# HCNet: Hierarchical Feature Aggregation and Cross-Modal Feature Alignment for Remote Sensing Image Captioning

Zhigang Yang, Qiang Li, *Member, IEEE*, Yuan Yuan, *Senior Member, IEEE*, Qi Wang, *Senior Member, IEEE*

***Abstract*—Remote sensing image captioning aims to describe the crucial objects from remote sensing images in the form of natural language. The inefficient utilization of object texture and semantic features in images, along with the ineffective cross-modal alignment between image and text features, are the primary factors that impact the model to generate high-quality captions. To alleviate this trouble, this paper presents a network for remote sensing image captioning, namely HCNet, including hierarchical feature aggregation and cross-modal feature alignment. Specifically, a hierarchical feature aggregation module is proposed to obtain a comprehensive representation of vision features, which is beneficial for producing accurate descriptions. Considering the disparities between different modal features, we design a cross-modal feature interaction module in the decoder to facilitate feature alignment. It can fully utilize cross-modal features to localize critical objects. Besides, a cross-modal feature align loss is introduced to realize the alignment between image and text features. Extensive experiments show our HCNet can achieve satisfactory performance. Especially, we demonstrate significant performance improvements of +14.15% CIDEr score on NWPU datasets compared to existing approaches. The source code is publicly available at https://github.com/CVer-Yang/HCNet.**

***Index Terms*—Remote sensing, image caption, feature aggregation, feature alignment, attention mechanism.**

## I. Introduction

REMOTE sensing image captioning (RSIC) is a comprehensive task that combines natural language processing and computer vision, which has attracted extensive interest due to its significant application, such as image retrieval [2], scene understanding [3], change detection [4] and other fields. Compared with tasks such as image super-resolution [5] [6], object detection [7], and semantic segmentation [8], image captioning not only recognizes objects in the image, but also captures the relationships of objects.

RSIC faces several special challenges: (1) Scale variations. Objects with the same category in remote sensing images may display considerable scale variations. For example, an airport may contain airplanes with different sizes. To generate accurate and meaningful captions, RSIC models need to integrate features across different scales and capture the relationship between them. (2) Cross-modal disparities. Given a remote sensing image, the model generates corresponding textual descriptions. There exist substantial modality differences between the input and output, requiring precise alignment of image and text features during the text generation and achieving feature fusion effectively. The above challenges have a significant impact on producing high-quality captions, which makes the RSIC more difficult than other remote sensing tasks.

The RSIC models typically can be split into two stages: vision feature extraction and text generation. In the vision feature extraction stage, these methods commonly employ convolutional neural networks (CNNs) [9] or vision Transformer [10] to extract features from images and encode them into high-dimensional feature vectors. In the text generation stage, recurrent neural networks or Transformer [11] are utilized to transfer the high-dimensional feature vectors into corresponding text. Currently, most methods design the multi-scale features integration, several attention mechanism combination [15], [16], and auxiliary tasks [17] to construct the models for text generation. Although these approaches have achieved promising results, the following problems still exist: (1) In the vision feature extraction stage, these feature fuse methods [12]–[14] cannot effectively utilize the texture and semantic information in different scale features, which results in inaccurate category descriptions in the generated captions. (2) During the text generation stage, the multi-modal features are directly fed into the long short-term memory (LSTM) network without effective alignment, which will cause semantic confusion and limit the ability of model to accurately locate the objects. Therefore, *how to effectively* ***obtain comprehensive vision representation*** *and* ***align cross-modal features*** *is significant for RSIC task.*

To alleviate these problems, this paper proposes a novel method for RSIC called HCNet, which utilizes hierarchical feature aggregation and cross-modal feature alignment. The model aims to acquire comprehensive visual features and produce precise hidden states. Firstly, we employ ResNet50 [18] as the encoder to extract features at multiple resolutions. This allows the model to implicitly capture rich and detailed information from the input images. To further strengthen the feature representation capability, a hierarchical feature aggregation module (HFAM) is proposed to effectively mines both texture and semantic features. During decoder stage, we utilize a dual-LSTM architecture to generate coherent and contextually relevant captions. Additionally, a cross-modal feature interaction module (CFIM) is designed to facilitate the feature alignment from different modalities. The module enables effective information exchange remote sensing images

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2041. All authors are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: zgyang@mail.nwpu.edu.cn, liqmges@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com) (Corresponding author: Qi Wang, Qiang Li.)

and text representations, leading to improved caption generation. Furthermore, a cross-modal feature alignment loss is introduced to enhance the quality of the produced captions. This loss guides HCNet to mitigate the disparity between image and text features. The main contributions of this work are summarized as follows:

- We propose an HCNet for RSIC task, which incorporates two key modules: HFAM and CFIM. We verify that the proposed model with the two main modules exhibits superior performance, which reveals effectiveness of two modules for model learning and optimization. Compared with the previous RSIC methods, the proposed HCNet can produce accurate and complete captions.
- The HFAM with texture completion submodule and semantic distribution submodule is proposed, which can provide rich structure features from shallow features and precious semantic features from deep features to reduce caption confusion.
- The CFIM is designed to enhance the alignment between image and different modal features, thereby enabling the model to generate accurate hidden states. Besides, we introduce a cross-modal feature align loss to minimize the differences between image and text features.

The rest of this paper is organized as follows. Section II provides related work in natural image captioning (NIC) and RSIC. In Section III, we describe the details of our proposed HCNet. The experimental results are analyzed in Section IV. Finally, Section V offers the conclusion.

## II. Related Work

In this section, we review the existing image captioning methods. They are roughly divided into NIC and RSIC.

### A. Natural Image Captioning

Currently, the mainstream NIC methods for caption generation are usually built by using the encoder-decoder framework. Here, the encoder is employed to extract features and the decoder generates corresponding text descriptions. For instance, Vinyals et al. [19] utilize CNN to convert image into visual features, and these features are then input into the LSTM network for text generation. Anderson et al. [20] introduce a dual-LSTM decoder, where Att-LSTM focuses on important image content and L-LSTM is responsible for word generation. It results in a substantial improvement in caption quality. Later, some researchers combine attention mechanism to enhance the caption quality. Xu et al. [21] integrate a positional attention mechanism into the decoder stage to selectively focus on information from different positions during word generation. Additionally, Chen et al. [22] introduce attention mechanism into the encoder, and leverage key image features to generate words. Similar methods have [23] and [24]. Inspired by the benefits of Transformer, a collaborative network is proposed in [25] to realize the fusion of region and grid features for captioning task. Wang et al. [26] utilize Swin Transformer [10] and Transformer respectively to extract image features and generate words. Kuo et al. [27] employ the shared encoder to obtain object features, grid features, and text features of images, while the hierarchical decoder is designed to weigh these input features. This method achieves an obvious enhancement in CIDEr. Luo et al. [28] develop a semantic diffusion network, and adopt a pre-trained retrieval model to obtain descriptions related to visual content. This approach demonstrates the potential of the diffusion model in image captioning. Ramos et al. [29] propose a caption generation framework that combines retrieval techniques, the model is light to train and has perfect domain adaptation capabilities. Ma et al. [30] combine diffusion model with large language model techniques to create a synthetic image-text dataset that improves the ability of model to describe images from different perspectives. Although these methods have made progress in the description of natural images, their performance can be significantly degraded for complex remote sensing images when applied directly. In this case, we can refer to these methods to design corresponding algorithms for RSIC.

### B. Remote Sensing Image Captioning

There are some differences between NIC and RSIC, including data source, content characteristics, linguistic representation. In the past decade, some improved RSIC methods have been proposed [31]. For instance, Huang et al. [12] combine the denoising strategy with feature fusion to improve the caption quality. Ma et al. [14] propose two feature fusion modules and introduce object detection to obtain rich visual feature representation. Similar methods have [13] and [42]. To integrate the cross-modal features accurately, Zhang et al. [15] propose the linguistic state guided attention in the decoder to remove irrelevant information in the fused feature maps. Similarly, Ye et al. [17] design a multi-label classification task to capture prior knowledge, and a semantic gate module is developed to steer the generation of hidden states. To enhance the interpretability of model, a word-sentence framework is developed in [32], which first extracts words from images and then generates sorted description sentences by Transformer. Hoxha et al. [33] propose an SVM-based decoder that converts visual vectors into sentences and can generate high-quality textual descriptions in small datasets. Additionally, researchers also propose some methods based on Transformer, such as Chen et al. [34] construct the Swin Transformer network to explore features of images, and use the Transformer network as a decoder to generate description sentences. To alleviate the over-fitting problem, a truncation cross entropy loss is proposed in [35] and achieves superior performance on small datasets. Lu et al. [36] take the audio input as active attention to generate accurate captions. Zhao et al. [37] provide regional-level annotations for remote sensing caption datasets and propose a method that fully utilizes of regional features with grid features. Recently, Hu et al. [38] employ large language model (LLM) techniques to explore the task of RSIC. Although existing RSIC methods have yielded satisfactory performances, they are still unable to obtain accurate captions. On the one hand, these methods fail to efficiently leverage the spatial information and semantic features in the fused feature maps. Therefore, the models cannot produce a comprehensive visual feature representation.

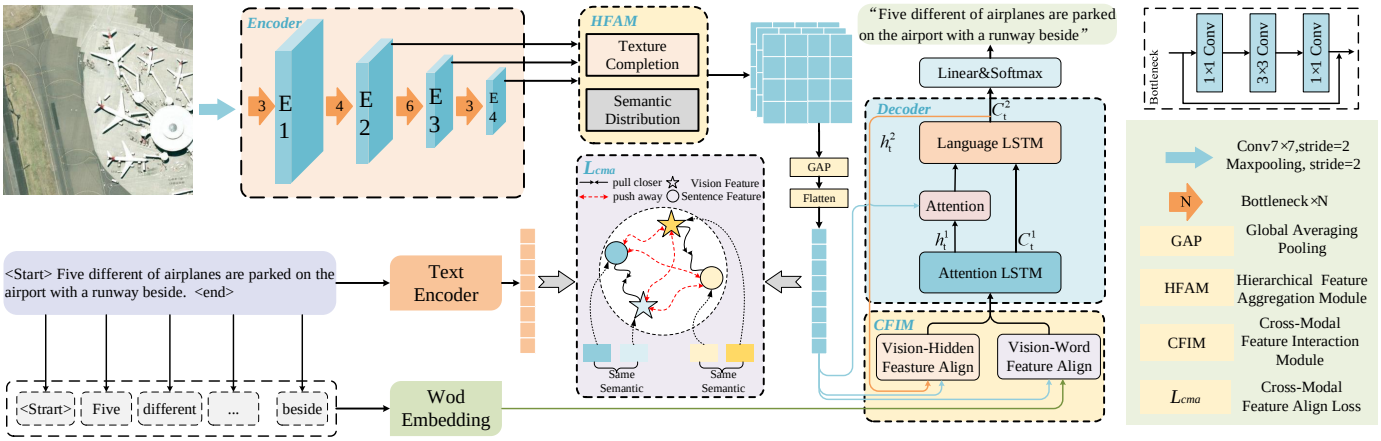



Fig. 1. The overall framework of the proposed HCNet, and it is divided into four parts: the encoder, the HFAM, the CFIM, and the dual-LSTM decoder.

On the other hand, these methods [15], [17], [34] do not fully aggregate cross-modal features in the decoder, which results in the generation of inaccurate hidden states. These factors make it difficult to generate accurate and complete captions.

## III. Proposed Method

### A. Overview

In remote sensing images, objects usually display significant scale variations. Besides, there exist substantial cross-modal disparities remote sensing images the input and output for RSIC task. To overcome these challenges, we concentrate on improving this task in crucial aspects: hierarchical feature aggregation and cross-modal feature alignment. Fig. 1 shows the overall framework of HCNet. Firstly, the pre-trained ResNet50 on the ImageNet dataset is employed as the encoder to extract features. It yields four feature maps with different resolutions, where they are denoted as E1, E2, E3, and E4, respectively. Subsequently, the feature maps E2, E3, and E4 are input into the HFAM, which obtains a comprehensive visual feature representation. Unlike this manner that previous approaches often concatenate the generated visual features, word features, and LSTM hidden states directly and feed the fusion features into the LSTM network, we recognize the significant disparities remote sensing images different modal features, and design a cross-modal feature interaction module that align visual features with different input features. Meanwhile, the cross-modal feature align loss is introduced to minimize the distance remote sensing images the visual features produced by HFAM and the sentence features encoded by LSTM. Finally, the aligned visual features, word features, and LSTM hidden states are embedded in the dual-LSTM network with Att-LSTM and L-LSTM to produce image captions.

### B. Hierarchical Feature Aggregation Module

In contrast to natural images, remote sensing images typically capture a broad geographical area from a top-down remote perspective, where ground objects varies greatly in size and shape. It means that objects in images have small size, which makes it difficult to determine the category of objects by its appearance alone. It tends to obtain inaccurate feature representation. Interestingly, the surrounding environment provides a valuable complement to the shape, orientation, and other features of objects. A natural manner is to combine local texture features and global semantic information to build model. To this end, a HFAM is designed to generate accurate vision features. The architecture of the HFAM is illustrated in Fig. 2. Specifically, we utilize several large kernels [39] with $15 \times 15$, $11 \times 11$, and $7 \times 7$ to explore the long-distance context from E2, E3, and E4. Subsequently, the feature maps are resized to the same size as E3 with 512 channels. To effectively fuse the multi-scale features, the *texture completion submodule* and *semantic distribution submodule* are developed. The process is formulated as

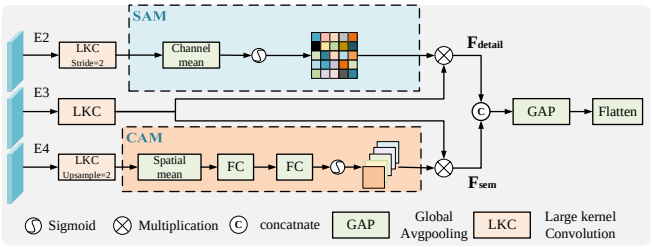



Fig. 2. Illustration of hierarchical features aggregation module (HFAM), where SAM means spatial attention mechanism and CAM means channel attention mechanism.

$$E_i^{'} = Conv(E_i), i \in \{2, 3, 4\}, \tag{1}$$

where $Conv(\cdot)$ is convolution operation.

**Texture Completion Submodule**: High-resolution feature map $E_2^{'}$ contains abundant detail features, which is conducive to help the model identify salient objects. To obtain the position weights of the hidden feature map, an average pooling operation on the channel dimension is employed, which is followed by a Sigmoid operation. The position weight is then utilized in a matrix multiplication to fuse $E_2^{'}$ and $E_3^{'}$. By doing so, the strategy achieves the information integration from the two feature maps according to position attention. It strengthens the representation of texture features in the fused feature maps.

The process can be defined as

$$F_{detail} = Sigmoid(Avg(E_2^{'})) \otimes E_3^{'}, \tag{2}$$

where $Avg(\cdot)$ represents the average pooling, and $\otimes$ denotes matrix multiplication.

**Semantic Distribution Submodule**: Low-resolution feature map $E_4^{'}$ can capture robust semantic information. It is conducive to learn the relationships remote sensing images objects in decoder. For that reason, we adopt global average pooling to condense the position feature. As a result, it obtains the channel weights of the hidden feature maps after two Full Connection (FC) layers and a Sigmoid operation. Finally, $E_3^{'}$ and $E_4^{'}$ are merged by matrix multiplication operation to yield semantic distribution. The process of semantic distribution submodule is represented as

$$F_{sem} = Sigmoid(FC(R(FC(Avg(E_4^{'}))))) \otimes E_3^{'}, \tag{3}$$

where $FC(\cdot)$ is FC operations, and $R(\cdot)$ is a Rectified Linear Unit (ReLU). Then we contact the texture feature map $F_{detail}$ and the semantic feature map $F_{sem}$, and a convolution operation is utilized to conduct feature fusion. This process produces a visual feature map $F_{vision}$ with the comprehensive feature representation. Finally, the vision feature map is reshaped to $\overline{V} \in \mathbb{R}^{196\times1024}$ by using an Avgpooling and Flatten operation, i.e.,

$$F_{vision} = Conv([F_{detail}, F_{sem}]), \tag{4}$$

$$\overline{V} = Flatten(Avg(F_{vision})), \tag{5}$$

where $[,]$ is the concatenation operation.

### C. Cross-Modal Feature Interaction Module

During the caption generation stage, the vision features $\overline{V}$ obtained by HFAM, the previous hidden state $h_{t-1}$ from LSTM network, and the word feature $F_{word}$ at the time $t$ are input into the LSTM simultaneously to generate hidden states $h_t$. Here, $h_{t-1}$ contains the text semantic information generated by the model before time $t$, while $F_{word}$ has word information at the current time $t$. Since there is the remarkable feature disparity remote sensing images different modalities, only using concatenation or other simple operations to combine these inputs may not be sufficient to achieve an effective interaction. To address this trouble, a cross-modal feature interaction module (CFIM) is designed, as shown in Fig. 3. Considering the different sources of input features, the module is mainly composed of two parts: *Hidden-Attention Guidance* and *Word-Attention Guidance*.

**Hidden-Attention Guidance:** The feature maps $h_{t-1}$ are obtained by initializing the visual features $\overline{V}$, which contains text features generated before the current time. To align visual features with hidden states $h_{t-1}$, the hidden-attention guidance submodule is developed. Firstly, we split the visual features into $\overline{V}_1 \in \mathbb{R}^{196\times512}$ and $\overline{V}_2 \in \mathbb{R}^{196\times512}$ based on the channel dimension. Then, the hidden features $h_{t-1}$ are transformed into $F_h \in \mathbb{R}^{1\times512}$ by Squeeze and Linear operations. We conduct the multiplication operation remote sensing images $\overline{V}_1$ and $F_h$, and add a Softmax operation to produce channel weights. On this basis, the visual features $\overline{V}_1$ are weighted by multiplication operation along the channel dimension by above weights. Finally, the input visual features and weighed visual feature are combined through skip connections to generate the hidden state guided visual features $F_{align1}$. The overall process is defined as

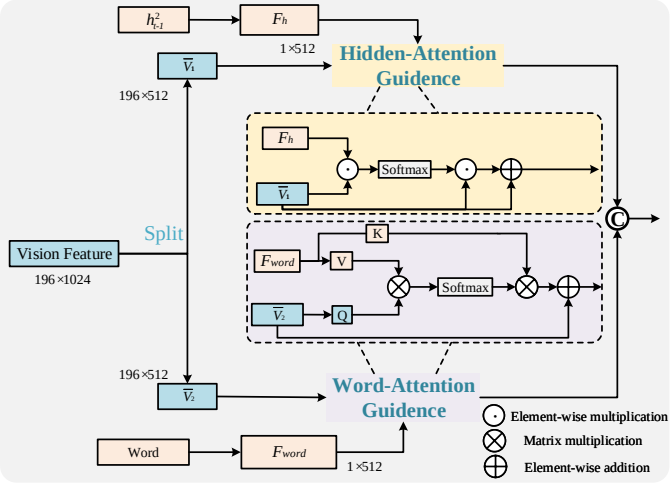



Fig. 3. Illustration of cross-modal feature interaction module (CFIM).

$$F_{align1} = Softmax(\overline{V}_1 \odot F_h) \odot \overline{V}_1 + \overline{V}_1, \tag{6}$$

where $\odot$ is element-wise multiplication.

**Word-Attention Guidance:** To align between visual features and word features, the word attention guidance submodule is designed. Specifically, we employ Squeeze and Linear operations to transform the word features into $\overline{V}_2 \in \mathbb{R}^{1\times512}$. Then, the visual features $\overline{V}_2$ and generated word features $F_{word}$ are fed into the multi-head cross-attention mechanism, where $\overline{V}_2$ obtains the query features $Q_i$, and $F_{word}$ yields the key features $K_i$ and value features $V_i$, i.e.,

$$Q_i = \overline{V}_2 W_i^Q, K_i = F_{word} W_i^K, V_i = F_{word} W_i^V \tag{7}$$

Here, $i$ is the number of heads. $W_i^Q$, $W_i^K$, and $W_i^V$ are learnable projection matrixes. Through cross attention mechanism, the visual features guided by word features is obtained. Finally, the visual features $F_{align2}$ aligned with the word features is generated by combining the $\overline{V}_2$ and weighted visual features through skip connections, i.e.,

$$F_{align2} = Concat(Softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i) + \overline{V}_2, \tag{8}$$

where $d_k$ is the dimension of $K_i$. The visual features $F_{align1}$ computed by hidden-attention guidance submodule and the visual vector $F_{align2}$ produced by word-attention guidance submodule are combined and transformed to generate the visual features $F_{align} \in \mathbb{R}^{196\times1024}$, i.e.,

$$F_{align} = Linear([F_{align1}, F_{align2}]). \tag{9}$$

### D. Dual-LSTM Decoder

Given the brief length of the remote sensing caption, we take dual-LSTM architecture [20] with Att-LSTM and L-LSTM as the decoder to decode in our approach. Firstly, the aligned

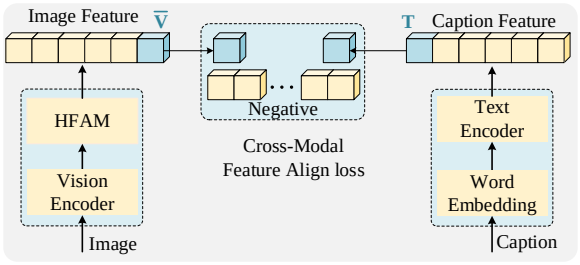



Fig. 4. Illustration of cross-modal feature align loss.

visual features $F_{align}$ generated by the CFIM, the word feature $F_{word}$ at the time $t$, and the hidden state $h_{t-1}^2$ output from the L-LSTM network are contacted and fed into the Att-LSTM to obtain the hidden states $h_t^1$ and $c_t^1$. The process is formulated as

$$h_t^1, c_t^1 = Att\text{-}LSTM(h_{t-1}^1, c_{t-1}^1, [h_{t-1}^2, F_{word}, F_{align}]). \tag{10}$$

Then, the attention mechanism is employed to refine the visual features, i.e.,

$$M = Softmax(W_a \tanh(W_v \overline{V} + W_h h_t^1)) \odot \overline{V}. \tag{11}$$

Subsequently, the refined visual features are merged with $h_t^1$ and input in the L-LSTM network to generate the hidden states $h_t^2$ and $c_t^2$, which is obtained by

$$h_t^2, c_t^2 = L\text{-}LSTM(h_{t-1}^2, c_{t-1}^2, [h_t^1, M]). \tag{12}$$

Finally, the next word is generated by performing FC layer and Softmax operation on $h_t^2$, which projects $h_t^2$ into the word space, i.e.,

$$s_t^* = Softmax(FC(h_t^2)). \tag{13}$$

### E. Loss Function

To train the model, we introduce two losses to optimize the RSIC model, i.e., cross entropy loss and cross-modal feature align loss. Specifically, the cross entropy loss is defined as

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log(p_\theta(s_t^* | s_{1:t-1}^*)), \tag{14}$$

where $s_1^*, s_2^*, \cdots, s_{t-1}^*$ represent the label words. The difference between image and text features makes it challenging for the model to obtain high-quality captions. Inspired by Contrastive Language-Image Pretraining (CLIP) technology [40], the cross-modal feature align loss is introduced to align the global vision features $\overline{V}$ in HFAM with the semantic features $T$ in the caption. As illustrated in Fig. 4, we utilize the LSTM as text encoder to extract the sentence semantic information $T$ from the caption. Then, the similarity between image-text pairs in a batch is computed by

$$\begin{aligned} \mathcal{L}_{I2T} &= -\log \frac{\exp(\overline{V}_i \odot T_i)}{\sum_{j=1}^{n} \exp(\overline{V}_i \odot T_j)}, \\ \mathcal{L}_{T2I} &= -\log \frac{\exp(\overline{V}_i \odot T_i)}{\sum_{j=1}^{n} \exp(\overline{V}_j \odot T_i)}, \end{aligned} \tag{15}$$

where $n$ represents the total number of the captions in a batch. The cross-modal feature align loss is defined as

$$\mathcal{L}_{cma} = (L_{I2T} + L_{T2I})/2. \tag{16}$$

According to above two losses, the total loss function of our method is

$$\mathcal{L}_{loss} = \lambda \mathcal{L}_{cap} + \mathcal{L}_{cma}, \tag{17}$$

where $\lambda$ is balance factor. Finally, the weight is set to 3 in this paper.

## IV. Experiments

In this section, extensive experiments on RSIC dataset are performed to demonstrate the effectiveness of HCNet. Firstly, the dataset, evaluation metrics, and experiment settings are introduced. Then, comparison and ablation study are conducted and analyzed. We also provide some visual results. Finally, we discuss the selection of hyperparameters in our experiments.

### A. Datasets

The experiment use Sydney, UCM, and NWPU caption datasets for evaluation. These datasets are derived from the scene classification datasets. The Sydney caption dataset contains 613 images with 0.5 meters/pixel, and each image is annotated with five captions. The UCM caption dataset includes 2,100 images with 0.3048 meters/pixel and covers 21 scenes, where 10,500 sentences are available in this dataset. The NWPU caption dataset contains 31,500 images and 157,500 sentences. To the best of our knowledge, it is currently the largest dataset for RSIC. We follow the [41] [42] to split the dataset, which aims to ensure a fair comparison.

### B. Evaluation Metrics

**BLEU** [43]: The generated sentences are evaluated by calculating the match of n-grams between the predicted captions and ground truth (GT). Here, n-gram is a sequence of n continuous words with values ranging from 1 to 4.

**METEOR** [44]: It mainly evaluates machine translation task. Compared to the BLEU metric, METEOR considers word variations and synonyms of the same stem, which allows for more flexibility in calculating similarities.

**ROUGE_L** [45]: It measures the similarity by calculating the F-measure given the longest common subsequence between the generated captions and references.

**CIDEr** [46]: The metric is used to evaluate image captioning task, which transforms the caption into a term frequency inverse-document frequency and weight n-gram. It focuses on the occurrence of keywords.

**Sm**: Sm metric is calculated as the arithmetic mean of BLEU-4, METEOR, ROUGE_L, and CIDEr metrics. It can comprehensively evaluate the quality of generated captions.

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 [31] | 0.4913 | 0.3412 | 0.2760 | 0.2314 | 0.1930 | 0.4201 | 0.9164 | 0.4402 |
| Sound-a-a [36] | 0.7093 | 0.6228 | 0.5393 | 0.4602 | 0.3121 | 0.5974 | 1.7477 | 0.7794 |
| Multimodal [41] | 0.6980 | 0.6130 | 0.5440 | 0.5050 | 0.3610 | 0.6370 | 2.2020 | 0.9262 |
| Soft Attenttion [31] | 0.7128 | 0.6239 | 0.5527 | 0.4924 | 0.3675 | 0.6913 | 2.0343 | 1.0471 |
| Hard Attention [31] | 0.7689 | 0.6613 | 0.5840 | 0.5170 | 0.3719 | 0.6842 | 1.9863 | 0.9541 |
| TCE loss [35] | 0.7937 | 0.7304 | **0.6717** | **0.6193** | **0.4430** | 0.7130 | 2.4042 | 1.0449 |
| Word-Sentence framework [32] | 0.7891 | 0.7094 | 0.6317 | 0.5625 | 0.4181 | 0.6922 | 2.0411 | 0.9285 |
| GVFGA+LSGA [15] | 0.7681 | 0.6846 | 0.6145 | 0.5504 | 0.3866 | 0.7030 | 2.4522 | 1.0231 |
| SVM-D BOW [33] | 0.7787 | 0.6835 | 0.6023 | 0.5305 | 0.3797 | 0.6992 | 2.2722 | 0.9704 |
| SVM-D CONC [33] | 0.7547 | 0.6711 | 0.5970 | 0.5308 | 0.3643 | 0.6746 | 2.2222 | 0.9480 |
| MLCANet [42] | **0.8310** | **0.7420** | 0.6590 | 0.5800 | 0.3900 | 0.7110 | 2.3240 | 1.0012 |
| MLAT∗ [13] | 0.7768 | 0.7034 | 0.6422 | 0.5862 | 0.3882 | 0.7167 | 2.3605 | 1.0129 |
| Post-processing [47] | 0.7837 | 0.6985 | 0.6322 | 0.5717 | 0.3949 | 0.7106 | **2.5553** | **1.0581** |
| RS-CapRet [50] | 0.7870 | 0.7000 | 0.6280 | 0.5640 | 0.3880 | 0.7070 | 2.3920 | 1.0127 |
| HCNet | 0.7686 | 0.7109 | 0.6573 | 0.6102 | 0.3980 | **0.7172** | 2.4714 | 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 [31] | 0.7016 | 0.6085 | 0.5496 | 0.5030 | 0.3464 | 0.6520 | 2.3131 | 0.9536 |
| Sound-a-a [36] | 0.7484 | 0.6837 | 0.6310 | 0.5896 | 0.3623 | 0.6579 | 2.7281 | 1.0845 |
| Multimodal [41] | 0.7100 | 0.5980 | 0.5530 | 0.4600 | 0.3430 | 0.6620 | 2.9260 | 1.0977 |
| Soft Attenttion [31] | 0.7454 | 0.6545 | 0.5855 | 0.5250 | 0.3886 | 0.7237 | 2.6124 | 1.0624 |
| Hard Attention [31] | 0.8157 | 0.7312 | 0.6702 | 0.6182 | 0.4263 | 0.7698 | 2.9947 | 1.2023 |
| TCE loss [35] | 0.8210 | 0.7622 | 0.7140 | 0.6700 | 0.4775 | 0.7567 | 2.8547 | 1.1897 |
| Word-Sentence framework [32] | 0.7931 | 0.7237 | 0.6671 | 0.6202 | 0.4395 | 0.7132 | 2.7871 | 1.1400 |
| GVFGA+LSGA [15] | 0.8319 | 0.7657 | 0.7103 | 0.6596 | 0.4436 | 0.7845 | 3.3270 | 1.3037 |
| SVM-D BOW [33] | 0.7635 | 0.6664 | 0.5869 | 0.5195 | 0.3654 | 0.6801 | 2.7142 | 1.0717 |
| SVM-D CONC [33] | 0.7653 | 0.6947 | 0.6417 | 0.5942 | 0.3702 | 0.6877 | 2.9228 | 1.1437 |
| MLCANet [42] | 0.8260 | 0.7700 | 0.7170 | 0.6680 | 0.4350 | 0.7720 | 3.2400 | 1.2787 |
| MLAT∗ [13] | 0.8226 | 0.7539 | 0.6959 | 0.6424 | 0.4268 | 0.7729 | 3.0826 | 1.2311 |
| Post-processing [47] | 0.7973 | 0.7298 | 0.6744 | 0.6262 | 0.4080 | 0.7406 | 3.0964 | 1.2186 |
| Clipcap ∗ [50] | 0.8213 | 0.7488 | 0.6872 | 0.6320 | 0.4232 | 0.7742 | 3.2391 | 1.2671 |
| PureT [26] | 0.8573 | 0.8020 | 0.7562 | 0.7129 | 0.4686 | 0.8201 | 3.4900 | 1.3729 |
| RS-CapRet [51] | 0.8430 | 0.7790 | 0.7220 | 0.6700 | 0.4720 | 0.8170 | **3.5480** | 1.3767 |
| HCNet | **0.8826** | **0.8335** | **0.7885** | **0.7449** | **0.4865** | **0.8391** | 3.5183 | **1.3972** |

### C. Experimental Settings

As for Sydney, UCM, and NWPU datasets, the images are resized to 224×224 pixels, and random flipping is employed to augment images during the training stage. The number of heads in CFIM is fixed to 8. The dimension of the word embedding is set to 512, and the hidden state dimensions of LSTM is set to 1000. During training stage, the number of the count words to make vocabulary is set to 4, and the max length of the generated sentence is set to 25. We use Adam optimizer to optimize the model. The learning rate is set to $10^{-4}$ for Sydney and UCM datasets, and $5 \times 10^{-4}$ for NWPU dataset. The batch size of the network is fixed to 64 and the maxepoch is set to 100. In the process of inference, we select the beam search value as 3, the random flipping and normalization are employed to augment images. All the experiments are implemented by PyTorch 1.10 in an NVIDIA RTX 3090.

### D. Comparison with Existing Methods

To illustrate the effectiveness of the HCNet, several existing methods are selected and compared with the proposed model on three datasets. Table I shows that the proposed method achieves competitive performance on Sydney dataset. Specifically, VLAD-LSTM utilizes hand-crafted features to generate caption of the image, which leads to the model with limited representation ability. Therefore, it cannot obtain ideal values. GoogLeNet combined by soft/hard attention generates high-quality captions, which benefits deep exploration. It results in great performance. Considering overfitting problem, TCE loss obtain better results on Sydney dataset. In contrast, our HCNet yields competitive performance, especially in BLEU-4, ROUGE_L, and CIDEr metrics. In our view, there are two main reasons for this performance. The one is that the introduction of the HFAM attains accurate and complete visual features. It facilitates the model in generating accurate descriptions. The other is that the proposed CFIM and loss $\mathcal{L}_{cma}$ allows the model to better align different modal features.

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 [41] | 0.7250 | 0.6030 | 0.5180 | 0.4550 | 0.3360 | 0.5910 | 1.1790 | 0.6402 |
| Soft Attention [31] | 0.7310 | 0.6090 | 0.5250 | 0.4620 | 0.3390 | 0.5990 | 1.1360 | 0.6340 |
| Hard Attention [31] | 0.7330 | 0.6100 | 0.5270 | 0.4640 | 0.3400 | 0.6000 | 1.1030 | 0.6267 |
| FC-Att+LSTM [48] | 0.7360 | 0.6150 | 0.5320 | 0.4690 | 0.3380 | 0.6000 | 1.2310 | 0.6595 |
| SM-Att+LSTM [48] | 0.7390 | 0.6170 | 0.5320 | 0.4680 | 0.3300 | 0.5930 | 1.2360 | 0.6567 |
| MLCANet [42] | 0.7540 | 0.6240 | 0.5410 | 0.4780 | 0.3370 | 0.6010 | 1.2640 | 0.6700 |
| BUTD$*$ [20] | 0.8718 | 0.7878 | 0.7206 | 0.6669 | 0.4304 | 0.7612 | 1.8866 | 0.9362 |
| AoANet$*$ [49] | 0.8755 | 0.7754 | 0.6917 | 0.6207 | 0.3907 | 0.7301 | 1.849 | 0.8976 |
| MLAT$*$ [13] | 0.8527 | 0.7674 | 0.7007 | 0.6479 | 0.4327 | 0.7496 | 1.8556 | 0.9214 |
| Clipcap$*$ [50] | 0.8394 | 0.7421 | 0.6619 | 0.5954 | 0.4142 | 0.7385 | 1.7265 | 0.8685 |
| PureT [26] | <u>0.8880</u> | <u>0.8031</u> | <u>0.7330</u> | <u>0.6750</u> | 0.4232 | 0.7584 | <u>1.9512</u> | <u>0.9519</u> |
| RS-CapRet [51] | 0.8710 | 0.7870 | 0.7170 | 0.6560 | <u>0.4360</u> | <u>0.7760</u> | 1.9290 | 0.9492 |
| HCNet | **0.8954** | **0.8251** | **0.7658** | **0.7168** | **0.4742** | **0.8053** | **2.0927** | **1.0222** |

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

| | Components | | | | | Metric | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Baseline | HFAM | R&Cat [13] | CFIM | $\mathcal{L}_{cma}$ | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4↑ | METEOR↑ | ROUGE_L↑ | Cider ↑ | Sm ↑ |
| Model1 | ✓ | | | | | 0.8392 | 0.7778 | 0.7273 | 0.6820 | 0.4486 | 0.7895 | 3.1416 | 1.2654 |
| Model2 | ✓ | ✓ | | | | 0.8658 | 0.8148 | 0.7690 | 0.7273 | <u>0.4855</u> | 0.8352 | 3.4343 | 1.3706 |
| Model3 | ✓ | | ✓ | | | 0.8597 | 0.8067 | 0.7584 | 0.7136 | 0.4608 | 0.8091 | 3.3437 | 1.3318 |
| Model4 | ✓ | | | ✓ | | 0.8429 | 0.7793 | 0.7263 | 0.6796 | 0.4433 | 0.8022 | 3.1529 | 1.2695 |
| Model5 | ✓ | ✓ | | ✓ | | <u>0.8805</u> | <u>0.8270</u> | <u>0.7796</u> | <u>0.7361</u> | 0.4779 | <u>0.8355</u> | <u>3.4953</u> | <u>1.3862</u> |
| HCNet | ✓ | ✓ | | ✓ | ✓ | **0.8826** | **0.8335** | **0.7885** | **0.7449** | **0.4865** | **0.8391** | **3.5183** | **1.3972** |

This can produce coherent captions in terms of language.

As seen in Table II, our method achieves the best performance for almost metrics on UCM dataset. Concretely, the BLEU-4 obtained by the proposed method reaches 74.49%, which is 7.49% higher than that of TCE loss. Meanwhile, our method yields 0.9% improvement on METEOR compared with TCE loss, which shows the generated captions that cover important information in the reference labels. Unlike the results on the above dataset, GVFGA+LSGA shows highest ROUGE_L and CIDEr values, which reveals that it has poor robustness. The Purt model achieves a competitive performance in the UCM dataset, which is due to the fact that the model is entirely based on the Transformer for extracting image features and generating descriptions, thus minimizing the cross-modal feature differences is vital for RSIC task. Meanwhile, the RS-CapRet model that based on the large language model technique achieves the best Cider, demonstrating the potential of the LLM for the RSIC task.

The above experimental results prove that our method exhibits significant advantages in comprehensive performance over existing methods, particularly in dealing with large-scale datasets. Similarly, our HCNet also demonstrates superior performance on NWPU dataset in Table III, especially in the case of the Cider metric, which outperformed the Purt model by 14.15%. The comparison experiment indicates the proposed method can address these scenes across datasets well.

### *E. Ablation Study*

In this section, we conduct ablation experiment on UCM dataset to demonstrate the influence of HFAM, CFIM, and cross-modal feature align loss function. Here, the combination of different parts are defined as Model.

**Effect of HFAM:** The proposed HFAM aggregates features at different scales in Model2 achieves significant improvements across all metrics compared with the Baseline (ResNet50+dual-LSTM) in Table IV. These results provide the evidences that the aggregation of multi-resolution features enables the model to acquire comprehensive feature representations. As a result, the model performs better in object classification and greatly improves the accuracy of generated captions. Furthermore, we also consider the feature fusion strategy [13], which adopts the multiple operations to fuse features with different resolutions. Compared with Model3, the Baseline combined with HFAM has a better performance. This improvement can be attributed to the detail and semantic information.

**Effect of CFIM:** Compared with Baseline, the Model4 with CFIM increases 1.27% in ROUGE_L. This improvement enhances the alignment effect of cross-modal features, which facilitates an effective interaction of visual and textual features in the decoder. CFIM enables the LSTM to generate accurate hidden states, which is conducive to guide accurate caption generation. Therefore, the generated sentences are closer to the ground-truth. In contrast with Model2 and Model3, Model5 demonstrates the improvements in most metrics when HFAM and CFIM are embedded in the model. These experiment results indicate that the integration of designed modules yields higher-quality captions.

**Effect of the $\mathcal{L}_{cma}$ Loss Function:** To minimize the difference between image and text features, cross-modal align loss

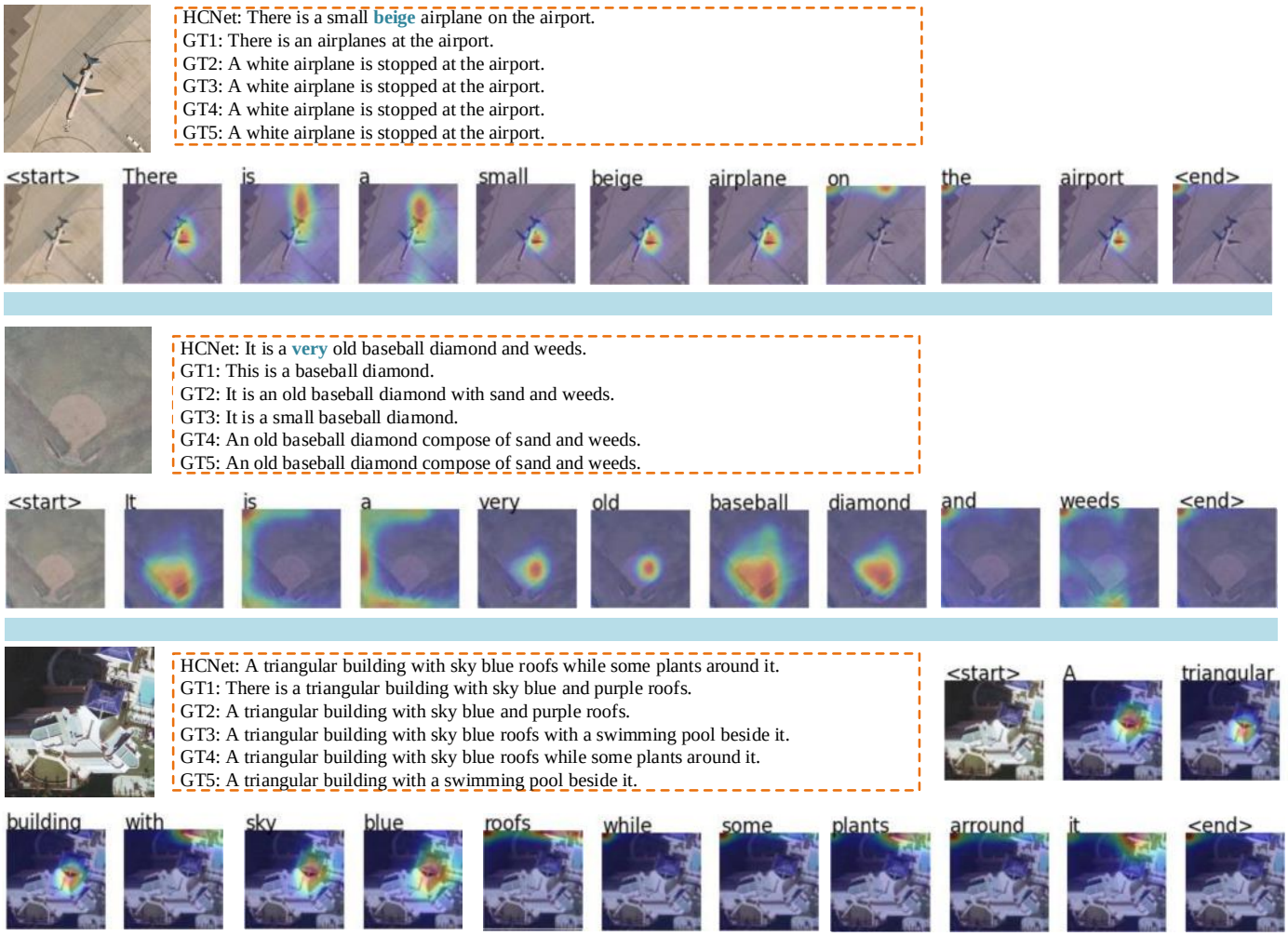



Fig. 5. Visual results of each word in the generated sentence. The closer the color is to red means that the model pays more attention to this area. The blue indicates the difference between our results and ground truth.

function is introduced in the Model5. Cross-modal features with the same semantics are further aligned under the guidance of this loss function to improve the performance of predicted captions. Therefore, our method obtains the highest values in all metrics, which reflects that the captions generated by HCNet are more closer to the references. This experiment demonstrates the importance of incorporating the cross-modal align loss function for RSIC task.

### *F. Visual Results*

To quantitatively show the generation process in captions, specific words, sentences, and corresponding attention maps are selected. Fig. 5 depicts the examples of sentences generated by HCNet. This figure lists three distinct scenes, including airplane, baseball diamond, and building. In the first sample, our approach generates the word "beige", which closely resembles the color "white". Throughout the description process, the model notably pay more attention to the location area of aircraft, when generating words related to aircraft, such as "a", "small", "airplane", and "airport". In the second sample, the model predicts the word "very", which enhances the coherence of the sentence. In the third sample, the model focuses on the corresponding location area in the image while generating the words "building" and "roofs". These samples presented in this figure demonstrate that our HCNet can provide precise descriptions of objects in the RSI.

To analyze the experimental results, several predicted results are selected to show. Fig. 6 displays five examples on Sydney, UCM, and NWPU datasets. These contents include experimental images, corresponding manual annotations (five in total), and image descriptions generated by Baseline and OurNet. Specifically As the images (A) and (B) show on Sydney dataset, the Baseline can identify "Many houses" in the image but fails to provide a comprehensive depiction of the global features. In contrast, our method effectively integrates multi-scale image features, which enhances the ability of feature representation. Consequently, it generates the "A residential area". As shown in the image (C), it is obvious that the description of "An empty of with" predicted by the Baseline contains a grammatical error. Different from this, the CFIM designed by our approach can effectively integrate visual and text features, which reduces the appearance of grammatical errors. As for the image (D), our model demonstrates the superior recognition ability, whereas the Baseline wrongly identifies "river" in the

## Sydney

**Base:** Many houses arranged neatly and some roads go through this area.
**HCNet:** A residential area with many houses arranged neatly and some roads go through this area.
**GT1:** A residential area with many houses arranged neatly and a wide highway beside .
**GT2:** There are many houses with red and orange roofs densely arranged while a wide highway beside .
**GT3:** A residential area with many houses arranged neatly and a wide road beside .
**GT4:** Lots of houses with red and orange roofs arranged densely while a wide highway beside .
**GT5:** Many houses with red and orange roofs arranged densely and a wide road beside .

(A)

**Base:** Many houses arranged neatly and divided by some roads.
**HCNet:** A residential area with many houses arranged neatly and divided by some roads.
**GT1:** A residential area with many houses arranged neatly .
**GT2:** Many houses arranged densely and divided into rectangles by some roads .
**GT3:** This is a residential area with many houses neatly arranged and divided into rectangles by some roads .
**GT4:** Lots of houses arranged neatly in the residential area and divided into rectangles by some roads .
**GT5:** There are many houses arranged neatly and divided into rectangles by some roads .

(B)

**Base:** An empty of with buildings in the industrial with some buildings beside.
**HCNet:** An industrial area with many white buildings densely arranged while a lawn beside.
**GT1:** An industrial area with many white buildings densely arranged while a lawn beside .
**GT2:** There is a lawn with a industrial area beside .
**GT3:** Many white houses arranged densely in the industrial area while a lawn beside .
**GT4:** There are many white buildings arranged densely in the industrial area with a lawn beside .
**GT5:** An industrial area with many white buildings densely arranged while a lawn beside .

(C)

**Base:** A straight runway with white mark lines on while a lawn beside.
**HCNet:** Two straight runways with white mark lines on while a river beside.
**GT1:** A wide runway with white mark lines on while a river beside .
**GT2:** A river with deep green water are beside a wide runway .
**GT3:** There is a wide runway with white mark lines on while a river beside .
**GT4:** There is a river beside a wide runway .
**GT5:** A wide runway with white mark lines and numbers on while a river beside .

(D)

**Base:** An industrial area with many white buildings and some roads go through.
**HCNet:** An industrial area with many white buildings densely arranged while a crossroad beside.
**GT1:** Many white buildings arranged densely in the industrial area while some houses beside .
**GT2:** This is an industrial area with many white buildings densely arranged while a residential area beside .
**GT3:** An industrial area with many white buildings densely arranged while a residential area beside .
**GT4:** An industrial area with a residential area beside .
**GT5:** Many white buildings in the industrial area while a residential area beside .

(E)

## UCM

**Base:** There are two straight freeways in the desert.
**HCNet:** It is a piece of farmland.
**GT1:** There is a piece of farmland .
**GT2:** There is a piece of cropland .
**GT3:** It is a piece of farmland .
**GT4:** It is a piece of cropland .
**GT5:** Here is a piece of farmland .

(F)

**Base:** Three airplanes are stopped dispersedly at the airport.
**HCNet:** There are two airplanes at the airport.
**GT1:** There are two airplanes at the airport .
**GT2:** Two different kinds of airplanes are stopped at the airport .
**GT3:** Two airplanes are stopped at the airport .
**GT4:** Two white airplanes are stopped at the airport .
**GT5:** Two different airplanes are stopped at the airport .

(G)

**Base:** A baseball diamond composed of sand and weeds.
**HCNet:** It is a very old baseball diamond compose of sand and 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 .

(H)

**Base:** This is a beach with blue sea and sand.
**HCNet:** This is a beach with blue sea and white sands.
**GT1:** This is a beach with blue sea and white sands .
**GT2:** Waves are slapping a white sand beach while some birds flying .
**GT3:** Waves are slapping a white sand beach over and over again and throw up white foam .
**GT4:** Waves beat on the beach over and over again .
**GT5:** Waves are slapping the white sand beach and throwing up white foams .

(I)

**Base:** There are some buildings with grey roofs.
**HCNet:** There are some buildings with white 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 .

(J)

## NWPU

**Base:** Two planes are in the parking lot .
**HCNet:** Two planes are in the parking lot next to the building .
**GT1:** A small plane and a big plane parked next to boarding bridges while some white facilities around them .
**GT2:** Here are two aircraft parked at the boarding port .
**GT3:** Two planes of different sizes are neatly parked next to the buildings in the airport .
**GT4:** Two planes are parked next to the terminal .
**GT5:** Two planes are in the parking lot next to the building .

(K)

**Base:** There are some green bushes beside the railway .
**HCNet:** There are some trains on the railway.
**GT1:** These railways have lots of trains on the tracks .
**GT2:** There are several rows of buildings on one side of the railway .
**GT3:** There are some trains on the railway .
**GT4:** There are many trains on the railway .
**GT5:** There are some green trees beside the railway with some trains .

(L)

**Base:** There is a basketball court in the middle of grass between a parking lot and many trees .
**HCNet:** There is a basketball court in the middle of grass next to a white building .
**GT1:** A basketball court between a purple building and a white building while a parking lot beside .
**GT2:** There is a basketball court directly below .
**GT3:** The parking lot is next to the lawn with a basketball court .
**GT4:** There is a basketball court in the middle of grass next to a parking lot .
**GT5:** A blue building is beside the basketball court .

(M)

**Base:** The beach with white sand and the color of the seawater is green .
**HCNet:** There are water marks on the beach
**GT1:** The beach with khaki sand and the color of the seawater is green .
**GT2:** There is a large area of vegetation next to the beach .
**GT3:** Green clear sea and light yellow beach .
**GT4:** The beach is made up of deep green seawater and a yellow coast .
**GT5:** There are water marks on the beach .

(N)

**Base:** The chaparral is in the bare land .
**HCNet:** The chaparral is scattered and sparse .
**GT1:** Some green chaparrals of different sizes grow in the desert and the dry river channel beside .
**GT2:** Only scattered weeds can be seen in this desert .
**GT3:** Chaparral is sparsely distributed in strips on the sand .
**GT4:** The chaparral is consist of sparse bushes in the bare land .
**GT5:** The chaparral is scattered and sparse .

(O)

Fig. 6. Examples of sentences generated by HCNet 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.

image as "lawn". Similarly, the proposed model obtains more comprehensive captions compared with the Baseline in the image (E). Furthermore, we also show the results on UCM and NWPU datasets. The experiments exhibit that our scheme can achieve higher accuracy and completeness of the captions.

### G. Hyperparameter Selection

In this section, we conduct hyperparameter selection experiment on UCM dataset to demonstrate the influence of $\lambda$ and the dimension of Word Embedding.

**Value of $\lambda$:** The weight of the loss function is an important hyperparameter in the model optimization phase. It is difficult to align the visual features obtained by the encoder with the same semantic information of the text features extracted by captions while the value is small. Conversely, a large value introduces optimization bias within the framework, impeding the generation of high-quality text descriptions. As shown in the Fig. 7, we set $\lambda$ from 1 to 4, and when the value is set to 3, the model achieves the best value in Sm metric, which proves that the model delivers the best comprehensive performance under this configuration.

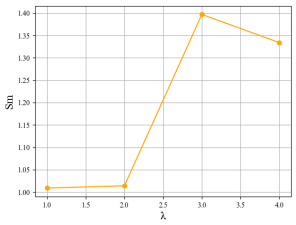



Fig. 7. Comparison results with the different value of $\lambda$ in the UCM dataset.

**Dimension of Word Embedding:** The dimension of word embedding stands as a crucial parameter in caption modeling. The model with small value of dimension fails to capture enough semantic information embedded in the text, and when the value is large, it decreases the representation ability of visual features. As shown in Fig. 8, when the value is set to 512, the model obtains the highest score.

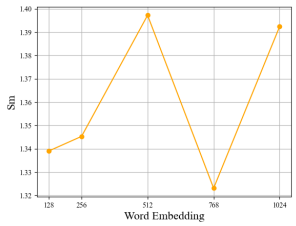



Fig. 8. Comparison results with with the different value of word embedding in the UCM dataset.

## V. Conclusion

This paper proposes a RSIC approach that leverages hierarchical feature aggregation and cross-modal feature alignment. To effectively utilize the features of different scales, the hierarchical feature aggregation module is designed to enhance feature representation. Additionally, the cross-modal feature interaction module is developed to align visual features with different input features in the decoder, which facilitates the model to generate accurate hidden states. Considering the disparities between two modal features, we introduce the cross-modal feature align loss to align vision features and text features, thus increasing the overall quality of generated captions. The experiments are conducted on three public datasets, and the results demonstrate the superiority of the proposed method in terms of quantitative and qualitative aspects. At present, the large model shows great advantages in computer vision. In the future, we will combine large model technology to improve the performance of the RSIC model.

## References

[1] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-Based Image Captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539-559, 2023.

[2] X. Tang, Y. Wang, J. Ma, X. Zhang, F. Liu and L. Jiao, "Interacting-Enhancing Feature Transformer for Cross-Modal Remote-Sensing Image and Text Retrieval," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1-15, 2023.

[3] L. Bashmal, Y. Bazi, F. Melgani, M. M. Al Rahhal and M. A. A. Zuair, " Language Integration in Remote Sensing: Tasks, datasets, and future directions," *IEEE Geosci. Remote Sens. Mag.*.

[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 Trans. Geosci. Remote Sensing*, vol. 60, pp. 1-20, 2022.

[5] Q. Li, M. Gong, Y. Yuan and Q. Wang, "RGB-Induced Feature Modulation Network for Hyperspectral Image Super-Resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1-11, 2023.

[6] 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.

[7] Z. Zheng, Y. Zhong, J. Wang, A. Ma and L. Zhang, "FarSeg++: Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13715-13729, 2023.

[8] F. Yang and C. Ma, "Sparse and Complete Latent Organization for Geospatial Semantic Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1809-1818.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1-9.

[10] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct, 2021, pp. 10012-10022.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017.

[12] W. Huang, Q. Wang and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning." *IEEE Geosci. Remote Sens. Lett.*, pp. 436-440, 2020.

[13] C. Liu, R. Zhao and Z. Shi, " Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1-5, 2022.

[14] X. Ma, R. Zhao and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, pp. 2001-2005, 2020.

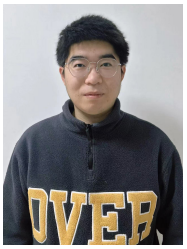
[15] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu and X. Sun, " Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1-16, 2021.

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

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

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

[19] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156-3164.
[20] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6077-6086.
[21] 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," *Int. Conf. Mach. Learn. (RMLR)*, 2015, pp. 2048-2057.
[22] L. Chen, H. Zhang, L. Nie, J. Shao, W. Liu and T. -S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5659–5667.
[23] L. Li, S. Tang S, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4133–4139.
[24] J. Lu, C. Xiong, D. Parikh and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 375-383.
[25] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 2286-2293.
[26] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 2585-2594.
[27] C. -W. Kuo and Z. Kira, "HAAV: Hierarchical Aggregation of Augmented Views for Image Captioning" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp.11039-11049.
[28] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, T. Mei, "Semantic-conditional diffusion networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23359–23368.
[29] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, "Smallcap: lightweight image captioning prompted with retrieval augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 2840–2849.
[30] F. Ma, Y. Zhou, F. Rao, Y. Zhang, and X. Sun, "Image captioning with multi-context synthetic data," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2024, pp. 4089–4097.
[31] X. Lu, B. Wang, X. Zheng and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Trans. Geosci. Remote Sensing*, vol. 56, no.4, pp, 2185-2195, 2018.
[32] Q. Wang, W. Huang, X. Zhang and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no.12, pp, 10532-10543, 2020.
[33] G. Hoxha and F. Melgani, " A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1-14, 2021.
[34] Z. Chen, J. Wang, A. Ma and Y. Zhong, " TypeFormer: Multiscale transformer with type controller for remote sensing image caption," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1-5, 2022.
[35] X. Li, X. Zhang, W. Huang and Q. Wang, " Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 6, pp. 5246-5257, 2020.
[36] X. Lu, B. Wang and X. Zheng, " Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 3, pp. 1985-2000, 2019.
[37] K. Zhao and W. Xiong, "Cooperative Connection Transformer for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp, 1-14, 2024.
[38] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, " RSGPT: A Remote Sensing Vision Language Model and Benchmark," 2023, *arXiv:2307.15266*.
[39] X. Ding, X. Zhang, J. Han and G. Ding, " Scaling Up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11953-11965, 2022.
[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, " Learning transferable visual models from natural language supervision," *International Conference on Machine Learning*, 2021.
[41] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1-5.
[42] 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 Trans. Geosci. Remote Sensing*, vol. 60, pp. 1-19, 2022.
[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, " Bleu: a method for automatic evaluation of machine translation," in *in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311-318.
[44] M. Denkowski and A. Lavie, " Meteor universal: Language specific translation evaluation for any target language" in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376-380.
[45] C.-Y. Lin, " Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004, pp. 74-81.
[46] R. Vedantam, C. Lawrence Zitnick, and D. Parikh," Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566-4575.
[47] G. Hoxha, G. Scuccato and F. Melgani, " Improving Image Captioning Systems With Postprocessing Strategies," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1-13, 2023.
[48] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li," Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, 2019.
[49] L. Huang, W. Wang, J. Chen and X. -Y. Wei," Attention on Attention for Image Captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 4633-4642.
[50] J. D. Silva, J. Magalh aes, D. Tuia, and B. Martins, " Large language models for captioning and retrieving remote sensing images," 2024, *arXiv:2402.06475*.
[51] R. Mokady, A. Hertz, and A. H. Bermano, " Clipcap: Clip prefix for image captioning," 2021, *arXiv:2111.09734*.



**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.

**Qiang Li** (Member, IEEE) is currently 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.

**Yuan Yuan** (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

**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.