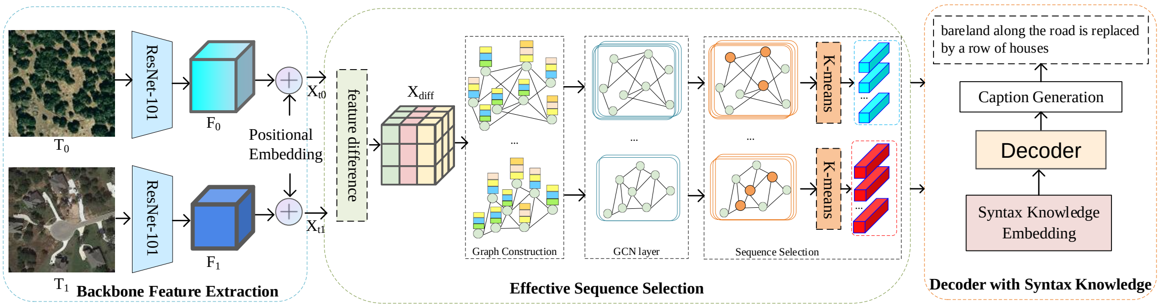

T0
T1
ResNet-101
ResNet-101
F0
F1
Positional Embedding
Xt0
Xt1
Backbone Feature Extraction
feature difference
Xdiff
Graph Construction
GCN layer
Sequence Selection
K-means
K-means
Effective Sequence Selection
bareland along the road is replaced by a row of houses
Caption Generation
Decoder
Syntax Knowledge Embedding
Decoder with Syntax Knowledge

Figure 1: Architecture of the proposed SGD-RSCCN.

et al., 2023b) introduced distance-sensitive self-attention (DSA), considered the original geometric distance between query key pairs in 2D images, and proposed multi-branch self-attention (MSA) to make up for the low-rank bottleneck. Since the traditional attention mechanism only uses a one-way flow from vision to language, the visual features of interest are usually independent of the target word state. Therefore, (Ariyo et al., 2019a) proposes to improve the traditional attention mechanism to a relationship-aware attention mechanism with two graph learning, namely, visual-to-visual homogeneous graph (HMG) and language-to-visual heterogeneous graph (HTG), which capture the internal relationship of visual features and the relationship between the target word and the visual features concerned, respectively.

2.2 Change Caption

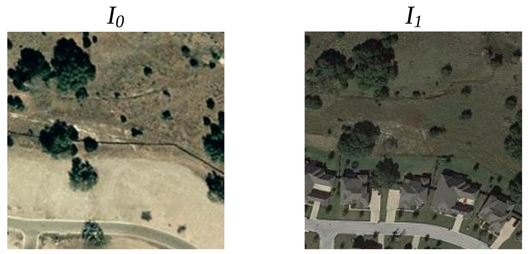
Compared with the traditional image captioning task, the change captioning task is more challenging because it needs to deeply understand the content of the two images and further describe the differences.

Pioneer work (Jhamtani and Berg-Kirkpatrick, 2018) describes changes based on monitoring scenarios. (Tan et al., 2019) describes in detail the editing conversion between the two images. (Ariyo et al., 2019c) proposes a fully convolutional CaptionNet (FCC), which uses an encoder-decoder architecture to perform visual feature extraction, calculate feature distance, and generate new sentences describing the measured distance. Combining a variety of deep learning techniques, (Ariyo et al., 2019b) proposed a multimodal end-to-end connection difference captioning model (SDCM) for capturing, aligning, and calculating the differences between the two image features. In order to generate accurate captions, (Chang and Ghamisi, 2023) proposed an attention caption network for dual-temporal remote sensing images. In order to improve the model's ability to perceive various changes in different scenarios, (Tu et al., 2023a) proposed a neighborhood comparison converter. Adjacent feature aggregation is designed to integrate adjacent contexts into each feature. A common feature distillation is designed to compare two images at the neighborhood level, and common attributes are extracted from each image to learn effective comparison information. (Yue et al., 2023) proposes an Intra- and Inter-representation Interaction Network (I3N) for learning fine-grained difference representations that are not affected by viewpoint changes. In order to make the change description model capture the actual change while ignoring the influence of perspective change, (Kim et al., 2021) proposed a view-independent change caption network (VACC) with circular consistency. In order to learn stronger visual and linguistic associations to obtain fine-grained visual differences, (Yao et al., 2022) proposed a modeling framework that follows a pre-trained fine-tuning paradigm. Aiming at the shortcomings of current remote sensing image change description methods in fully extracting and utilizing multi-scale information, (Liu et al., 2023a) proposed a progressive scale-aware network (PSNet) to solve the problem. (Huang et al., 2022) proposed an instance-level fine-grained differential captioning (IFDC) model.

3 Model

The proposed method follows the encoder-decoder architecture and is used for remote sensing image change description generation. In this section, we first explain the overview of the model, then describe the architecture of the visual feature extractor in detail in Section 3.1, and describe the description generation in Section 3.2.

As shown in Figure 1, our SGD-RSCCN consists of three main modules: backbone feature extrac-



Caption:
"bareland", "along", "the", "road", "is", "replaced",
"by", "a", "row","of", "houses"
Syntax Dependency:
["nsubjpass","prep","det","pobj","auxpass","ROOT",
"agent","det","pobj","prep","pobj","punct"]

Figure 2: The syntax-dependent knowledge diagram.

tion module, effective sequence selection module (ESS) and decoder with syntax knowledge (DSK). Given the input of the dual-temporal image, we first use the shared backbone network to extract the feature map of the given image pair. Next, we input the features into the effective sequence selection module, that is, each pixel of the feature map is regarded as a graph node, and the graph neural network is proposed to model the structured information and learn the features of the change description directly from the original remote sensing data. We can mine the top-K effective sequence from the graph and use the clustering algorithm to refine it. Finally, a description generator based on grammatical prior knowledge is used to obtain a more accurate description of changes.

3.1 Effective Sequence Selection Module (ESS)

In remote sensing images, the spatial relationship between pixels is crucial for understanding the scene. GNN (Graph Neural Network) can model the spatial relationship by learning the connection pattern between nodes, thus providing richer spatial context information. At the same time, through the information transmission mechanism between nodes, the feature embedding of each node (or pixel) can be learned. The embeddings can capture the complex attributes of objects in remote sensing images, such as shape, size, and texture, and they can automatically identify and learn the features that are most useful for change description tasks. Specifically, the nodes in the graph represent the pixels in the image features, and the edges represent the relationship between them. GNN can effectively capture the complex relationship between these nodes and extract structured features.

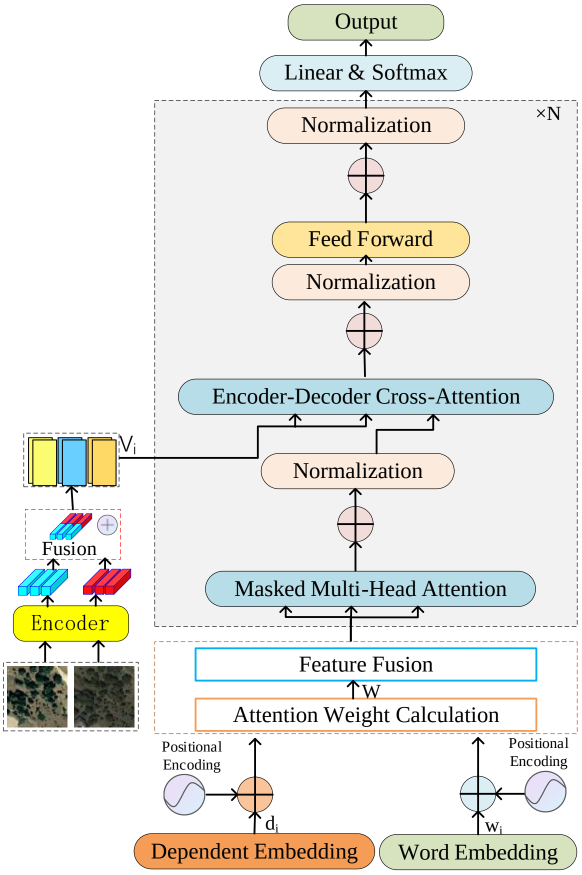


Figure 3: Outline of the decoder from visual features to description sentences.

To better capture the detailed information of the changing area, the two feature maps are subtracted and the absolute value is taken to obtain the difference feature map $X_{diff} = |X_1 - X_2| \ \epsilon \ (R)^{HW*C}$. Firstly, we construct an undirected weight graph $G = \{V, E\}$ by treating each feature point as a graph node v_i and the spatial relationship between nodes i and j as an edge $e_{i,j} = (v_i, v_j) \ \epsilon \ E$. Then, our effective sequence selection task can be regarded as the node selection in the graph G. We use GCN (Graph Convolutional Network) to learn graph-based structured information and obtain the reliable confidence of graph nodes for effective node selection. Specifically, we first calculate the adjacency matrix $A \in R^{HW \times HW}$ (Equation 1), which is used to measure the interaction between node pairs in the graph.

$$if\ A_{i,j} = \begin{cases} 0 \\ \overline{x_i}.\overline{x_j},\ if\ v_i, v_j\ are\ adjacent \end{cases} \tag{1}$$

Among them, $\overline{x_i}, \overline{x_j} \in \overline{X}$ respectively represent the characteristics of the node v_i and v_j. Structured information can be modeled and propagated in the graph through the GCN module. For the calculation defined in each layer of GCN, it can be expressed as Equation 2:

$$H^{(l+1)} = \sigma \left(\left(\widetilde{D} \right)^{-\frac{1}{2}} (I + A) \left(\widetilde{D} \right)^{-\frac{1}{2}} \right) H^{(l)} W^{(l)} \tag{2}$$

Among them, I is the unit matrix, $\widetilde{D}$ is the diagonal matrix, $\left(\widetilde{D} \right)_{ij} = (\sum)_j \left(\widetilde{A} \right)_{ij}$ and $\bar{A} = I{+}A$. $W^{(l)}$ represents learnable parameters. After the GCN layer processing, each element in the final output $P = H^{(L)} \in R^{H \times W \times 1}$ corresponds to a rough change confidence. Among them, the larger the value of the feature point, the greater the probability that the region is a changing region. In order to select feature sequences with high confidence, we record their position coordinates in P and obtain the first K minimum values. The valid tokens come from the features representing the unchanged region on the original feature map, and the two feature maps construct two sets of tokens from the same region. In order to further reduce the number of tokens, we finally use the K-means algorithm on X_1 and X_2 to obtain tokens centered on the class center L (L < < K) for the two branches, that is, $T_1,\ T_2 \in R^{L \times C}$, L represents the length of each group of tokens, C representing the channel dimension. After many experiments and parameter adjustments, in the experiment, we finally used the initial class center K value of 20, and further selected the class center with L value of 2 for the effective sequence.

3.2 Decoder with Syntax Knowledge(DSK)

In the generation of change descriptions, most of the work focuses on learning accurate change representations to generate descriptions, while ignoring the use of syntactic knowledge. In order to help the model distinguish the changed objects and their references in the real description during training, we propose a decoder module based on prior knowledge to eliminate the grammatical structure ambiguity in the change description.

First, we briefly introduce the dependencies between words. In natural language processing, dependency analysis refers to the process of checking the dependency between the linguistic units (such as words) of a sentence to determine its grammatical structure. That is to say, grammatical dependency refers to the concept that words are connected to each other through directed links. The verb is regarded as the structural center of the clause structure and is marked as the root “root”. All other syntactic words are directly or indirectly connected to the root “root” through directed links.

As shown in Figure 2, the corresponding change in the image is described as : “bareland along the road is replaced by a row of houses”. We observe that change description usually consists of two parts: semantic change and reference, which makes it contain complex syntactic structures. However, the subject “bareland” and its predicate “is replaced” are separated by the attribute describing the object “road”. In the case, the word “road” is closer to “is replaced” than “bareland”. In the training process, if the model does not understand the grammatical relationship between words, it may learn wrong information from the real caption. According to the literature research, the existing methods ignore the problem. In fact, if the model notices the direct dependence between “bareland” and “is replaced”, the above misunderstanding can be avoided. Therefore, it is necessary to introduce the grammatical dependency knowledge of text modality in the training process to help the model understand the grammatical structure of the description sentence.

Specifically, we load the small English model of the Spacy library "en_core_web_sm", which contains the language rules and resources required for syntactic analysis. By reading the existing data files containing images and their corresponding sentences, for each sentence of each image, we extract the original text and use Spacy for syntactic analysis to extract the dependency of each word in the sentence. Then the dependency list of each sentence is added to the annotation data of the corresponding image. Finally, a JSON file containing image description dependencies is output, which will be used to train the image description generation model to help the model understand the sentence structure and generate a more accurate description. By reading the obtained dependency data files, all unique dependency labels are extracted and indexes are assigned to them, and finally a corresponding dependency vocabulary file is generated.

We use the decoder with syntax knowledge shown in Figure 3 to generate the change description. Specifically, each decoder consists of N stacked Transformer decoding blocks. Each block

Method	B-1	B-2	B-3	B-4	M	R	C
Base	69.11	57.99	50.36	44.08	27.06	57.26	91.51
Base+ESS	74.27	63.25	53.88	45.69	32.96	64.60	110.55
Base+ESS+DSK	**77.15**	**66.80**	**58.07**	**50.27**	**34.07**	**65.88**	**118.42**

Table 1: Change description results on the Dubai-CC dataset. 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. All values are reported as percentage (%).

Method	B-1	B-2	B-3	B-4	M	R	C
Base	77.99	70.67	65.14	61.17	37.47	70.50	125.83
Base+ESS	82.38	72.87	65.25	59.22	38.57	73.06	131.33
Base+ESS+DSK	**84.17**	**75.16**	**68.05**	**62.48**	**39.18**	**74.24**	**136.27**

Table 2: Change description results on the LEVIR-CC dataset. 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. All values are reported as percentage (%).

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. At the same time, the corresponding syntax dependency is also expressed as the one-hot vector d_i. The description decoder takes the attention fusion features of w_i and d_i as input, and the masked multi-head attention mechanism embeds the word and grammar dependencies. And calculate the fusion embedding feature $\hat{E}\left[W\right]$. Then, through Encoder-Decoder cross-attention, $\hat{E}\left[W\right]$ 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 Equation 3.

$$p_i = softmax\left(W^T h_i^N + b_i\right) \tag{3}$$

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.

4 Experimental results

4.1 Dataset

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

4.2 Evaluation indicators

Following the most advanced change description methods, we use four common indicators to evaluate the accuracy of all methods, namely BLEU-N (where N = 1,2,3,4), ROUGE-L, METEOR and CIDEr-D. 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.

4.3 Implementation details

The method based on the PyTorch framework is trained and evaluated on the NVIDIA A100 or V100. We use ResNet-101 pre-trained to extract image features. During training, we use the Adam optimizer with the initial learning rate of 0.0005. At the same time, 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 evaluate the performance of the model on the whole test data set.

4.4 Ablation studies

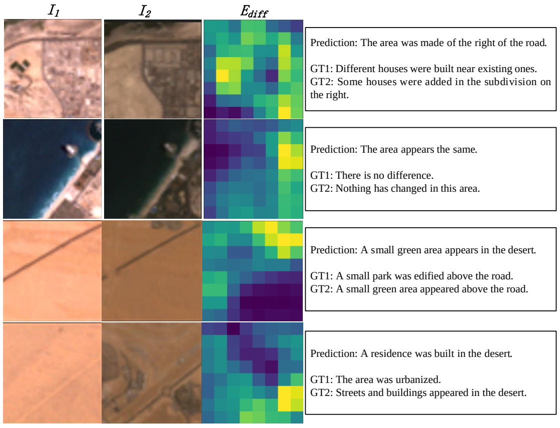
To clarify the contribution of each module of the proposed network, we conducted ablation experiments. The baseline does not contain any mod-

Method	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**
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
PSNet (2023b)	-	-	-	-	-	-	-
Prompt-CC (2023c)	-	-	-	-	-	-	-
RSICCformer (2022b)	67.92	53.61	41.37	31.28	25.41	51.96	66.54
SGD-RSCCN	**77.15**	**66.80**	**58.07**	**50.27**	**34.07**	**65.88**	**118.42**
SOTA	↑ 13.59	↑ 24.60	↑ 40.37	↑ 60.71	↑ 34.08	↑ 26.79	↑ 77.97
Average	↑ 16.06	↑ 20.63	↑ 22.58	↑ 23.09	↑ 11.27	↑ 17.17	↑ 56.01

Table 3: Comparisons experiments on the Dubai-CC dataset. 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. All values are reported as percentage (%).

Method	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**
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 (2023b)	83.86	75.13	67.89	62.11	38.80	73.60	132.62
Prompt-CC (2023c)	83.66	75.73	**69.10**	**63.54**	38.82	73.72	**136.44**
RSICCformer (2022b)	**84.72**	**76.27**	68.87	62.77	**39.61**	74.12	134.12
SGD-RSCCN	84.17	75.16	68.05	62.48	39.18	**74.24**	136.27
SOTA	–	–	–	–	–	↑ 0.16	–
Average	↑ 1.84	↑ 1.75	↑ 2.11	↑ 2.60	↑ 1.21	↑ 2.20	↑7.55

Table 4: Comparisons experiments on the LEVIR-CC dataset. 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. All values are reported as percentage (%).



Prediction: The area was made of the right of the road.

GT1: Different houses were built near existing ones.
GT2: Some houses were added in the subdivision on the right.

Prediction: The area appears the same.

GT1: There is no difference.
GT2: Nothing has changed in this area.

Prediction: A small green area appears in the desert.

GT1: A small park was edified above the road.
GT2: A small green area appeared above the road.

Prediction: A residence was built in the desert.

GT1: The area was urbanized.
GT2: Streets and buildings appeared in the desert.

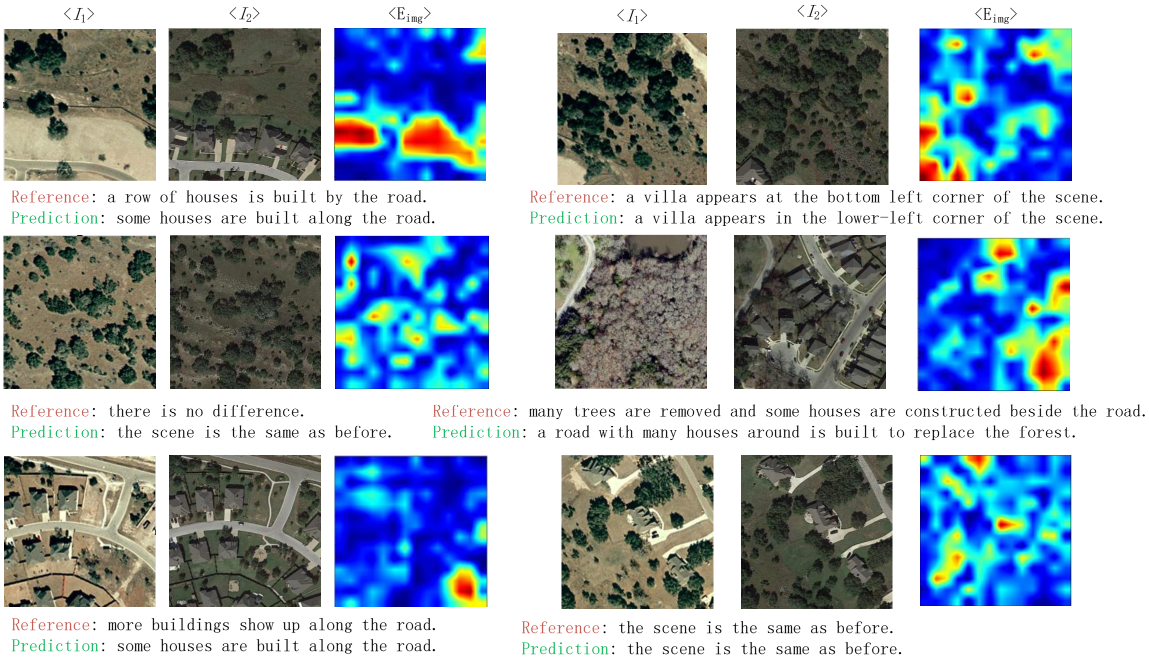
Figure 4: Visualized image embeddings and change captioning examples generated by SGD-RSCCN in the Dubai-CC dataset. GT1 and GT2 represent reference description 1 and reference description 2 respectively.

ules. The experimental results are shown in Table 1 and Table 2, where Table 1 focuses on the Dubai-CC dataset and Table 2 addresses the LEVIR-CC dataset.

On the Dubai-CC dataset, it can be seen that after the introduction of the ESS module, all indicators have been greatly improved. BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L and CIDEr-D increased by 5.16%, 5.26%, 3.52%, 1.61%, 5.9%, 7.34% and 18.99%, respectively. After further introducing the DSK module, BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L and CIDEr-D increased by 2.88%, 3.55%, 4.19%, 4.58%, 1.11%, 1.28% and 7.87%, respectively. On the LEVIR-CC dataset, it can be seen that the introduction of the ESS module increases the BLEU-1, BLEU-2, BLEU-3, METEOR, ROUGE-L, and CIDEr-D scores by 4.39%, 2.2%, 0.11%, 1.1%, 2.56%, and 5.5%, respectively. Further introduction of the DSK module increases each score by 1.79%, 2.29%, 2.8%, 3.26%, 0.61%, 1.18%, and 3.94%, respectively.

4.5 Comparison to State-of-the-Art

Table 3 and Table 4 show the performance evaluation of the proposed SGD-RSCCN model with three natural image change captioning methods and three remote sensing change captioning methods using the Dubai-CC and LEVIR-CC datasets. The results show that the SGD-RSCCN model is su-



Reference: a row of houses is built by the road.
Prediction: some houses are built along the road.

Reference: a villa appears at the bottom left corner of the scene.
Prediction: a villa appears in the lower-left corner of the scene.

Reference: there is no difference.
Prediction: the scene is the same as before.

Reference: many trees are removed and some houses are constructed beside the road.
Prediction: a road with many houses around is built to replace the forest.

Reference: more buildings show up along the road.
Prediction: some houses are built along the road.

Reference: the scene is the same as before.
Prediction: the scene is the same as before.

Figure 5: Visualized image embeddings and change captioning examples generated by SGD-RSCCN in the LEVIR-CC dataset.

perior to the comparison method in terms of key evaluation indicators on the two datasets. On the Dubai-CC dataset with a spatial resolution of 30 meters, our SGD-RSCCN model is 8.66% higher than the best performing RSICCformer method on METEOR and 51.88% higher on CIDEr-D. In the LEVIR-CC dataset with higher spatial resolution, SGD-RSCCN is 1.15% higher than the best performing RSICCformer method on CIDEr-D.

In addition, on the Dubai-CC dataset, our method improves the value of each indicator by an average of 11% - 56% among all comparison methods. The value of CIDEr-D has the largest average improvement among all comparison methods, up to 56 %. On the LEVIR-CC dataset, our method improves the value of each indicator by more than 1% on average in all comparison methods. The value of CIDEr-D is increased by an average of 7.55% in all comparison methods. It shows that the change description generated by SGD-RSCCN is significantly better than other methods in quality, showing higher language fluency and fit with image content.

4.6 Qualitative Results

To evaluate the quality of our model, we visualize the change description and prediction generated by the description decoder, as shown in Figure 4 and Figure 5, where I_1 and I_2 represent the images captured in Time 1 and Time 2, respectively. E_{img} is image embedding, and E_{diff} is the difference image embedding extracted by the encoder.

In Figure 4, it can be observed that in the first image pair, we correctly identify the location of the change region on the right side of the road, while in the third image pair, our SGD-RSCCN accurately identifies the number of regional changes. In addition, as shown in the second image pair, our SGD-RSCCN shows a significant ability to accurately distinguish scene changes.

In Figure 5, the description generated by SGD-RSCCN is visualized with the actual reference description, including the description of the changed image pair and the actual unchanged image pair. It can be observed that SGD-RSCCN can effectively distinguish the actual change from the irrelevant change, such as the third image pair (horizontal view) and the last image pair. In the second image pair, we accurately identify the changed object and its specific orientation, while in the fourth image pair (horizontal view), SGD-RSCCN accurately gives the logical relationship of the change.

It shows that our SGD-RSCCN is excellent in identifying the location of changing objects, their attributes and the relationship between objects.

5 Conclusion

The SGD-RSCCN proposed significantly improves the accuracy of remote sensing image change information extraction and the naturalness of description by integrating scene graph construction and dependency syntax analysis. By introducing attention mechanism and scene graph construction technology, the model can better understand and represent complex remote sensing image scenes while capturing dynamic change processes. In addition, the decoder based on prior knowledge uses dependency parsing to enhance the model's compliance with grammatical rules, making the generated description more in line with the norms of natural language, and improving the readability and comprehensibility of the description. Extensive experiments on LEVIR-CC and Dubai-CC datasets verify the effectiveness of the method, and show superior performance.

Limitations

Although the study has made significant progress in the description of remote sensing image changes, there are still challenges in dealing with complex and irregular sentence patterns. Future work will explore the combination of more advanced dependency models and context-aware mechanisms. In addition, extending the model to multilingual and cross-domain applications is also an important direction for future research. Through continuous optimization, the remote sensing image change description network is expected to play a greater role in a wider range of application scenarios.

Acknowledgements

This work was supported in part by Natural Science Foundation of China under Grant 62162047, Natural Science Foundation of Inner Mongolia under Grant 2023MS06017, Inner Mongolia Youth Science and Technology Talents support project under Grant NJYT24034, Inner Mongolia Autonomous Region higher Education Carbon Peak and Carbon Neutral Research Project under Grant STZX202322, the Self-project Program of Engineering Research Center of Ecological Big Data, Ministry of Education, and the fund of Supporting the Reform and Development of Local Universities (Disciplinary Construction) and the special research project of First-class Discipline of Inner Mongolia A. R. of China under Grant YLXKZX-ND-036.

References

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, Muhammad Umar Aftab, Eatedal Alabdulkreem, Bulbula Kumeda, Edward Yellakuor Baagyere, and Zhiquang Qin. 2019b. 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. 2019c. Fully convolutional captionnet: Siamese difference captioning attention model. *IEEE Access*, 7:175929–175939.

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.

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.

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.

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, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2023b. 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. 2022a. 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, Hao Chen, Zhengxia Zou, and Zhenwei Shi. 2022b. 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. 2023c. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–18.

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, 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^{2}transformer: 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, 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.

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.

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.