# Semantic-Spatial Collaborative Perception Network for Remote Sensing Image Captioning

Qi Wang, *Senior Member, IEEE,* Zhigang Yang, Weiping Ni, Junzhen Wu, Qiang Li, *Member, IEEE*

***Abstract*—Image captioning is a fundamental vision-language task with wide-ranging applications in daily life. Existing methods often struggle to accurately interpret the semantic information in remote sensing images due to the complexity of backgrounds. Target region masks can effectively reflect the shape characteristics of targets and their potential interrelationships. Therefore, incorporating and fully integrating these features can significantly improve the quality of generated captions. However, researchers are hindered by the lack of relevant datasets that contain corresponding object masks. It is natural to ask: How to efficiently introduce and utilize object masks? In this paper, we provide potential target masks for the publicly available remote sensing image caption (RSIC) datasets, enabling models to utilize the regional features of targets for RSIC. Meanwhile, a novel RSIC algorithm is proposed that combines regional positional features with fine-grained semantic information, abbreviated as S$^2$CPNet. To effectively capture the semantic information from image and position relationship from mask respectively, the semantic and spatial feature enhance sub-modules are introduced at the ends of encoder branches, respectively. Furthermore, the cross view feature fusion module is designed to integrate regional features and semantic information efficiently. Then, a target recognition decoder is developed to enhance the ability of model to identify and describe critical targets in images. Finally, we improve the caption generation decoder by adaptively merging textual information with visual features to generate more accurate descriptions. Our model achieve satisfactory results on three RSIC datasets compared with existing method. The related datasets and code will be open-sourced in https://github.com/CVer-Yang/SSCPNet.**

***Index Terms*—Remote sensing, image captioning, cross view, attention mechanism**

## I. Introduction

THE goal of remote sensing image captioning task is to translate the content of given remote sensing images into text [1] [2], enabling humans to intuitively grasp the crucial information contained in the images. This task combines computer vision and natural language processing, which holds significant application value in various fields [3] [4], such as intelligent security, military intelligence collection, geographic information updating, etc.

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2041, Grant 62471394, and Grant 62301385. Qi Wang, Zhigang Yang and Qiang Li are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: crabwq@gmail.com, zgyang@mail.nwpu.edu.cn, liqmges@gmail.com) (Corresponding author: Qiang Li.)

Weiping Ni and Junzheng Wu are with the department of remote sensing, Northwest Institute of Nuclear Technology, Xi'an 710072, P.R. China. (e-mail: niweiping@nint.ac.cn, wujunzheng@nint.ac.cn)

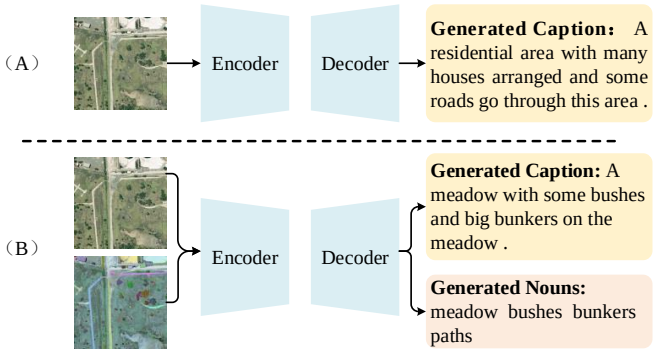



Fig. 1. The motivation of the proposed S$^2$CPNet. **Top:** Existing RSIC algorithms primarily rely on an encoder-decoder framework, where the encoder extracts visual features from the image and then feeds them into a language decoder to generate corresponding textual descriptions. **Bottom:** Unlike these methods, we propose a novel algorithm that takes the image and its corresponding target mask as inputs. The extracted comprehensive visual features are fed into a dual-branch decoder, which separately generates the categories of the ground targets and the textual descriptions.

Currently, RSIC faces the following challenges: **(1) Limited visual feature expression ability.** Remote sensing images are typically captured from an overhead perspective, encompassing numerous objects within a confined area. The complexity of backgrounds and the small account of targets make it challenging for models to accurately extract target features. Additionally, the existence of objects with various scales weakens the ability of model to analyze the relationships between different geospatial targets. This further hinders the model from generating precise captions. Therefore, models need to accurately extract semantic features of targets and understand the spatial relationships between geospatial entities. **(2) Semantic confusion in description generation.** Both the accuracy of visual content and the fluency of the generated description are key factors in evaluating the quality of the generated captions. Therefore, it is essential to efficiently utilize both visual and textual features in the text generation process. However, structural and semantic discrepancies between cross-modal data weaken the ability of model to fully leverage these features. Consequently, effectively integrating visual and text features to produce satisfactory language descriptions remains a critical challenge.

Inspired by that Qu et al. [5] introduced the RSIC task, many studies are then proposed. Considering the challenge that models cannot accurately extract target features from complex remote sensing images, researchers employ some strategies to alleviate this problem, such as multi-scale feature integration [6] [7] and the introduction of attention mechanisms [8] [9].

As shown in the top of Fig. 1, these methods primarily focus on extracting fine-grained features from images and feeding them into decoders to produce captions. Different from the above methods, we improve on the following three aspects to generate high-quality captions.

Firstly, since fine-grained features fail to supply the overall characteristics of targets, which poses difficulties in meeting the demands of model for accurate shape features and spatial relationships. In contrast to fine-grained features, regional features can offer an alternative perspective by focusing on the overall characteristics of significant targets. Consequently, in natural image captioning, numerous researchers combine fine-grained features with regional features to produce descriptions. However, the application of regional features in RSIC faces significant challenges due to the lack of corresponding annotated datasets. Creating these datasets is labor-intensive and costly. Although object detection annotations can provide the location of targets, they also may introduce extraneous noise, such as background elements, which hinder the generation of precise semantic descriptions. Compared to object detection annotations, object masks can offer a more accurate representation of target locations and shapes. Therefore, the efficient acquisition of target segmentation annotations is crucial for generating visual features with stronger expression ability.

Secondly, designing auxiliary task [10] is an effective approach for models in learning more robust feature representations. In the fields of RSIC, existing multi-task models mainly enhance feature extraction capabilities by incorporating scene classification task. Indeed, scene classification task can facilitate the learning of global image features. However, this task cannot guide the model to capture local details of images. On the other hand, RSIC tasks require models to capture the relationship between the geospatial targets and their corresponding words, and the scene classification task cannot provide this information. Therefore, we introduce the multi-target classification task, as shown in the bottom of Fig. 1. By introducing this task, the model can achieve an efficient alignment of targets with textual features.

Thirdly, during the caption generation stage, most methods utilize Transformer for text modeling and employ cross Transformer to facilitate interactions between visual and textual features. However, this structure also introduces confusion in the description generation process. Specifically, in generated descriptions, some crucial entity nouns are directly related to the image content, such as “airplane”, and “building”. In contrast, conjunctions like “and” can be inferred simply without introducing visual features. In this case, using cross-attention mechanism to model meaningless words with visual features is unhelpful and may even impair the ability to align cross-modal features. Therefore, it is essential to design an adaptive fusion mechanism that efficiently utilizes features from different models and filters out irrelevant information, which can contribute to generating precise and coherent descriptions.

To alleviate these aforementioned challenges, we extend the existing RSIC dataset and design a novel RSIC algorithm. In terms of dataset, we combine the Segment Anything Model (SAM) [11] to automatically generate target masks corresponding to potential targets, which can provide the overall contour and location information of the crucial targets. We also extract the entity nouns in the caption to form a key noun sequence. In terms of algorithm, we propose an RSIC method that combines the target mask and fine-grained semantic information. Firstly, the remote sensing image and the corresponding coarse-grained annotation are taken as the input of the model, and the key feature representation of remote sensing image targets is enhanced by designing a semantic and spatial enhancement sub-module. Then a cross-view feature fusion module is designed to effectively fuse the fine-grained semantic information of the image with the corresponding regional features. In addition, we introduce a multi-target classification decoder to enhance the understanding ability of remote sensing images. Finally, we propose a gated-guide Transformer decoder, it can generate linguistically fluent and correct descriptions based on textual and visual features adaptively. The main contributions of this work are summarized as follows:

- We provide pixel-level annotations of remote sensing images for the potential targets based on publicly available RSIC datasets. These datasets form the foundation for the model to generate accurate descriptions by combining region features with fine-grained semantic features. Additionally, we extract the nouns in the caption to form a sequence of categories.
- The spatial and semantic enhancement sub-modules are designed at end of encoder, which can realize the feature enhancement of potential target semantic information and location features. Meanwhile, a Cross-view feature fusion (CVFF) module is designed to realize the full interaction between semantic information and regional features. It provides rich and complete visual features for the decoder.
- The multi-target classification auxiliary task is introduced to enhance the ability of model to understand potential targets, which can align the visual features with the crucial nouns. Meanwhile, we propose a gated guided caption decoder, which can adaptively utilize the visual features with the generated descriptions to help the model produce accurate and fluent text descriptions.

The rest of this paper is organized as follows. Section II provides related work in natural image captioning and RSIC. In Section III, we describe the details of processed dataset and our proposed $S^2$CPNet. The experimental results are analyzed in Section IV. Finally, Section V offers the conclusion.

## II. Related Work

In this section, we review the existing natural image captioning and the RSIC methods.

### A. Natural Image Captioning

Image captioning task aim to generate textual description that correspond to the content of images. With the advancement of deep learning technique [12] [13], many researchers combine neural network techniques to generate captions. Vinyals et al. [14] use convolution network to extract image

features and employ LSTM network to generate captions. On this bais, Xu et al. [15] integrate an attention mechanism during decoding, enabling the model to dynamically focus on relevant visual aspects. Yao et al. [16] further enhance visual feature extraction by using graph neural networks to model relationships between objects. Integrating textual information related to the image has proven beneficial for generating accurate descriptions. Inspired by this, Mun et al. [17] adopt text retrieval methods to guide the model in identifying key visual elements. Similarly, Lu et al. [18] introduce an adaptive attention network to effectively concentrate on target-related features during caption generation.

On the other hand, Anderson et al. [19] propose a bottom-up and top-down visual attention mechanism, which enables the model to locate attention targets and object regions accurately. Huang et al. [20] extend this idea by integrating attention mechanisms into Transformer networks to filter out irrelevant visual features. Fei [21] introduce a self-supervised method to improve the ability to focus on relevant image regions. Luo et al. [22] propose a diffusion model based approach for caption generation. The method leverages image retrieval to obtain semantic information, this information is then fed into the diffusion model to generate corresponding sentences.

Current large models show great advantages in Vision-Language tasks. Li et al. [23] develop a multimodal encoder-decoder structure that effectively aligns pretrained visual and textual feature encoders. They [24] introduce the Query Transformer (Q-Former) to optimize the alignment of visual and verbal features, achieving strong performance across various visual-verbal tasks. Recognizing that conventional image description algorithms often require extensive data and parameters for optimal learning, Ramos et al. [25] propose a retrieval-based approach that generates high-quality descriptions without extensive training, demonstrating robust domain adaptation capabilities. Although the above methods can obtain satisfactory results in natural scenes, they also exhibit certain limitations in RSIC tasks. Compared with natural images, remote sensing images contain more targets, requiring the model to have a stronger understanding of the feature targets. In addition, the model needs to accurately capture the positional relationships between the targets to generate comprehensive captions.

### *B. Remote Sensing Image Captioning*

Inspired by NIC methods, numerous RSIC methods have emerged [26]. Unlike natural images, remote sensing images contain lots of objects with varying orientations. Therefore, models not only need to effectively identify objects within images, but also accurately capture the relationships between them. Considering the difficulty in extracting visual features, Huang et al. [27] propose a multi-scale feature aggregation network and design denoising operations to eliminate redundant information from the aggregated features. Yuan et al. [28] present an innovative approach that combines a multi-level attention mechanism with graph neural networks to model image features. Similar methods have [29] [30]. Zhang et al. [31] extract both global and local features, and combine these features to generate accurate caption. To address the challenges that Transformer network do not well in capture local information, Meng et al. [32] design an attention gating mechanism to achieve precise aggregation of semantic and regional features, and a group attention mechanism is incorporated in the decoder to enhance the extraction of local features. Meanwhile, Du et al. [33] introduce a novel deformable attention mechanism to explore the relationship between foreground objects and background features.

Designing appropriate auxiliary tasks can assist models in extracting robust visual features. Zhao et al. [34] offer a network that can generate pixel-level segmentation results for key targets and image descriptions simultaneously. To facilitate multi-level interpretation of remote sensing images, Zhao et al. [35] provide a dataset that includes semantic segmentation, instance segmentation, and textual descriptions. They also design a joint optimization model for multi-task cooperative training. Ye et al. [36] introduce a multi-label classification task to incorporate prior knowledge, and then feed it into the decoder to enhance the accuracy of generated captions Kandala et al. [37] introduce the image scene classification task to guide the model in generating robust visual features. In view of the lack of explainability of RSIC methods, Wang et al. [38] divide the RSIC task into word prediction and word order tasks. It perform a multi-classification task in the image to predict potential words, and further generates captions by word sorting.

Combining regional features can yield high-quality visual representations. Meng et al. [39] utilize ResNet and Fast-RCNN networks to extract corresponding scene-level and object-level features from images. They then employ a graph neural network to model the correlations between different objects, enabling the generation of coherent descriptions. Furthermore, Zhao et al. [40] manually annotate target regions in the RSIC dataset and improve the decoder to enhance the efficiency of model in utilizing fine-grained and regional features. Large language model (LLM) technology offers significant advantages in the field of natural language processing. Hu et al. [41] propose a large remote sensing model for RSIC and remote sensing visual question answering tasks, they also provide a fine-grained RSIC dataset. Yang et al. [42] design a two-stage vision-language training method aimed at aligning visual features with textual information. They align the features of remote sensing images and text, and subsequently input the visual features into a LLM to generate corresponding textual descriptions.

Different from these approaches, we take remote sensing images with corresponding segmentation masks as inputs and thus can obtain richer visual features. Meanwhile, an multi-target classification task and improved decoder are introduced to fully utilize the visual and image text. The proposed method can obtain high-quality captions.

## III. Proposed Dataset and Method

In this section, we describe the processing of the dataset, and then introduce the overall structure of the designed semantic-spatial collaborative perception RSIC model with details of each module.

### A. Dataset Presentation

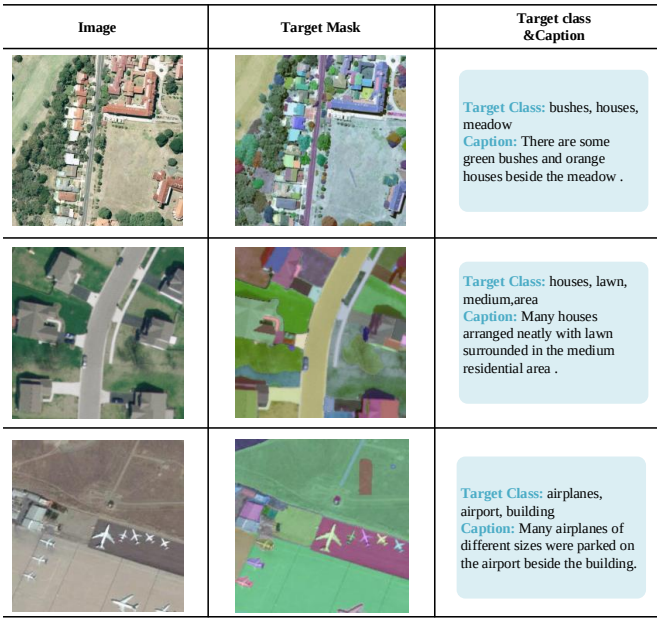



Fig. 2. The expanded Sydney caption, UCM caption, and NWPU caption datasets. These datasets comprise the original remote sensing images, potential target masks, captions for each image, and corresponding sequences of target nouns.

As shown in Fig.2, we further expand the caption dataset to provide segmentation mask of potential targets, and generate target class sequence by extracting meaningful nouns from the caption. Then, we describe the details of the dataset process.

**Potential Target Mask Generation:** To capture the positional relationship and overall shape features of potential targets in images, we provide a corresponding segmentation mask for each remote sensing image in an automated form. Specifically, the remote sensing image is input into the ViT_h model of the SAM, which produces the corresponding mask of potential objects. The generated segmentation mask is then combined with the input image. By combining images with masks, the generated image effectively represents the shape information of the target and the relative relationships between objects.

**Multi-target Class Sequence Generation:** Considering that the target masks lack relevant semantic information, we provide each remote sensing image with a corresponding target category sequence to help the model understand the semantic information of the image. Specifically, we use NLTK [43] techniques to extract nouns from the captions. However, due to semantic ambiguity, some of the extracted nouns may be meaningless. To address this problem, we implement a manual filtering process to eliminate nouns with a word frequency lower than five, as well as any meaningless nouns, such as "one" and "size". By introducing a multi-target classification task, the model can effectively align visual features with key word features, which guarantees the accuracy of the generated caption.

### B. Overview of $S^2$CPNet

The overall structure of the proposed model is illustrated in Fig. 3. The model comprises six key components: dual-view encoder, semantic feature enhancement sub-module, spatial feature enhancement sub-module, cross-view feature fusion module, target categorization generation decoder, and gated guided caption generation decoder. The remote sensing images and their corresponding coarse segmentation mask are processed by a weight-sharing ResNet50 [44] network, which serves as the encoder for image feature extraction. The fine-grained semantic features and potential target spatial features are then input to the semantic and spatial feature enhancement module to strengthen the expression ability of semantic information and target location features. Subsequently, the enhanced visual features are input to the cross-view feature fusion module, which facilitates the full fusion of fine-grained semantic features and location information. Finally, the aggregated features are input into the target category generation decoder and the gated guide caption generation decoder. These decoders generate the corresponding classification categories and comprehensive descriptions of the input image.

### C. Semantic-Spatial Feature Enhance

To capture complete visual features, the remote sensing image with the corresponding segmentation mask is utilized as input and fed into ResNet50 network. In each branch, the model can generate five feature maps with different resolutions. Feature map $E_5$ extracted by the remote sensing image feature extraction branch contains accurate semantic information, while the mask feature extraction branch captures rich object contour features $M_5$. To strengthen the semantic and position information conveyed by these features, we introduce a semantic enhancement sub-module after the remote sensing image branch, which aims to reinforce the semantic expression. Simultaneously, a spatial enhancement sub-module is implemented after the mask branch to enhance the contour and location information of the potential target.

The structure of the semantic feature enhancement sub-module is illustrated in Fig. 3. Initially, a global AvgPooling operation is adopted to compress the position information of the feature $E_5$ and generate global visual information. Then, a sequence of operations comprising Convolution, ReLU, and another Convolution is applied to the generated features to eliminate redundant information. The Sigmoid function is then utilized to obtain the weights for each channel. The semantic enhancement of the feature map $F_{sem}$ is achieved by multiplying the input semantic features by these channel weights. The overall process is defined as

$$W_{sem} = \delta(Conv(Relu(Conv(Avgpool(E_5))))), \quad (1)$$

$$F_{sem} = W_{sem} \otimes E_5, \quad (2)$$

where $Avgpool(\cdot)$ represents Avgpooling operation, $Conv(\cdot)$ represents convolution operation, $\delta(\cdot)$ represents Sigmoid operation, and $\otimes$ represents element-wise multiplication.

The structure of spatial feature enhancement sub-module is illustrated in Fig. 3. Initially, the low-frequency feature

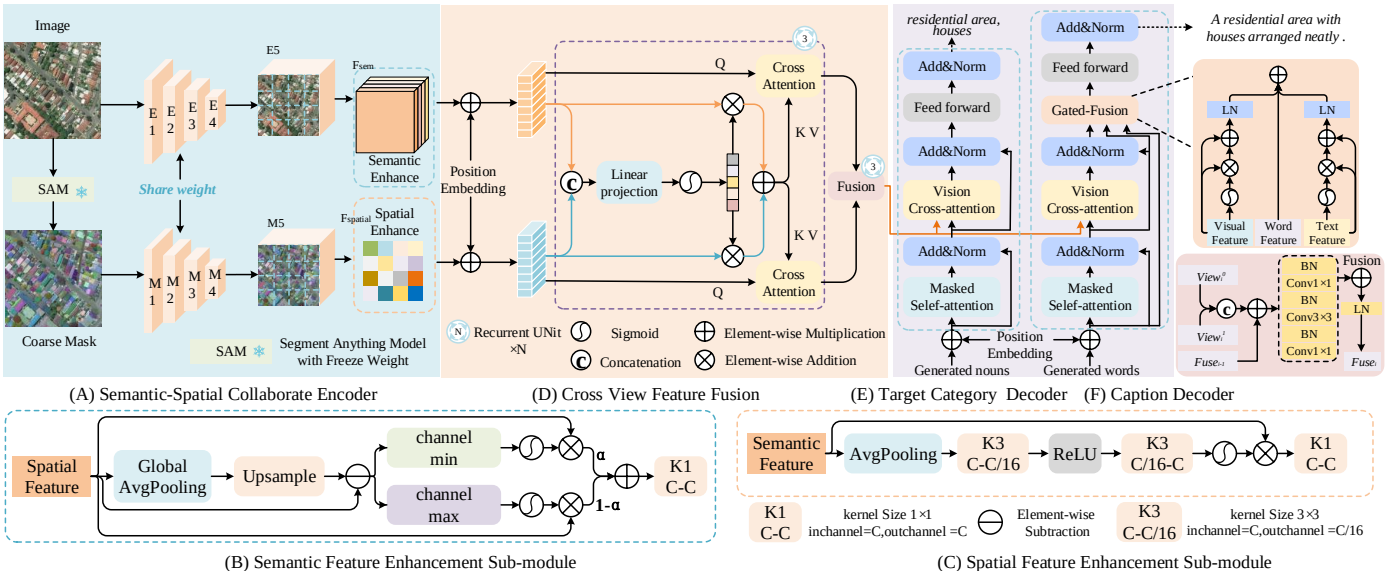



Fig. 3. The overall framework of the proposed S$^2$CPNet, and it is divided into some key parts: the encoder, semantic enhance sub-module, spatial enhance sub-module, cross view feature fusion, category generation decoder and gated guided caption generation decoder.

vectors of the feature map are obtained through an Avgpooling operation, followed by an upsampling operation to restore the resolution of $M_5$. Next, the high-frequency features of the feature map are acquired through a subtraction operation. Further, the channel information of the high-frequency feature map is compressed using both channel min and channel max operations. Sigmoid operations are applied on these feature maps to obtain the corresponding position weights. These positional weights are then multiplied with the input feature maps to achieve positional weight. To fuse these features, a trainable parameter $\alpha$ with an addition operation is introduced, and the fused features are finally passed through a convolution operation to further aggregate the positional information and obtain $F_{spatial}$. The overall process is defined as

$$W_{back} = \delta(min(M_5 - (Up(pool(M_5))))), \quad (3)$$

$$W_{fore} = \delta(max(M_5 - (Up(pool(M_5))))), \quad (4)$$

$$F_{spatial} = Conv(\alpha * W_{back} \otimes M_5 + (1-\alpha) * W_{fore} \otimes M_5), \quad (5)$$

where $pool(\cdot)$ represents global avgpooling, $min(\cdot)$ represents channel min operation, and $max(\cdot)$ represents channel max operation.

AvgPooling and reshape operations are performed on the refined feature map to generate features $V_1 \in \mathbb{R}^{196\times2048}$, $V_2 \in \mathbb{R}^{196\times512}$ respectively. Subsequently, the learnable 2-D position embedding algorithm is employed to incorporate positional information into these feature maps.

### D. Cross-view Feature Fusion Module

The feature maps generated by the encoder provide different information. To integrate these features effectively, we propose a cross-view feature fusion module that employs a coarse-to-fine feature fusion strategy. By incorporating this module, semantic information is effectively fused with positional features to generate a comprehensive visual feature representation.

The structure of the CVFF module is illustrated in Fig. 3, it consists of three recurrent units of the same structure. Initially, the feature maps $V_1$ and $V_2$ are concatenated along the channel dimension to obtain the coarse aggregated feature. A linear operation is then applied to compress the channel dimension, facilitating crucial interactions between the input features. Subsequently, a Sigmoid function combined with channel splitting is employed to obtain positional weights for each input feature map. These input feature maps are then multiplied by their respective positional weights to enhance the positional information. Finally, the weighted feature maps are summed to generate fused feature maps with intermediate granularity. To further enhance the integration of visual information across different feature maps, we employ a cross-attention mechanism to facilitate comprehensive interaction between input feature and the fused feature. Specifically, the visual features $V_1$ and $V_2$ serve as query, and the aggregated feature map acts as key and value. This approach enables the model to capture complementary visual features from the fused maps and minimize the influence of irrelevant features. Moreover, skip connection is implemented to accelerate the convergence of model. In this way, we obtain three sets of feature maps with precise semantic information and complete structural features respectively.

Finally, a concatenation with Resblock is employed to merge these features. The resblock is composed of Convolution with kernel size of 3× 3 and Batch Normalization operations. Specifically, the features from the different views $View_i^0$ and $View_i^1$ are initially aggregated using a concatenation operation and fused with the aggregated feature map $Fuse_{i-1}$ of the previous stage using an addition operation. Then, the fused fature is input to a Resblock block to generate the aggregated features $Fuse_i$ of the current stage. Through these operations, the model can effectively fuse semantic information with spatial details and obtain complete visual feature representations.

### E. Target Category Generation Decoder

Accurately recognizing the target category in an image is crucial for generating high-quality captions. The SAM model can produce the region and specific contours of the target effectively. However, the generated mask lacks corresponding semantic information, which poses challenges in accurately predicting the category of the target. To address this issue, we introduce a multi-target classification task. This task guides the model to correctly identify the target category in remote sensing images through a classification loss, thereby enhancing the comprehension of visual features. The structure of the module is illustrated in Fig. 3. First, we employ the Word Embedding algorithm [45] to encode the sequence of noun features. Subsequently, a self-attention mechanism is applied to encode noun sequences, which enables the model to understand the relationships between targets. Then, the noun features and fused visual features generated by the CVFF are fed into a cross-attention mechanism. Here, the noun sequence features serve as query, and the fused visual feature maps act as the key and value. This approach allows the model to correlate the image content with relevant noun sequences and predict the categories of targets present in the image.

### F. Gated Guide Caption Generation Decoder

Some words in the caption are inherently related to other words in the caption and require minimal visual content for accurate reasoning, such as conjunctions. In contrast, some words have a significant correlation with the visual content of an image. However, these words are fed into the cross attention mechanism to interact with the visual features, which cause part of the modeled information being useless or even harmful to the generation of high-quality captions. Therefore, we design a gated guide caption generation decoder. The structure of the caption decoder is shown in Fig. 3. Similar to the target category generation decoder, the input word $F^i_{Word}$ is embedded by using the Word Embedding algorithm and a self-attention mechanism. Subsequently, the fused feature maps and textual feature are fed into the cross attention mechanism. This decoder performs a Sigmoid operation on the text features $F^i_{Text}$ and visual features $F^i_{Visual}$ to determine the importance of cross-modal features for the current inference word. The feature maps are then multiplied accordingly. Finally, an addition operation followed by a linear operation is used to fuse these features. The overall process is defined as

$$F^i_{Caption} = LN(\delta(F^i_{Visual}) \otimes F^i_{Visual}) + LN(\delta(F^i_{Text}) \otimes F^i_{Text}) + F^i_{Word}, \quad (6)$$

where $LN$ is the LayerNorm operation, and $i$ represents the ith word in the caption. This module allows the model to adaptively utilize both image and text features to improve the ability to generate accurate and fluent captions.

### G. Loss Function

The RSIC can be reviewed as a temporal word classification task. We adopt the binary cross-entropy as the loss function. The definition of this loss function is shown below:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log(p_\theta(s^*_t | s^*_{1:t-1})), \quad (7)$$

where $s^*_1, s^*_2, \cdots, s^*_{t-1}$ represent the label words in caption. Similarly, the binary cross entropy is adopted as the loss function $L_{word}$ for the multi-target classification task. The total loss function of the proposed model is defined as

$$L_{loss} = L_{caption} + L_{word} \quad (8)$$

## IV. Experiments

In this section we first introduce the datasets and evaluation metrics used by the algorithm, followed by the specific details of the algorithm. At the same time, we analyze the performance of the algorithm on three datasets and provide some visualization examples. Finally, the results of the ablation experiments are analyzed.

### A. Datasets

To verify the effectiveness of this algorithm, we conduct experiments on Sydney, UCM and NWPU datasets. These datasets come from remote sensing image scene classification tasks. The Sydney dataset contains 613 remote sensing images across 7 scene categories. The UCM caption dataset includes 21 scene categories, with each category represented by 100 images at a resolution of 256×256 pixels. NWPU is the largest dataset for remote sensing image caption that covers 45 scene categories and includes a total of 31,500 high-resolution remote sensing images, each with a resolution of 256×256 pixels. All the above datasets are divided into training set, validation set and test set according to the original division.

### B. Evaluation Metrics

In this paper, we choose BLEU1-4, METEOR, ROUGE_L, CIDEr, Sm metrics to evaluate the quality of generated descriptions.

BLEU (Bilingual Evaluation Understudy): The BLEU metric is usually used in the field of machine translation. It calculates the score by comparing the overlapping n-grams between the generated text and the references.

METEOR (Metric for Evaluation of Translation with Explicit Ordering): It combines exact word matching with near-synonym matching and takes into account the information of word order for a more comprehensive evaluation of the similarity between the generated text and the reference text.

ROUGE_L (Recall-Oriented Understudy for Gisting Evaluation): ROUGE_L is commonly used for text summarization that focuses on the content overlap between the generated text and the reference text.

CIDEr (Consensus-based Image Description Evaluation): The metric is used to evaluate the image captioning task, which focuses on the occurrence of keywords.

Sm: This metric is obtained by calculating the average of BLEU-4, METEOR, ROUGE_L, and CIDEr, it can assess the overall quality of the generated caption.

TABLE I
EXPERIMENTS RESULTS ON SYDNEY DATASET. THE BOLD AND UNDERLINED INDICATE THE BEST AND SECOND BEST RESULTS. THE ∗ DENOTES THE RESULTS OF OUR RE-IMPLEMENTED.

| Method | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr↑ | Sm↑ |
|---|---|---|---|---|---|---|---|---|
| VLAD-LSTM [46] | 0.4913 | 0.3412 | 0.2760 | 0.2314 | 0.1930 | 0.4201 | 0.9164 | 0.4402 |
| Sound-a-a [47] | 0.7093 | 0.6228 | 0.5393 | 0.4602 | 0.3121 | 0.5974 | 1.7477 | 0.7794 |
| Multimodal [5] | 0.6980 | 0.6130 | 0.5440 | 0.5050 | 0.3610 | 0.6370 | 2.2020 | 0.9262 |
| Soft Attenttion [46] | 0.7128 | 0.6239 | 0.5527 | 0.4924 | 0.3675 | 0.6913 | 2.0343 | 1.0471 |
| Hard Attention [46] | 0.7689 | 0.6613 | 0.5840 | 0.5170 | 0.3719 | 0.6842 | 1.9863 | 0.9541 |
| TCE loss [48] | 0.7937 | 0.7304 | **0.6717** | **0.6193** | **0.4430** | 0.7130 | 2.4042 | 1.0449 |
| Word-Sentence framework [38] | 0.7891 | 0.7094 | 0.6317 | 0.5625 | 0.4181 | 0.6922 | 2.0411 | 0.9285 |
| GVFGA+LSGA [31] | 0.7681 | 0.6846 | 0.6145 | 0.5504 | 0.3866 | 0.7030 | 2.4522 | 1.0231 |
| SVM-D BOW [49] | 0.7787 | 0.6835 | 0.6023 | 0.5305 | 0.3797 | 0.6992 | 2.2722 | 0.9704 |
| SVM-D CONC [49] | 0.7547 | 0.6711 | 0.5970 | 0.5308 | 0.3643 | 0.6746 | 2.2222 | 0.9480 |
| MLCANet [50] | **0.8310** | **0.7420** | 0.6590 | 0.5800 | 0.3900 | 0.7110 | 2.3240 | 1.0012 |
| MLAT [51]∗ | 0.7768 | 0.7034 | 0.6422 | 0.5862 | 0.3882 | **0.7167** | 2.3605 | 1.0129 |
| Post-processing [52] | 0.7837 | 0.6985 | 0.6322 | 0.5717 | 0.3949 | 0.7106 | **2.5553** | **1.0581** |
| RS-CapRet [53] | 0.7870 | 0.7000 | 0.6280 | 0.5640 | 0.3880 | 0.7070 | 2.3920 | 1.0127 |
| $S^2$CPNet | 0.7770 | 0.7208 | 0.6597 | 0.5998 | 0.4023 | 0.7163 | 2.4787 | 1.0492 |

TABLE II
EXPERIMENTS RESULTS ON UCM DATASET. THE BOLD AND UNDERLINED INDICATE THE BEST AND SECOND BEST RESULTS. THE ∗ DENOTES THE RESULTS OF OUR RE-IMPLEMENTED.

| Method | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr↑ | Sm↑ |
|---|---|---|---|---|---|---|---|---|
| VLAD-LSTM [46] | 0.7016 | 0.6085 | 0.5496 | 0.5030 | 0.3464 | 0.6520 | 2.3131 | 0.9536 |
| Sound-a-a [47] | 0.7484 | 0.6837 | 0.6310 | 0.5896 | 0.3623 | 0.6579 | 2.7281 | 1.0845 |
| Multimodal [5] | 0.7100 | 0.5980 | 0.5530 | 0.4600 | 0.3430 | 0.6620 | 2.9260 | 1.0977 |
| Soft Attenttion [46] | 0.7454 | 0.6545 | 0.5855 | 0.5250 | 0.3886 | 0.7237 | 2.6124 | 1.0624 |
| Hard Attention [46] | 0.8157 | 0.7312 | 0.6702 | 0.6182 | 0.4263 | 0.7698 | 2.9947 | 1.2023 |
| TCE loss [48] | 0.8210 | 0.7622 | 0.7140 | 0.6700 | 0.4775 | 0.7567 | 2.8547 | 1.1897 |
| Word-Sentence framework [38] | 0.7931 | 0.7237 | 0.6671 | 0.6202 | 0.4395 | 0.7132 | 2.7871 | 1.1400 |
| GVFGA+LSGA [31] | 0.8319 | 0.7657 | 0.7103 | 0.6596 | 0.4436 | 0.7845 | 3.3270 | 1.3037 |
| SVM-D BOW [49] | 0.7635 | 0.6664 | 0.5869 | 0.5195 | 0.3654 | 0.6801 | 2.7142 | 1.0717 |
| SVM-D CONC [49] | 0.7653 | 0.6947 | 0.6417 | 0.5942 | 0.3702 | 0.6877 | 2.9228 | 1.1437 |
| MLCANet [50] | 0.8260 | 0.7700 | 0.7170 | 0.6680 | 0.4350 | 0.7720 | 3.2400 | 1.2787 |
| MLAT∗ [51] | 0.8226 | 0.7539 | 0.6959 | 0.6424 | 0.4268 | 0.7729 | 3.0826 | 1.2311 |
| Post-processing [52] | 0.7973 | 0.7298 | 0.6744 | 0.6262 | 0.4080 | 0.7406 | 3.0964 | 1.2186 |
| Clipcap ∗ [54] | 0.8213 | 0.7488 | 0.6872 | 0.6320 | 0.4232 | 0.7742 | 3.2391 | 1.2671 |
| PureT [55] | 0.8573 | 0.8020 | 0.7562 | 0.7129 | 0.4686 | 0.8201 | 3.4900 | 1.3729 |
| RS-CapRet [53] | 0.8430 | 0.7790 | 0.7220 | 0.6700 | 0.4720 | 0.8170 | 3.5480 | 1.3767 |
| $S^2$CPNet | **0.8722** | **0.8183** | **0.7729** | **0.7302** | **0.4822** | **0.8345** | **3.5482** | **1.3987** |

### C. Experimental Settings

The model is based on pytorch 1.8 framework, and all the experiments are conducted on an NVIDIA 3090 GPU. The batchsize is set to 48, and the Adam is employed as the optimizer. The initial learning rate of the model is set to 1e-4 on the Sydney, UCM dataset and 2e-4 on the NWPU dataset. The number of cross attention units in the CVFF is set to 3. During the training process, if the METROE not increase for continuous 5 epoch, the learning rate will adjust to one-fifth the current value. The model will terminate early if the METEOR fails to increase for continuous 20 epoch.

### D. Comparison with Existing Methods

**Results on the Sydney dataset:** Table I illustrates that the proposed $S^2$CPNet demonstrates competitive performance on the Sydney dataset across various metrics such as BLEU-4, ROUGE_L, CIDEr, and Sm. Notably, the TCE loss model achieves the highest BLEU-4 and METEOR scores. This success can be attributed to the small amount of the Sydney dataset and the fact that the model incorporation a truncated loss function, which helps model alleviate overfitting issues. The GCFGA+LSGA method excels in the CIDEr metric because it integrates global and local features of remote sensing images to obtain accurate and comprehensive semantic information. By leveraging these dual perspectives, GCFGA+LSGA captures the intricate details necessary to generate high-quality captions. Compared to these methods, our $S^2$CPNet integrates regional and semantic features of potential targets within the images, enabling it to capture more complete and precise target features. Consequently, it achieves satisfactory performance across multiple metrics.

**Results on the UCM dataset:** From Table II, it can be observed that our model achieves the best performance across all metrics on the UCM dataset. The Puret model also demonstrates competitive results due to its fully Transformer-based architecture for extracting visual and textual features, which effectively reduces the semantic gap between cross-modal features. Our proposed model enhances the understanding of

TABLE III
EXPERIMENTS RESULTS ON NWPU DATASET. THE BOLD AND UNDERLINED INDICATE THE BEST AND SECOND BEST RESULTS. THE ∗ DENOTES THE RESULTS OF OUR RE-IMPLEMENTED.

| Method | BLEU-1↑ | BLEU-2↑ | BLEU-3↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr↑ | Sm↑ |
|---|---|---|---|---|---|---|---|---|
| Multimodal [5] | 0.7250 | 0.6030 | 0.5180 | 0.4550 | 0.3360 | 0.5910 | 1.1790 | 0.6402 |
| Soft Attention [46] | 0.7310 | 0.6090 | 0.5250 | 0.4620 | 0.3390 | 0.5990 | 1.1360 | 0.6340 |
| Hard Attention [46] | 0.7330 | 0.6100 | 0.5270 | 0.4640 | 0.3400 | 0.6000 | 1.1030 | 0.6267 |
| FC-Att+LSTM [56] | 0.7360 | 0.6150 | 0.5320 | 0.4690 | 0.3380 | 0.6000 | 1.2310 | 0.6595 |
| SM-Att+LSTM [56] | 0.7390 | 0.6170 | 0.5320 | 0.4680 | 0.3300 | 0.5930 | 1.2360 | 0.6567 |
| MLCANet [50] | 0.7540 | 0.6240 | 0.5410 | 0.4780 | 0.3370 | 0.6010 | 1.2640 | 0.6700 |
| BUTD∗ [19] | 0.8718 | 0.7878 | 0.7206 | 0.6669 | 0.4304 | 0.7612 | 1.8866 | 0.9362 |
| AoANet∗ [20] | 0.8755 | 0.7754 | 0.6917 | 0.6207 | 0.3907 | 0.7301 | 1.8490 | 0.8976 |
| MLAT∗ [51] | 0.8527 | 0.7674 | 0.7007 | 0.6479 | 0.4327 | 0.7496 | 1.8556 | 0.9214 |
| Clipcap∗ [54] | 0.8394 | 0.7421 | 0.6619 | 0.5954 | 0.4142 | 0.7385 | 1.7265 | 0.8685 |
| PureT [55] | **0.8880** | 0.8031 | 0.7330 | 0.6750 | 0.4232 | 0.7584 | 1.9512 | 0.9519 |
| RS-CapRet [53] | 0.8710 | 0.7870 | 0.7170 | 0.6560 | 0.4360 | 0.7760 | 1.9290 | 0.9492 |
| $S^2$CPNet | 0.8801 | **0.8081** | **0.7470** | **0.6965** | **0.4646** | **0.7914** | **2.0392** | **0.9979** |

visual features by incorporating a target category generation decoder. This decoder helps the model to better comprehend and classify visual content, facilitating the production of more accurate linguistic descriptions. Additionally, our model introduces a gated guided caption generation decoder, which optimally leverages both visual and textual features. This design allows the model to adaptively fuse different features, which can generate more accurate and coherent descriptions. By integrating these components, our model outperforms existing methods in producing precise and contextually relevant captions.

**Results on the NWPU dataset:** Table III indicates that our model achieves the best performance on the NWPU dataset as well. The RS-CapRet model based on LLM technology, which obtains the second score in metrics such as METEOR and ROUGE_L, demonstrating the potential of LLM technology in RSIC tasks. Our proposed $S^2$CPNet outperforms the PureT model by achieving a 2.15% higher score in BLEU-4, indicating that the descriptions generated by our method are more precise. The consistent performance of our model across different datasets, regardless of their scale, further attests to the robustness and versatility of our approach.

### *E. Visualization results*

To fully showcase the advantages of this method, we select some typical remote sensing scenes with corresponding caption, as illustrated in Fig. 4. These results show that the proposed method is able to generate more comprehensive and accurate linguistic descriptions compared to existing methods.

On the Sydney dataset, as depicted Fig. (B), the Baseline model generates statements with repetitive words. In contrast, the $S^2$CPNet model not only produces smoother statements, but also generates more complete descriptions containing crucial elements such as rivers, residential areas, and houses. Meanwhile, the target category decoder can produce a category sequence for the presence of the target. The enhancement can be attributed to the gated fusion attention mechanism designed in our model, which effectively leverages both image and text features. This leads to more coherent descriptions and significantly reduces the issue of word repetition.

On the UCM dataset, as shown in Fig. (F), the $S^2$CPNet model generates more complete and accuracy descriptions compared to the Baseline model. For instance, in Fig. (H), the Baseline model incorrectly predicts white buildings as gray, and our $S^2$CPNet model can accurately describe them. This improvement is due to the CVFF module integrated into our model, which acquires more comprehensive visual features, thereby enhancing the accuracy of the generated caption. In addition, the multi-target classification task helps the model recognize the feature of "baseball", "diamond", "sand" and "weed", which further enhances the ability of model to comprehend and interpret images. This capability is particularly important in complex scenes.

On the NWPU dataset, as illustrated in Fig. (K), the $S^2$CPNet model generates captions that are more comprehensive in content compared with the Baseline model. The descriptions encompass a broader range of elements and provide a clearer depiction of the scene. Furthermore, in Fig. (N), the Baseline model incorrectly predicts the number of basketball courts, demonstrating a significant shortcoming. In contrast, our model leverages the spatial information by introducing corrseponding mask. Therefore, the proposed $S^2$CPNet is more sensitive to the number of targets and produces reliable captions.

### *F. Ablation Study*

In this section, we design ablation in the UCM dataset to demonstrate the effectiveness of the design module.

**Only baseline:** We utilize two variants as baselines: the first variant (Baseline1) employs ResNet50 as the encoder and a Transformer network as the decoder, while the second variant (Baseline2) replaces the standard Transformer network with a gated guide Transformer mechanism. As shown in Table IV, the improved Transformer decoder gains improvements in BELU4, METEOR, ROUGE_L, and CIDEr metrics compared to Baseline1. These enhancements are attributed to the ability of model to fully leverage both visual and textual features, thereby significantly improving the accuracy of the generated descriptions. Consequently, we adopt the ResNet50 with a gated-guided Transformer as our Baseline.

## Sydney

**Base:** A residential area with many houses arranged and some roads go through this area.
**SSCPNet:** A meadow with some **bushes** and big bunkers on the meadow.
**Nouns:** meadow bushes bunkers paths
**GT1:** A meadow with some green bushes on it while some roads passed by .
**GT2:** A meadow with bushes and roads on it while some houses beside .
**GT3:** This is a meadow with some green bushes on it while some roads passed by .
**GT4:** Some green bushes and paths on the meadow with some houses beside .
**GT5:** There are some green bushes and paths on the meadow while some houses beside .

**Base:** A residential area with many green waters **with a wide wide river**.
**SSCPNet:** A river with dark green waters on the residential area **with many houses on the bank**.
**Nouns:** river waters river
**GT1:** A wide river with deep green waters and some boats on it .
**GT2:** There are some white boats on the river with some houses on the bank .
**GT3:**Some small boats on the river with some houses on the bank .
**GT4:**There are some small boats on the wide river with some houses on the bank .
**GT5:** A wide river with deep green waters and some boats on it .

**Base:** There are many white buildings on the river **with a river on the river**.
**SSCPNet:** There are many white buildings on the banks of a curved river.
**Nouns:** river banks river
**GT1:** A river goes through a industrial area with many white buildings on the banks of the river.
**GT2:** There are many white buildings on the banks of a river.
**GT3:** Lots of white buildings are on the bankes of a curved river .
**GT4:** Lots of white buildings are on the bankes of a curved river.
**GT5:** There are many white buildings on the banks of a curved river.

**Base:** A straight white **white** marking lines on the straight **straight straight straight straight straight straight straight straight straight straight**.
**SSCPNet:** This is a part of mark lines on it.
**Nouns:** part mark lines lawns
**GT1:** A narrow runway with ssome lawns beside .
**GT2:** A rectangular runway goes through the lawn .
**GT3:** There is a narrow rectangular runway going through the lawn .
**GT4:** A rectangular runway goes through the lawn .
**GT5:** A narrow rectangular runway goes through the lawn .

**Base:** There is **a big green with some green water** with some green river.
**SSCPNet:** This is a part of deep green sparkling sea.
**Nouns:** part ocean water beach
**GT1:** A narrow runway on the river bank .
**GT2:** There is a narrow curved runway on the river bank .
**GT3:** A narrow curved runway are on the river bank .
**GT4:** There is a narrow runway on the river bank while a lawn beside.
**GT5:** A narrow runway goes through a lawn with a river beside .

## UCM

**Base:** This is a old baseball diamond.
**SSCPNet:** It is a very old baseball diamond **compose of sand and weeds**.
**Nouns:** baseball diamond sand weeds
**GT1:** This is a baseball diamond
**GT2:** It is an old baseball diamond with sand and weeds.
**GT3:** It is a small baseball diamond.
**GT4:** An old baseball diamond compose of sand and weeds .
**GT5:** An old baseball diamond compose of sand and weeds .

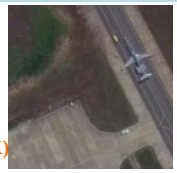
**Base:**This is a dense residential area with lots of houses arranged neatly.
**SSCPNet:**Some buildings with grey roofs.
**Nouns:** buildings roofs
**GT1:**There is are some buildings with grey and orange roofs.
**GT2:** Some buildings with grey and orange roofs.
**GT3:** There are some buildings with different colors of roofs.
**GT4:** There are some buildings with plants and cars beside them.
**GT5:** Some buildings and palnts.

**Base:** Some buildings with **grey** roofs.
**SSCPNet:**Some buildings with **white** roofs.
**Nouns:** buildings roofs
**GT1:** There is are some buildings with white roofs .
**GT2:** Some buildings with white roofs .
**GT3:** There are some buildings with some plants arround them.
**GT4:** There are some buildings with some plants beside them .
**GT5:** Buildings and plants.

**Base:** This is a chaparral.
**SSCPNet:** Some grey plants are scattered **in the loess ground**.
**Nouns:** plants ground
**GT1:** This is a chaparral .
**GT2:** Lots of plants scattered on the ground.
**GT3:** Lots of plants scattered in the loess ground.
**GT4:** There are some grey plants scattered on the ground.
**GT5:** Some grey plants are scattered in the loess ground .

**Base:** This is a medium residential area with houses arranged neatly.
**SSCPNet:** There are lots of houses arranged neatly **and a road goes through them**.
**Nouns:** area houses roads
**GT1:** This is a dense residential area .
**GT2:** It is a dense residential area with lots of houses arranged in lines.
**GT3:** There ars lots of houses arranged in lines and a road goes through them.
**GT4:** Lots of houses arranged in lines and a road goes through them .
**GT5:** There are lots of houses arranged in lines and a road goes through them.

## NWPU



**Base:** The runway is on the bare land .
**SSCPNet:** The runway is on the bare land **next to the grass** .
**Nouns:** runway bare land grass
**GT1:** A gray plane on the runway and the lawn beside .
**GT2:** A grey plane is on the runway by the lawn .
**GT3:** There is an airplane on the runway with a large lawn by the runway .
**GT4:** A plane is parked on the runway next to the grass.
**GT5:** There is a plane on the runway beside the grass.

**Base:** **Two** planes are parked on the runway .
**SSCPNet:** **A** plane parked on the open area **with some yellow marks**.
**Nouns:** plane ground
**GT1:** A plane parked in a line on the airport with some marks .
**GT2:** A white plane was parked on the instruction line.
**GT3:** An airplane parked in an open area with many containers next to it .
**GT4: A** plane is parked on the open space .
**GT5:** There is 1 plane on the ground marked.

**Base:** There is **a** white house on the edge of the baseball field .
**SSCPNet:** There are **two** baseball diamonds in the grass surrounded by the houses and roads.
**Nouns:** baseball diamonds trees road
**GT1:** Two adjacent baseball diamonds on the lawn with some trees and some roads beside .
**GT2:** There are four baseball fields in the picture .
**GT3:** There are two baseball diamonds in the grass along the roads surrounded by brushwood.
**GT4:** There are two baseball diamonds between many trees and a road .
**GT5:** Many trees are beside the two baseball diamonds .

**Base:** There are **two** basketball courts in the middle of grass next to some trees.
**SSCPNet:** There are **four** baseball diamonds in the grass surrounded by the roads.
**Nouns:** basketball courts lawn trees road
**GT1:** Four baseball diamonds adjacent to each other make up a circle and some buildings beside.
**GT2:** This baseball field looks like a semicircle with two white houses beside it.
**GT3:** There are four basebale Diamonds in The grass along the roads.
**GT4:** There are four baseball diamonds field in a circle.
**GT5:** Several buildings are between four baseball diamonds.

**Base:** The **grass** is beside the beach.
**SSCPNet:** **Many buildings** are beside the beach .
**Nouns:** beach coast buildings beach
**GT1:** The type of beach is rocky.
**Nouns:** coast and the residential area beside.
**GT2:** The sea has ripples.
**GT3:** Warm ocean and houses on the beach .
**GT4:** There are many buildings and grass beside the beach.
**GT5:** Some buildings are beside the beach .

Fig. 4. Examples of sentences generated by $S^2$CPNet from Sydney dataset (first row), UCM dataset (second row), and NWPU dataset (third row). The red words and blue words indicates the mistakes and advantages of our method compared to Baseline.

TABLE IV
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATE THE BEST RESULTS.

| Components | | | | Metric | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | ResNet50 | Transformer | Dual-Gated Transformer | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr ↑ | Sm ↑ |
| Baseline1 | ✓ | ✓ | | **0.8589** | **0.8029** | **0.7481** | 0.6942 | 0.4501 | 0.8026 | 3.1708 | 1.2794 |
| Baseline2 | ✓ | | ✓ | 0.8536 | 0.7989 | **0.7481** | **0.7016** | **0.4560** | **0.8040** | **3.3007** | **1.3155** |

TABLE V
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATES THE BEST RESULTS.

| Components | | | | | | Metric | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Baseline | Objcet masks | SSFE | CVFF | Nouns Predict | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr ↑ | Sm ↑ |
| Model1 | ✓ | | | | | 0.8536 | 0.7989 | 0.7481 | 0.7016 | 0.4506 | 0.8040 | 3.3307 | 1.3217 |
| Model2 | ✓ | ✓ | | ✓ | | 0.8606 | 0.8056 | 0.7606 | 0.7208 | 0.4629 | 0.8125 | 3.4428 | 1.3597 |
| Model3 | ✓ | ✓ | | ✓ | ✓ | 0.8499 | 0.8061 | 0.7688 | **0.7332** | 0.4662 | 0.8018 | 3.4063 | 1.3518 |
| $S^2$CPNet | ✓ | ✓ | ✓ | ✓ | ✓ | **0.8722** | **0.8183** | **0.7729** | 0.7302 | **0.4822** | **0.8345** | **3.5482** | **1.3987** |

TABLE VI
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATES THE BEST RESULTS.

| Components | | | | | | Metric | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Baseline | Object masks | add | contact | CVFF | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr ↑ | Sm ↑ |
| Model2-A | ✓ | ✓ | ✓ | | | 0.8458 | 0.7835 | 0.7282 | 0.6776 | 0.4417 | 0.7971 | 3.2148 | 1.2828 |
| Model2-B | ✓ | ✓ | | ✓ | | 0.8412 | 0.7841 | 0.7315 | 0.6871 | **0.4662** | 0.8018 | 3.4063 | 1.3403 |
| Model2 | ✓ | ✓ | | | ✓ | **0.8606** | **0.8056** | **0.7606** | **0.7208** | 0.4629 | **0.8125** | **3.4428** | **1.3597** |

TABLE VII
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATES THE BEST RESULTS.

| Components | | | | Metric | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Model3 | Semantic Enhance | Spatial Enhance | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | CIDEr ↑ | Sm ↑ |
| Model4-A | ✓ | ✓ | | 0.8621 | 0.8132 | 0.7663 | 0.7214 | 0.4613 | 0.8138 | 3.3598 | 1.3339 |
| Model4-B | ✓ | | ✓ | 0.8714 | **0.8325** | **0.7941** | **0.7579** | 0.4752 | 0.8264 | 3.4396 | 1.3747 |
| $S^2$CPNet | ✓ | ✓ | ✓ | **0.8722** | 0.8183 | 0.7729 | 0.7302 | **0.4822** | **0.8345** | **3.5482** | **1.3987** |

**Introducing Segmentation Masks and CVFF:** Based on Baseline, we take the remote sensing image and the corresponding target segmentation mask as the input to the model and use the CVFF module to realize the fusion of these features. As shown in Table V, incorporating the segmentation masks and utilizing the CVFF module to aggregate fine-grained semantic features with contour features enables the model to obtain more comprehensive visual features. Consequently, the captions generated by the model show significant improvements, with increases of 1.92% in the BLEU4 metric and 11.21% in the CIDEr metric, respectively.

In addition, we also design different feature fusion methods to replace the cross-view feature fusion module, as shown in table VI, including feature addition (Model2-A) and feature contaction (Model2-B). However, the performance of the models all show a decrease compared to Model2, which proves that simple fusion operations are insufficient for effectively integrating semantic and contour features.

**Introduction of multi-target classification task:** As shown in table V, based on Model2, we introduce the task of multi-target classification to achieve feature alignment between visual features and nouns. This enhancement improves the semantic comprehension of remote sensing images. As a result, the model shows an improvement of 1.24% in the BLEU-4 metric and 0.33% in the METEOR metric, indicating that the produced predictions are more consistent with the ground truth.

**Introduction of semantic and spatila feature enhance sub-module:** As shown in table VII, based on Model3, we incorporate a semantic feature enhancement sub-module and a spatial feature enhancement sub-module separately. The experimental results demonstrate a 1.2% improvement in the ROUGE_L score with the inclusion of the spatial feature enhancement sub-module. Furthermore, the integration of this sub-module leads to consistent improvements across various evaluation metrics, with the Sm indicator showing a notable increase of 2.29%. This indicates that the overall quality of the captions generated by the model improves significantly. It is due to the fact that the spatial feature enhancement sub-module can strengthen the boundary regions of objects and provide the model with sufficient target position relations, which is crucial for the RSIC task. Moreover, the combination of Model3

with both spatial and semantic feature enhancement sub-modules further enhances the performance of model, which demonstrates the effectiveness of this combination.

## V. Conclusion

This paper proposes a RSIC algorithm that combines fine-grained semantic information with region regions, named as $S^2$CPNet. In terms of dataset, we provide masks for potential targets in the RSIC dataset, allowing the model to leverage regional features effectively. To fully align visual features with textual features, we use NLTK to extract nouns from captions and form a sequence of target categories. In terms of algorithms, the semantic enhancement and spatial enhancement sub-modules are introduced to enhance target semantic features and spatial features. Then, a cross-view feature fusion module is developed to effectively integrate fine-grained semantic features with regional features, which can generate a comprehensive visual feature. Further, the multi-target classification task is introduced to align visual features with text features. In this way, the model can accurately describe the objects in the image. Finally, we propose an improved decoder that employs gated guide mechanism to utilize visual and textual features efficiently. The proposed approach produces accurate and fluent captions on the RSIC dataset, and ablation experiments demonstrate the necessity of these improvements.

In the future, we will focus on building a new RSIC dataset that provides both detection and segmentation of targets as well as image caption. This will further promote the development of the RSIC.

## References


[1] H. Deng, Y. Xie, Q. Wang, J. Wang, W. Ruan, W. Liu, and Y.-J. Liu, "Cdkm: Common and distinct knowledge mining network with content interaction for dense captioning," *IEEE Transactions on Multimedia*, pp. 1–15, 2024.

[2] Z. Yang, Q. Li, Y. Yuan, and Q. Wang, "Hcnet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[3] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 4, pp. 297–312, 2019.

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

[5] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International conference on computer, information and telecommunication systems (Cits)*. IEEE, 2016, pp. 1–5.

[6] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 436–440, 2021.

[7] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 2001–2005, 2021.

[8] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2020.

[9] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[10] X. Huang, J. Wang, Y. Tang, Z. Zhang, H. Hu, J. Lu, L. Wang, and Z. Liu, "Segment and caption anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 405–13 417.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[12] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, 2023.

[13] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, 2022.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[16] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[17] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[18] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[19] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[20] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.

[21] Z. Fei, "Attention-aligned transformer for image captioning," in *proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 607–615.

[22] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei, "Semantic-conditional diffusion networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 359–23 368.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[24] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[25] Y. Ge, X. Zeng, J. S. Huffman, T.-Y. Lin, M.-Y. Liu, and Y. Cui, "Visual fact checker: Enabling high-fidelity detailed caption generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 033–14 042.

[26] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[27] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 436–440, 2020.

[28] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2019.

[29] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 2001–2005, 2020.

[30] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[31] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[32] L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, "A multiscale grouping transformer with clip latents for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[33] R. Du, W. Cao, W. Zhang, G. Zhi, X. Sun, S. Li, and J. Li, "From plane to hierarchy: Deformable transformer for remote sensing image captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

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

[35] D. Zhao, B. Yuan, Z. Chen, T. Li, Z. Liu, W. Li, and Y. Gao, "Panoptic perception: A novel task and fine-grained dataset for universal remote sensing image interpretation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[36] X. Ye, S. Wang, Y. Gu, J. Wang, R. Wang, B. Hou, F. Giunchiglia, and L. Jiao, "A joint-training two-stage method for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[37] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multilabel classification for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[38] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 532–10 543, 2020.

[39] L. Meng, J. Wang, Y. Yang, and L. Xiao, "Prior knowledge-guided transformer for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[40] K. Zhao and W. Xiong, "Cooperative connection transformer for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[41] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *arXiv preprint arXiv:2307.15266*, 2023.

[42] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image-text alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[43] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

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

[47] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1985–2000, 2019.

[48] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5246–5257, 2020.

[49] G. Hoxha and F. Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[50] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "Nwpu-captions dataset and mlca-net for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

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

[52] G. Hoxha, G. Scuccato, and F. Melgani, "Improving image captioning systems with postprocessing strategies," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

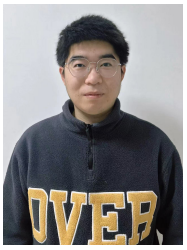
[53] J. D. Silva, J. Magalhães, D. Tuia, and B. Martins, "Large language models for captioning and retrieving remote sensing images," *arXiv preprint arXiv:2402.06475*, 2024.

[54] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[55] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2585–2594.

[56] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.


**Qi Wang** (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/)



**Zhigang Yang** is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.

**Weiping Ni** received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2004, the M.S. degree from the National University of Defense Technology, Changsha, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2016. Since 2014, he has been a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an. His research interests include remote sensing image processing, automatic target recognition, and computer vision.

**Junzheng Wu** received the B.Sc. degree in automation from Tsinghua University, Beijing, China, in 2008, the M. Sc. degree in signal and information processing from Northwest Institute of Nuclear Technology, Xi'an, China, in 2011, and the Ph. D. in information and communication engineering from National University of Defense Technology, Changsha, China, in 2022, respectively. Currently, he is an Associate researcher with the department of remote sensing, Northwest Institute of Nuclear Technology, Xi'an, China. His research interests include processing of remote sensing images and machine learning.

**Qiang Li** (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.