CADFormer: Fine-Grained Cross-Modal Alignment and Decoding Transformer for Referring Remote Sensing Image Segmentation

Maofu Liu, Xin Jiang, and Xiaokang Zhang, *Senior Member, IEEE*

***Abstract*—Referring remote sensing image segmentation (RRSIS) is a challenging task, aiming to segment specific target objects in remote sensing images based on a given language expression. Existing RRSIS methods typically employ coarse-grained unidirectional alignment approaches to obtain multimodal features, and they often overlook the critical role of language features as contextual information during the decoding process. Consequently, these methods exhibit weak object-level correspondence between visual and language features, leading to incomplete or erroneous predicted masks, especially when handling complex expressions and intricate remote sensing image scenes. To address these challenges, we propose a fine-grained cross-modal alignment and decoding Transformer, CADFormer, for RRSIS. Specifically, we design a semantic mutual guidance alignment module (SMGAM) to achieve both vision-to-language and language-to-vision alignment, enabling comprehensive integration of visual and textual features for fine-grained cross-modal alignment. Furthermore, a textual-enhanced cross-modal decoder (TCMD) is introduced to incorporate language features during decoding, using refined textual information as context to enhance the relationship between cross-modal features. To thoroughly evaluate the performance of CADFormer, especially for inconspicuous targets in complex scenes, we constructed a new RRSIS dataset, called RRSIS-HR, which includes larger high-resolution remote sensing image patches and semantically richer language expressions. Extensive experiments on the RRSIS-HR dataset and the popular RRSIS-D dataset demonstrate the effectiveness and superiority of CADFormer.**

***Index Terms*—Cross-modal alignment, referring image segmentation (RIS), remote sensing.**

I. Introduction

In recent years, with the rapid development of Earth observation techniques, the combination of remote sensing and deep learning has become a popular research topic [1], [2]. Deep learning has made significant progress in various remote sensing tasks, including image captioning [3], visual question answering [4], semantic segmentation [5], [6], [7], image recognition [8], [9], [10], visual grounding [11], etc. Despite these advancements, the task of referring remote sensing image segmentation (RRSIS) remains an area of exploration. It combines computer vision and natural language processing, aiming to segment the target objects described by a natural language expression in remote sensing images. Compared to traditional remote sensing semantic segmentation and remote sensing instance segmentation [12], RRSIS is more flexible, allowing users to extract specific target objects of interest from images based on their needs. It holds great potential in various fields such as land use classification [13], disaster response [14], military intelligence generation [15], environmental monitoring [16], and agricultural production [17].

Since RRSIS is an emerging task in the field of RS, there has been relatively little exploration in this area. Yuan et al. [18] first introduced the concept of RRSIS and proposed the first RRSIS dataset, RefSegRS, along with a language-guided cross-scale enhancement module (LGCE), which aims to improve segmentation performance on small and sparsely distributed objects. Furthermore, to address the issue of complex spatial scales and orientations in remote sensing image scenes, Liu et al. [19] proposed the rotated multiscale interaction network (RMSIN) and constructed a large-scale RRSIS dataset, called RRSIS-D, based on the remote sensing visual grounding dataset RSVG [20]. The proposed RRSIS-D dataset is a new large-scale benchmark for RRSIS tasks and fully advances the research of RRSIS. However, we observe that the referring target objects in the remote sensing images of the RRSIS-D dataset are quite salient, and the referring text descriptions are relatively simple and brief, as shown in Fig. 2(b). This may decrease the challenge of the RRSIS task, allowing some RIS methods, which perform well on natural images, to still yield good results on the RRSIS-D dataset, even outperforming some RRSIS methods. This drives us to consider whether RRSIS models can still effectively segment inconspicuous targets from very high-resolution remote sensing images with complex language expressions. Therefore, we built a new RRSIS dataset based on the RSVG-HR dataset [21], named RRSIS-HR. The RRSIS-HR dataset consists of seven object categories and contains 2650 image-text-label triplets. Compared to the RRSIS-D dataset, the remote sensing images in RRSIS-HR have higher resolution and cover larger regions with complex backgrounds and less

Received 30 March 2025; revised 18 May 2025; accepted 31 May 2025. Date of publication 4 June 2025; date of current version 20 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42371374 and in part by the "14th Five Year Plan" Advantageous and Characteristic Discipline Project of Hubei Province (China) under Grant 2023D0302. *(Corresponding author: Xiaokang Zhang.)*

Maofu Liu and Xin Jiang are with the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China (e-mail: liumaofu@wust.edu.cn; jx@wust.edu.cn).

Xiaokang Zhang is with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: natezhangxk@gmail.com).

Datasets and source codes will be available at https://github.com/zxk688/CADFormer.

Digital Object Identifier 10.1109/JSTARS.2025.3576595

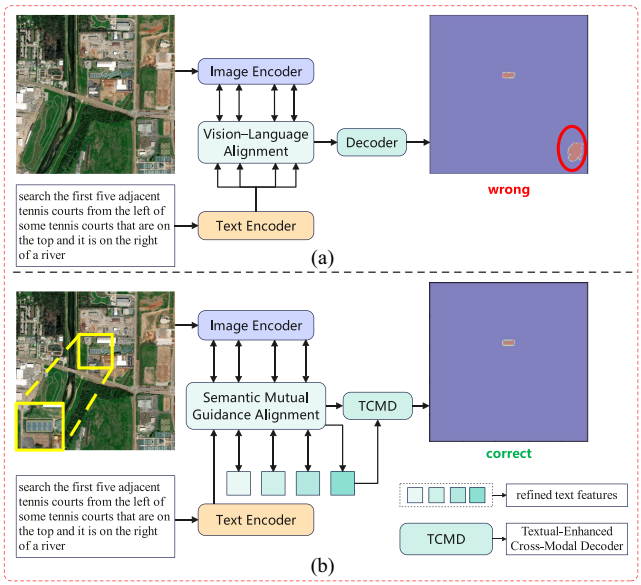

Fig. 1. Motivation of the proposed approach. The input remote sensing image-text pair is from our proposed RRSIS-HR dataset. (a) Existing RRSIS methods use coarse-grained unidirectional alignment from vision to language and a simple standard decoder. (b) Our proposed CADFormer uses semantic mutual guidance alignment and a TCMD.

prominent objects. Furthermore, the language expressions in RRSIS-HR are longer, more complex and semantically richer, frequently describing multiple object categories, detailed spatial relationships, and contextual information, as shown in Fig. 2(b). The increased linguistic and visual complexity poses substantial challenges for cross-modal learning and computation [22], [23].

Previous RRSIS methods [18], [19] follow the basic architectural strategy in Fig. 1(a), which adopts simple pixel-word attention [24] to align language and visual features, directly integrating original language features derived from BERT [25] into multiscale visual features throughout the alignment process. In this process, visual features are progressively refined under the guidance of original language features, while language features remain fixed throughout. This represents a coarse-grained alignment strategy from vision to language that neither utilizes visual context to refine language representations nor incorporates linguistic feedback to enhance visual comprehension, ultimately failing to establish a dynamic cross-modal feedback loop. Ideally, with pixel-word vision-language alignment, language and visual features should exhibit high feature similarity when referring to the same object [26]. However, achieving this alignment is not straightforward because language expressions can be highly complex and diverse. When confronted with high-resolution remote sensing image scenes and longer, semantically richer language expressions, these methods exhibit weak object-level correspondence and often struggle to produce optimal segmentation results. In addition, these methods only consider the alignment and interaction between language and visual features during the multimodal feature fusion process. During the decoding phase, cross-modal features are gradually decoded using a simple segmentation head [24], [27] to produce the final prediction. However, the importance of text guidance during the decoding process is often overlooked, which may lead to the loss of crucial fine-grained details and suboptimal segmentation results due to the lack of explicit semantic constraints in refining region boundaries and resolving ambiguities.

To address the challenges mentioned above, we propose CADFormer, a novel RRSIS method from the perspective of fine-grained cross-modal alignment and decoding, as shown in Fig. 1(b). Specifically, we introduce a semantic mutual guidance alignment module (SMGAM) to enhance semantic relevance between cross-modal features. This module performs both language-guided vision alignment and vision-guided language alignment, fully integrating visual and language features from the perspective of mutual guidance alignment. In this process, the two modalities guide each other and dynamically adjust. Visual features are progressively refined guided by language, while language features are simultaneously adjusted guided by vision, dynamically adapting to different visual content. This approach differs from previous methods [18], [19], [27] that relied solely on semantic features derived from BERT [25] throughout the alignment process. Through this mechanism of mutual guidance and collaborative adaptation between modalities, SMGAM achieves strong object-level correspondence between visual and language features for fine-grained cross-modal alignment. In addition, we design a textual-enhanced cross-modal decoder (TCMD), which accepts the refined multiscale visual features and language features as input and leverages a Transformer decoder for further processing. The refined language features are used as contextual information to guide the model in processing the refined multiscale visual features, thus enhancing the interaction between cross-modal features and ensuring more accurate predicted segmentation masks. In summary, the main contributions of this article are as follows.

1) We propose a novel RRSIS method named CADFormer for handling complex remote sensing image scenes with semantically richer language expressions. Specifically, the SMGAM enhances the semantic relevance between visual and language features by modeling their mutual dependencies and achieves fine-grained cross-modal alignment. In addition, the TCMD leverages refined language features as contextual guidance for decoding and segmentation, resulting in more accurate predictions.
2) We construct a new RRSIS benchmark, RRSIS-HR, which contains high-resolution remote sensing images with fine-grained language expressions, posing challenges for RRSIS methods in handling complex scenes.
3) We conduct extensive experiments on the RRSIS-HR and RRSIS-D datasets. The experimental results show that our proposed method, CADFormer, outperforms the majority of existing RRSIS methods, demonstrating the effectiveness of CADFormer and its superiority in handling high-resolution remote sensing image scenes with fine-grained language expressions.

II. Related Work

A. Referring Image Segmentation

Referring image segmentation (RIS) aims to segment specific target objects in images based on natural language expressions, making it a typical multimodal task that has gained increasing

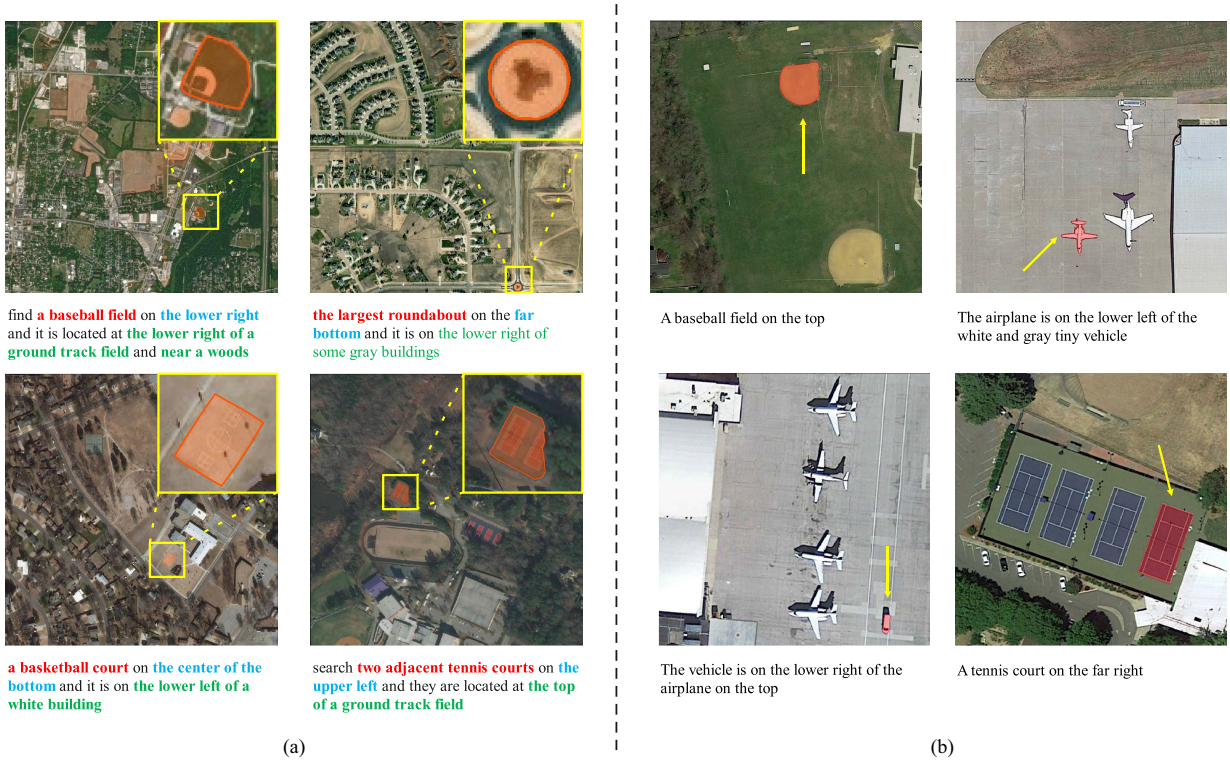

Fig. 2. Typical examples of our proposed RRSIS-HR datasets and public RRSIS-D datasets. (a) RRSIS-HR dataset. The red, blue, and green fonts in the language expressions represent categories, absolute positions, and relative position relationships, respectively. (b) RRSIS-D dataset.

attention. Early RIS research [28] focused on using convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to extract visual and language features, respectively, and then simply fused these features through concatenation to generate the final predictions. Subsequent work improved this process by using recurrent neural networks (RNNs) [29], [30], or dynamic networks [31] to progressively refine segmentation masks. The emergence of Transformer [32] architectures and attention mechanisms revolutionized RIS methods by providing significant fusion capabilities for multimodal integration. For example, CMSA [33], BRINet [34], and VLT [35] leverage crossmodal self/interactive attention to deeply fuse language and visual features, capture longrange dependencies, and produce query vectors that precisely retrieve target information from images. To promote cross-modal integration, LAVT [24] introduced a language-aware attention mechanism into the image encoding process, assisting early fusion of cross-modal features and improving segmentation accuracy. Recently, VPD [36] explored semantic information in diffusion models for RIS, while RefSegformer [37], ReLA [38], and DMMI [39] focused on improving model robustness with the proposal of the generalized RIS task. In the generalized RIS task, a referring expression can refer to an arbitrary number of target objects, including multiple targets or even no target at all. However, unlike natural images where isolated and prominent subjects dominate, the expansive coverage of remote sensing images inevitably captures densely clustered small-scale targets with multiscale spatial distributions, which are frequently obscured by cluttered backgrounds. These inherent characteristics not only amplify the technical challenges for precise localization, but also limit the generalization capability of existing RIS methods in achieving satisfactory performance.

B. *Referring Remote Sensing Image Segmentation*

In recent years, RIS tasks in the domain of remote sensing have attracted significant attention from researchers. Research in this field is still in its early stages and remains relatively scarce. Yuan et al. [18] first introduced the RIS task into the remote sensing domain by constructing the first RRSIS dataset, RefSegRS, and proposing the LGCE module to adaptively fuse deep and shallow visual features, thereby improving the segmentation of small and scattered objects. Liu et al. [19] proposed the RMSIN, based on the LAVT [24] framework, to address scale and rotation variations in remote sensing images by jointly modeling intrascale and cross-scale image-text interactions. They also introduced a large-scale RRSIS-D dataset, further advancing RRSIS research. To explicitly address the domain gap between vision and language, DANet [40] introduced an explicit alignment strategy to narrow inter-domain affinity distributions, along with a reliable proxy alignment module to enhance multimodal perception and suppress noisy interference. Recently, FIANet [27] decoupled referring expressions into object-specific and spatial location texts and integrated them with visual features through a fine-grained image-text alignment module to obtain more discriminative multimodal representations. However, these methods uniformly employ the original semantic features extracted by BERT [25] to interact with multiscale visual features throughout

the alignment and fusion process, and some approaches [18], [19] omit textual features during the decoding phase. In contrast, our CADFormer method incorporates cross-modal alignment across four stages. Throughout this process, we progressively obtain refined language features and refined visual features based on semantic mutual guidance, which serve as input for subsequent stages. We fully leverage language features, thereby establishing a robust object-level correspondence between visual and language features.

III. New Benchmark

A. RRSIS-HR Dataset Construction

We introduce a new RRSIS dataset, named RRSIS-HR. Inspired by SAM [41] and RMSIN [19], we adopt a semi-automatic method, using bounding boxes and SAM to generate pixel-level segmentation masks, significantly reducing the cost of manual annotation. Specifically, we follow the steps below to construct the RRSIS-HR dataset.

1) *Step 1:* We collect remote sensing images, referring text descriptions, and corresponding visual grounding bounding boxes from the RSVG-HR [21] dataset. By leveraging the bounding boxes provided by the RSVG-HR dataset and the bounding box hinting function of SAM, we obtain preliminary pixel-level masks for all referring target objects in the dataset. However, due to the significant domain gap between natural images and remote sensing images, SAM may generate unsatisfactory segmentation masks when applied to partial images, requiring refinement and optimization.
2) *Step 2:* To obtain more accurate fine-grained pixel-level segmentation masks, we optimize the masks generated by SAM in Step 1 through manual verification. First, three annotation experts in the field of remote sensing, drawing on their expertise, developed a set of annotation standards. Following the standards, they carefully examined the masks generated by SAM, identifying and filtering out problematic masks. Subsequently, each annotator used the image segmentation semi-automatic annotation tool to optimize the problematic masks. This process includes refining boundaries, adjusting sizes, correcting errors, and resolving occlusion issues. When initial corrections are uncertain, a consensus process is triggered. This process involves independent review and cross-checking by three annotators and discussion among multiple experts to make a final decision. Through this human–computer collaborative semi-automatic annotation method, we achieve high-precision mask annotations while significantly reducing the cost of manual annotation.
3) *Step 3:* Finally, to improve the compatibility of RRSIS-HR with various RIS models, we convert the annotation format to the RefCOCO dataset [42] format for later use.

B. Data Analysis

Through the semi-automatic annotation method, we have successfully constructed the RRSIS-HR dataset, which consists of 2650 image-text-mask triplets and includes seven object categories. Each remote sensing image has a size of 1024×1024 pixels, containing varying scales and details, and covers an area ranging from 0.06 km^2 to 25 km^2 [43]. The average length of the language descriptions is 19.6 words, with a minimum of 6 words and a maximum of 41 words. Specifically, a language expression contains one or more object categories, which requires RRSIS models to accurately identify target objects from more object categories. Fig. 2 shows some visual examples of RRSIS-HR and RRSIS-D datasets. Compared to the RRSIS-D dataset, although the RRSIS-HR dataset is not large in scale, it contains higher-resolution remote sensing images, with longer and more complex language expressions, making it a challenging dataset that includes complex remote sensing scenes for RRSIS methods.

IV. Methodology

A. Overview

The framework of our proposed CADFormer is illustrated in Fig. 3. It mainly consists of three parts: the image and text encoders, the SMGAM, and the TCMD. Given an image $I \in R^{H\times W\times 3}$ and a language expression $E \in \{e_i\}, i = \{0, 1 \dots N\}$, where H and W denote the height and width of the input image, respectively, and N represents the length of the language expression. We utilize the Swin Transformer [44] as the visual backbone network to extract multiscale visual features from the input image. The visual features at each stage are denoted as $V_i \in R^{H_i\times W_i\times C_i}, i \in \{1, 2, 3, 4\}$, where H_i, W_i, and C_i represent the number of height, width, and channel of the feature map from the ith stage, respectively. In addition, we use BERT [25] as the text encoder to extract language features represented as $L_1 \in R^{N\times C_t}$, where N and C_t denote the length of the sentence and the channel number, respectively. The language features from different stages and the multiscale visual features are progressively fed into the SMGAM. This module aligns cross-modal semantics and generates enhanced visual and linguistic representations. The enhanced visual and linguistic representations are then fed into the TCMD to predict pixel-level segmentation masks. We will introduce each module in detail as follows.

B. Semantic Mutual Guidance Alignment Module

In prior works [18], [19], [24], cross-modal alignment was limited to a coarse-grained unidirectional alignment from vision to language. The language features fused at each stage were directly derived from the initial sentence features extracted by BERT [25]. We consider this a coarse-grained alignment strategy that does not fully exploit the potential of language features. When the language expressions increase in length and complexity, existing approaches often encounter challenges in accurately distinguishing and localizing target objects across multiple categories. To address the challenge, we propose the SMGAM, which comprises two submodules, namely the language-guided vision-language alignment (LGVLA) submodule and the vision-guided language-vision alignment (VGLVA)

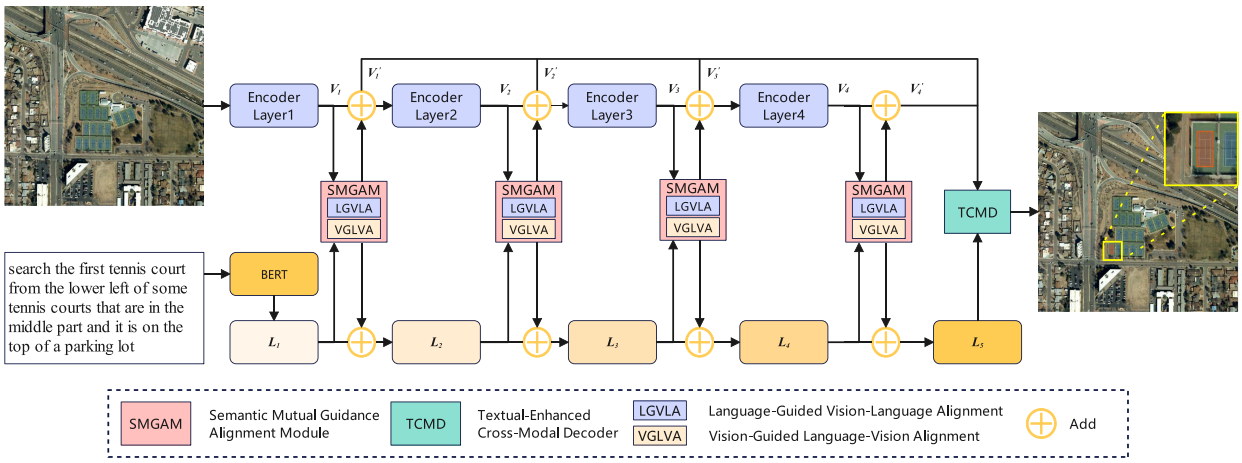

Fig. 3. Overview of our proposed CADFormer framework. The model first aligns multiscale visual features and text features progressively through the SMGAM. Then, the refined text features are used as contextual information to query the refined multiscale visual features in the TCMD, retrieving and aggregating target object information to generate the prediction results.

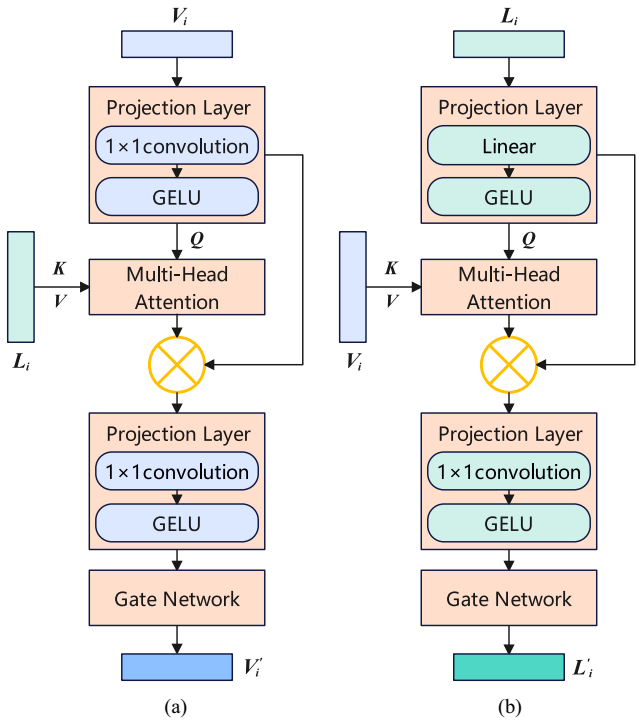

Fig. 4. Illustration of the proposed SMGAM. (a) LGVLA Submodule. (b) Vision-Guided Language-Vision Alignment Submodule.

submodule, as shown in Fig. 4. Both submodules take multiscale visual features V_i and stage-specific language features L_i as input. The LGVLA submodule produces refined visual features V_i', while the VGLVA submodule produces refined language features L_i'. Through a four-stage semantic mutual guidance alignment process, progressively refined visual and language features are obtained, achieving fine-grained cross-modal alignment. We will describe these two submodules in detail as follows.

1) Language-Guided Vision-Language Alignment: During the stage i, the LGVLA submodule takes the language features L_i and visual features V_i as input, as shown in Fig. 4(a). To better adapt to subsequent alignment tasks, visual features V_i are first passed through a projection layer, where they are mapped into a new feature space. The projection layer consists of a 1×1 convolutional layer followed by a GELU activation function, denoted as $\mathrm{Proj}(\cdot)$. The process can be formulated as: $\hat{V}_i \longleftarrow \mathrm{Proj}(V_i)$. Next, the language features interact with the visual features through a multihead attention [32] layer. For cross-modal interaction at each stage, the multihead attention layer performs cross-attention to obtain enhanced visual representations, where the visual features act as the query, and the language features serve as the key and value. Although the key and value are both derived from the language features, they are projected into different feature spaces through separate learnable transformations. Specifically, the attention layer initially calculates the similarity between the visual features and the language features through a scaled dot-product operation, which aligns each visual element with each language element. This computation results in a cross-modal similarity matrix M_{vl}^i that quantifies the relevance and interaction strength between the two modalities. The formulation is as follows:

$$M_{vl}^i = \mathrm{Softmax}\left(\frac{\hat{V}_i W_q^i \cdot \left(L_i W_k^i\right)^T}{\sqrt{C_i}}\right) \tag{1}$$

where W_q^i and W_k^i are the linear projections. Each of them is implemented as a 1×1 convolution with C_i output channels. Subsequently, we utilize the similarity matrix M_{vl}^i to integrate object-relevant details from the visual features into the language features and then multiply the result with the projected visual features, yielding language-guided visual features A_{vl}^i. The resulting features A_{vl}^i are further processed through another projection layer, followed by a language gate [24] to produce the refined visual features. The process is specifically described as follows:

$$A_{vl}^i = M_{vl}^i L_i W_v^i \otimes \hat{V}_i \tag{2}$$

$$V_i' = \mathrm{Gate}\left(\mathrm{Proj}\left(A_{vl}^i\right)\right) \tag{3}$$

$$\mathrm{Gate}(x) = \mathrm{MLP}(x) \otimes x \tag{4}$$

where W_v^i is the linear projection implemented as a 1×1 convolution, $\otimes$ denotes element-wise multiplication and $\mathrm{Proj}(\cdot)$ represents the projection layer, which consists of a 1×1 convolutional layer followed by a GELU activation function. $\mathrm{Gate}(\cdot)$ refers to a language gate and x is its input variable. MLP is a two-layer perceptron. The first layer is a linear layer followed by a ReLU activation function, while the second layer is a linear layer followed by a Tanh activation function. Subsequently, the refined visual features V_i' generated by the LGVLA submodule are merged with the original input features V_i. After passing through the next stage of the Swin Transformer layer, they are transformed into visual features for the next stage. The process can be described as follows:

$$V_{i+1} = \mathrm{SwinStage}_{i+1}\left(V_i' + V_i\right) \tag{5}$$

where $\mathrm{SwinStage}_i$ denotes the ith stage of the Swin Transformer [44], which primarily consists of downsampling operations and MLP layers.

2) Vision-Guided Language-Vision Alignment: Similar to the LGVLA submodule, the VGLVA submodule also takes language features and visual features from different stages as input and progressively performs cross-modal interactions between the two modalities. However, unlike the LGVLA submodule, the VGLVA submodule performs a different process and focuses on enhancing the language features through a series of iterative refinements guided by the visual context, as shown in Fig. 4(b). Specifically, the module employs the multihead attention [32] mechanism to iteratively align and refine the language features based on visual information. The output refined language features iteratively interact with the visual features of subsequent stages and this process continues until the refined language features have fully engaged with the multiscale visual features across all stages. During the ith stage, the language features first pass through a projection layer consisting of a linear layer followed by a GELU activation function, denoted as $\mathrm{Proj}(\cdot)$. The process is described as follows: $\hat{L}_i \longleftarrow \mathrm{Proj}(L_i)$. Subsequently, the multihead attention layer performs cross-attention to complete the cross-modal interaction between language and visual features, where the language features serve as the query, while the visual features act as the key and value, which are projected into different feature spaces through separate projection transforms. Specifically, the attention layer initially calculates the similarity between each language element and each visual element through a scaled dot-product operation to obtain the cross-modal similarity matrix M_{lv}^i as follows:

$$M_{lv}^i = \mathrm{Softmax}\left(\frac{\hat{L}_i W_q^i \cdot \left(V_i W_k^i\right)^T}{\sqrt{C_i}}\right) \tag{6}$$

where W_q^i and W_k^i are the linear projection matrices. Each of them is implemented as a 1×1 convolution with C_i output channels. Then, we utilize this similarity matrix M_{lv}^i to integrate object-relevant information from the language features into the visual features and multiply the result by the projected language features to obtain vision-guided language features. The resulting output is then passed through a projection layer, followed by a gate network similar to the language gate [24], to produce the refined language features. This gate network is still denoted as $\mathrm{Gate}(\cdot)$. The process is described as follows:

$$A_{lv}^i = M_{lv}^i V_i W_v^i \otimes \hat{L}_i \tag{7}$$

$$L_i' = \mathrm{Gate}\left(\mathrm{Proj}\left(A_{lv}^i\right)\right) \tag{8}$$

where W_v^i is the linear projection implemented as a 1×1 convolution, $\otimes$ denotes element-wise multiplication. $\mathrm{Gate}(\cdot)$ is a gate network with the same computation as equation (4), consisting of MLP [44] followed by multiplication. After that, the refined language features generated by the VGLVA submodule are merged with the input features L_i and transformed into the language features for the next stage, serving as the input for the subsequent SMGAM stage. The process is described as follows:

$$L_{i+1} = L_i + L_i'. \tag{9}$$

The refined language features obtained at the final stage are denoted as L_5.

C. Textual-Enhanced Cross-Modal Decoder

Previous works [18], [19], [24], [27] used only refined multiscale visual features as input during mask prediction, without fully leveraging the language features. These methods typically perform cross-modal interaction only before decoding, limiting their ability to further utilize language information during the decoding process. As a result, fine-grained details crucial for accurate segmentation may be lost. Furthermore, the lack of deep interaction between the semantic details in the language features and the visual features hinders the ability of the model to effectively distinguish between different categories or precisely segment specific target objects within the same category.

In contrast, our TCMD utilizes refined multiscale visual features and refined language features L_5 as input for mask prediction. The introduction of language features enables the model to focus on important information about the target objects at each stage of the decoding process, thereby enhancing its ability to capture fine-grained details. Specifically, the refined language features serve as contextual information to gradually guide the feature decoding process, retrieving and aggregating target object information from the refined visual features. This process incorporates a multilayer interaction mechanism of the Transformer decoder [32], where visual features are first reshaped as a sequence and aligned in dimensions with language features. Subsequently, the two modalities are fused through the Transformer decoder, where the self-attention mechanism dynamically incorporates contextual language information to enhance the semantic representation of visual features. Overall, through progressively integrating visual and language features across decoding layers, TCMD enables the model to capture the fine-grained relationships between modalities more accurately, thereby enhancing its understanding of complex scenes and improving the quality of mask prediction. Specifically, our decoding process follows a top–down approach, integrating refined multiscale visual features with refined language features, as shown in Fig. 5. In Fig. 5, the light blue rectangles of varying sizes represent refined multiscale visual features V_i' with different dimensions and we denote them as

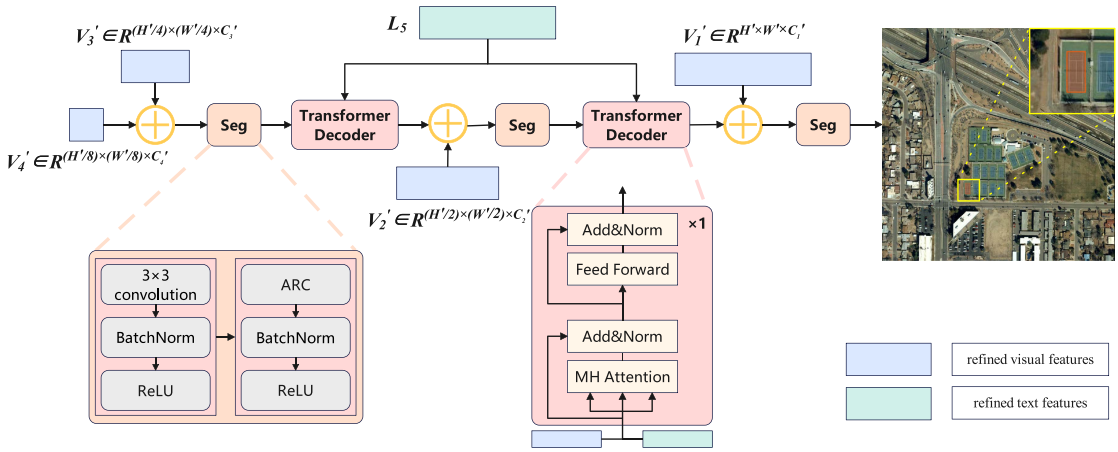

Fig. 5. Illustration of the proposed TCMD. MH Attention denotes the multihead attention layer. ARC denotes the adaptive rotated convolution. The light blue rectangles of varying sizes represent refined multiscale visual features with different dimensions.

$V_i' \in R^{(H'/2^{i-1})\times(W'/2^{i-1})\times C_i'}, i \in \{1,2,3,4\}$, where $H'/2^{i-1}$, $W'/2^{i-1}$, and C_i' represent the number of height, width and channels of the feature map. The overall decoding process can be described as follows:

$$Y_4 = V_4' \tag{10}$$

$$Y_i = \text{Transformer}\left(\text{Seg}\left(\left[Y_{i+1}; V_i'\right]\right), L_5\right), i = \{2,3\} \tag{11}$$

$$Y_1 = \text{Seg}\left(\left[Y_2; V_1'\right]\right) \tag{12}$$

where $[;]$ represents the concatenation operation along the channel dimension. Before concatenation, bilinear interpolation is applied to ensure spatial consistency between the feature maps. $\text{Seg}(\cdot)$ consists of two 3×3 convolutional layers, batch normalization, and ReLU activation functions to enhance the nonlinearity of the segmentation feature space. In addition, one of the 3×3 convolutional layers is replaced by an adaptive rotated convolution layer [19] to leverage directional information in the feature space, thereby eliminating redundancy and improving the accuracy of boundary details. $\text{Transformer}(\cdot)$ represents the Transformer decoder layer. The final feature map is projected into two class score maps using a 1×1 convolution. Finally, bilinear interpolation is employed to upsample the results to match the resolution of the input image.

D. Loss Function

In remote sensing images, the scarcity of target pixels compared to the abundance of background pixels creates a notable class imbalance. This imbalance can lead traditional cross-entropy loss functions to bias the model towards learning background features, ultimately reducing the effectiveness of target region detection. To address this issue, we adopt a combined loss function consisting of cross-entropy loss and Dice loss as our training objective

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{cross-entropy}}\left(Y, \hat{Y}\right) + (1-\lambda) \cdot \mathcal{L}_{\text{dice}}\left(Y, \hat{Y}\right) \tag{13}$$

where λ is the hyperparameter that balances the two loss functions, set to 0.9. $\hat{Y}$ represents the predicted results, and Y denotes the ground truth.

V. Experiments

A. Implementation Details

1) Experiment Settings: We use PyTorch to implement our method. Similar to previous methods [18], [19], during the experiments, we use the Swin Transformer [44] as the visual backbone, pretrained on ImageNet22 K, and the base BERT model from HuggingFace's library [45] as the text encoder. We employ the AdamW optimizer with a weight decay of 0.01 and an initial learning rate of 0.00005, with the learning rate decaying according to a polynomial schedule. The batch size is set to 2, and each model is trained for 40 epochs on an NVIDIA GeForce RTX 3090 GPU. During both training and testing phases, all images are resized to 480×480 pixels.

2) Evaluation Metrics: Following the prior research [18], [19], [24], we use mean intersection over union (mIoU), overall intersection over union (oIoU), and Precision@X (Pr@X) as evaluation metrics. Precision@X refers to the percentage of test samples for which the IoU between the predicted result and the ground truth exceeds a threshold X. It is used to evaluate the accuracy at a specific IoU threshold and reflects the method's performance in object localization. The mIoU and oIoU can be formulated as follows:

$$\text{mIoU} = \frac{1}{M}\sum_t I_t / U_t \tag{14}$$

$$\text{oIoU} = \sum_t I_t / \sum_t U_t \tag{15}$$

where t is the index of the image-language-label triplets and M represents the size of the dataset. I_t and U_t are the intersection and union areas of predicted and ground-truth regions.

3) Compared Methods: To evaluate the effectiveness of our proposed CADFormer, we compared it with several state-of-the-art methods of RIS for both natural images and remote sensing images on the test sets of RRSIS-D and RRSIS-HR. The results of the different methods are shown in Tables II and III, respectively. For a fair comparison, we reimplemented some of the state-of-the-art methods, including LAVT [24], LGCE [18],

TABLE I
DETAIL OF THE RRSIS-HR DATASET

Category	train	val	test	Total
Baseball field	420	53	52	525
Basketball field	190	24	24	238
Ground track field	170	22	21	213
Roundabout	524	66	66	656
Swimming pool	152	19	19	190
Storage tank	170	22	21	213
Tennis court	492	62	61	615
Total	2118	268	264	2650

RMSIN [19], and FIANet [27], with a total of 40 training epochs for both RRSIS-D and RRSIS-HR. For some earlier released methods, we used the results reported in RMSIN [19].

B. Dataset

We conducted experiments on two datasets, including the publicly available RRSIS-D dataset and the RRSIS-HR dataset constructed by us. Detailed information about these datasets is as follows.

1) RRSIS-D: The RRSIS-D dataset is built on the DIOR-RSVG [20] dataset and contains 20 object categories. The dataset contains a total of 17 402 image-language-label triplets, with 12 181 for training, 1740 for validation, and the remaining 3481 for testing, which is a large benchmark. The image size in this dataset is 800×800 pixels with spatial resolutions ranging from 0.5 to 30 m. The average length of the language expressions is 6.8 words. Some visual examples of the RRSIS-D dataset are shown in Fig. 2(b).

2) RRSIS-HR: The RRSIS-HR dataset contains very high-resolution remote sensing images and longer, semantically richer language expressions. Specifically, the dataset contains 2650 image-language-label triplets and 7 object categories in total. The training set has 2118 triplets, the validation set has 268 triplets, and the remaining 264 triplets are in the test set. Each remote sensing image has a size of 1024×1024 pixels, containing varying scales and details, and covers an area ranging from 0.06 km^2 to 25 km^2. The average length of the language descriptions is 19.6 words, with a minimum of 6 words and a maximum of 41 words. More detailed information about the RRSIS-HR dataset can be found in Table I, and some visual examples are shown in Fig. 2(a).

C. Results and Analysis

Table II presents the overall results of different methods on the RRSIS-D dataset. It can be observed that our method achieves the best performance across multiple evaluation metrics, including mIoU, oIoU, and precision scores from Pr@0.5 to Pr@0.8. Notably, our method outperforms the second-best RMSIN by 1.42% in mIoU, 0.59% in oIoU, and 1.98% in Pr@0.5, and 1.78% in Pr@0.6. This result suggests that our method achieves strong segmentation performance at low and medium overlap thresholds. However, its performance at the stricter Pr@0.9 threshold is less competitive, potentially due to the limitations of the model in capturing high-precision details, as it prioritizes overall segmentation accuracy. Fig. 6 shows the visual segmentation results of different methods on the RRSIS-D test set, along with the corresponding IoU scores. It can be observed that, compared to the model RMSIN [19], our CADFormer exhibits superior segmentation performance across various remote sensing scenes. Specifically, our method produces more accurate pixel-level segmentation masks with higher IoU scores for different ground objects at various scales while significantly reducing misclassification errors.

We further evaluated our proposed method on the RRSIS-HR dataset. Considering the complexity of the dataset, we selected several state-of-the-art RRSIS methods and LAVT [24] as comparison models. The results are shown in Table III. It can be observed that our method achieves the best performance across all metrics. Specifically, our CADFormer outperforms the second-best RMSIN [19] by 10.98% in Pr@0.6, 11.01% in mIoU, and 7.32% in oIoU. Fig. 7 shows the visual segmentation results of different methods on the RRSIS-HR test set, along with the corresponding IoU scores. For better clarity, we marked the approximate locations of the target objects with yellow boxes in the first row of the original images. In addition, we highlighted the clearly incorrect predicted areas with yellow circles. As can be seen, our method CADFormer demonstrates superior segmentation performance by more accurately segmenting the target objects in complex remote sensing scenes. For instance, in the first column of Fig. 7, RMSIN [19] not only predicts the target object, but also mistakenly predicts other nonspecific category objects, as indicated by the yellow circles. In the third and fourth columns, RMSIN incorrectly predicts the storage tank on the right and the roundabout in the middle, respectively. In contrast, our CADFormer correctly predicts the target objects specified by the language expressions, which aligns with the task requirements of RRSIS. The quantitative and qualitative results on the RRSIS-HR test set indicate that when the resolution of the remote sensing images is very high, and the target objects are hidden in complex backgrounds with more complex language expressions, previous RRSIS methods fail to deliver satisfactory performance. However, our CADFormer can achieve accurate target segmentation, demonstrating its effectiveness and superiority.

D. Ablation Study

We conducted ablation experiments on the test sets of both the RRSIS-HR and RRSIS-D datasets to verify the effectiveness of the core modules in our method.

1) Effectiveness of SMGAM and TCMD: To evaluate the effectiveness of our proposed SMGAM and TCMD, we performed ablation studies on all combinations of SMGAM and TCMD, as illustrated in Table IV. The first row presents the experimental results of the model using traditional cross-modal alignment in LAVT [24], without SMGAM and TCMD, which only reaches 60.05% and 39.48% mIoU on the RRSIS-D and RRSIS-HR datasets, respectively. As can be seen in the second row, the introduction of SMGAM improves the mIoU by 2.34% on the RRSIS-D dataset and by 7.16% on the RRSIS-HR dataset. In the third row, we add TCMD for the base model. The results indicate that the introduction of TCMD leads to improvements

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE RRSIS-D TEST SET

Methods	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
RRN [30]	51.07	42.11	32.77	21.57	6.37	45.64	66.43
CMSA [33]	55.32	46.45	37.43	25.39	8.15	48.54	69.39
LSCM [46]	56.02	46.25	37.70	25.28	8.27	49.92	69.05
CMPC [47]	55.83	47.40	36.94	25.45	9.19	49.24	69.22
BRINet [34]	56.9	48.77	39.12	27.03	8.73	49.65	69.88
CMPC+ [48]	57.65	47.51	36.97	24.33	7.78	50.24	68.64
LAVT [24]	62.32	56.16	47.06	36.37	21.49	55.48	74.68
LGCE [18]	70.44	63.95	53.86	41.94	**24.53**	61.01	76.55
FIANet [27]	72.16	65.81	54.67	41.60	24.48	62.58	76.48
RMSIN [19]	72.22	65.84	55.44	42.26	24.10	62.35	76.67
CADFormer (Ours)	**74.20**	**67.62**	**55.59**	**42.37**	23.59	**63.77**	**77.26**

The best result is bold.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE RRSIS-HR TEST SET

Methods	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
LAVT [24]	23.11	20.08	13.64	5.3	0.38	22.78	27.94
FIANet [27]	31.06	27.65	22.35	15.15	1.89	27.13	28.89
LGCE [18]	35.98	31.06	23.86	15.15	3.79	33.48	38.20
RMSIN [19]	50.00	46.97	39.77	29.92	6.44	43.70	45.97
CADFormer (Ours)	**61.74**	**57.95**	**48.86**	**35.61**	**13.26**	**54.71**	**53.29**

The best result is bold.

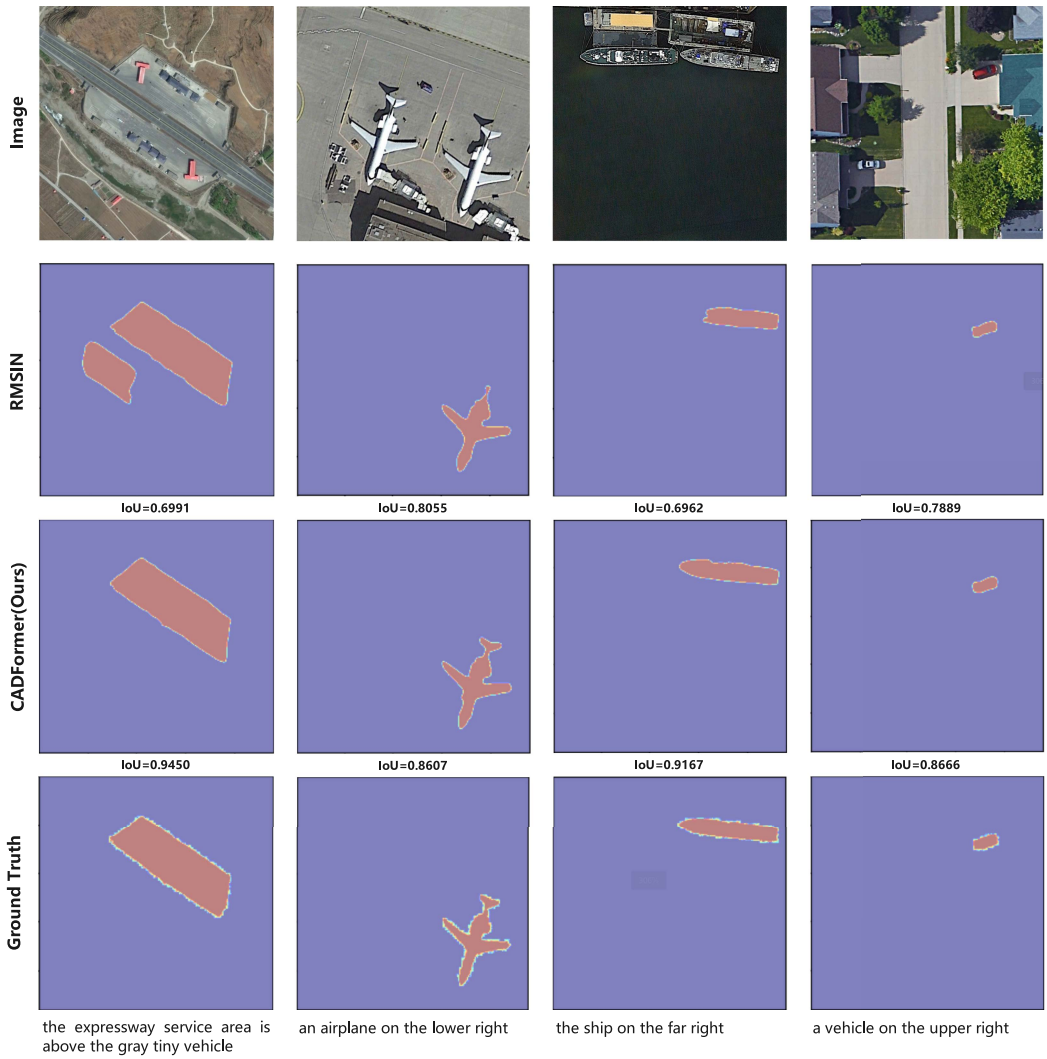

Fig. 6. Qualitative comparisons of different methods on the RRSIS-D test set. From top to bottom: original image, predictions by RMSIN, predictions by CADFormer, ground truth, and language expressions.

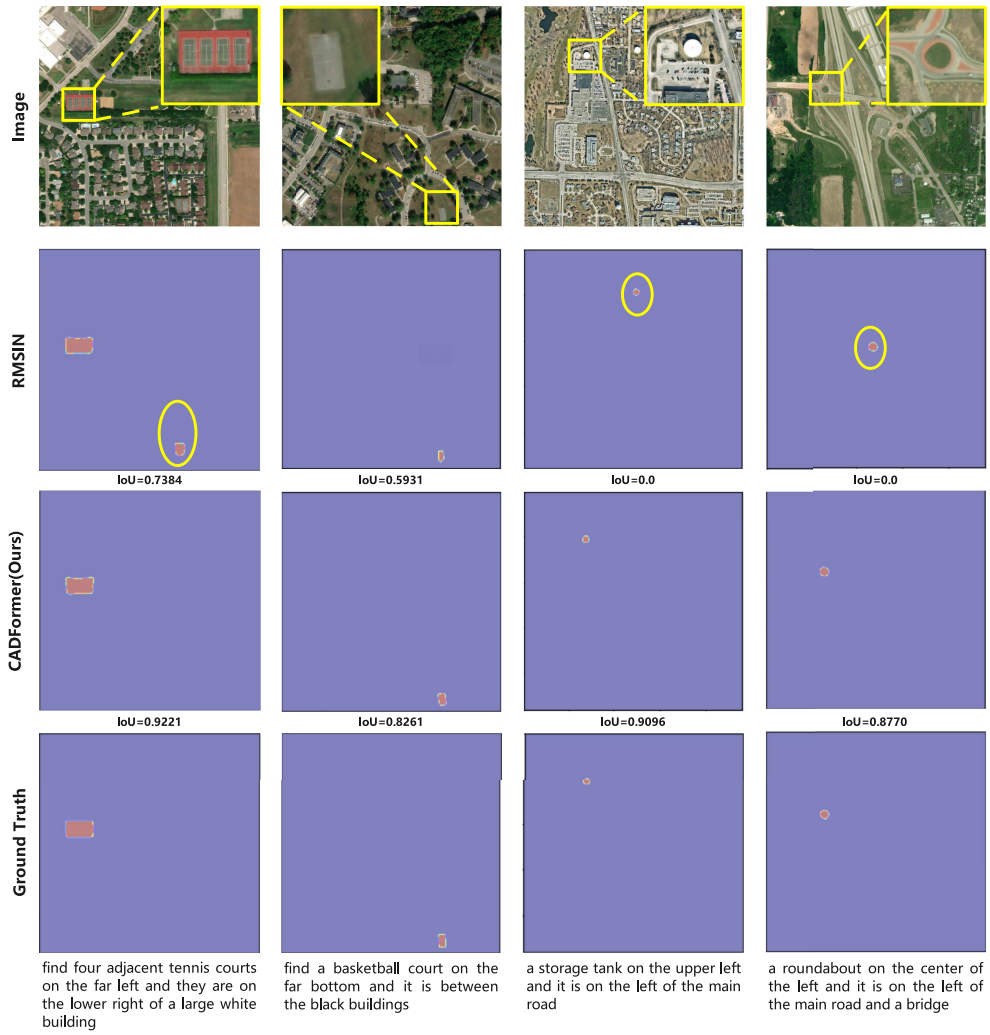

Fig. 7. Qualitative comparisons of different methods on the RRSIS-HR test set. From top to bottom: original image, predictions by RMSIN, predictions by CADFormer, ground truth, and language expressions. Yellow boxes indicate the approximate location of the target objects, while yellow circles highlight obvious incorrect predictions.

TABLE IV
ABLATION STUDIES ON THE SMGAM AND TCMD

SMGAM	TCMD	RRSIS-D			RRSIS-HR		
		Pr@0.6	mIoU	oIoU	Pr@0.6	mIoU	oIoU
✗	✗	62.31	60.05	76.37	40.15	39.48	41.92
✓	✗	65.76	62.39	**77.33**	48.86	46.64	48.96
✗	✓	66.53	63.61	76.89	55.68	53.03	53.02
✓	✓	**67.62**	**63.77**	77.26	**57.95**	**54.71**	**53.29**

Bold values indicate the best performance among all methods.

TABLE V
ABLATION STUDIES ON TWO SUBMODULES OF SMGAM

VGLVA	LGVLA	RRSIS-HR			
		Pr@0.6	Pr@0.7	Pr@0.8	mIoU
✗	✗	37.12	31.44	25.00	37.17
✓	✗	40.15	35.23	28.03	41.48
✗	✓	50.00	42.80	31.44	49.23
✓	✓	**57.95**	**48.85**	**35.61**	**54.71**

Bold values indicate the best performance among all methods.

in Pr@0.6, mIoU, and oIoU. The fourth row shows the complete model, our CADFormer. Although the value of oIoU is slightly lower by 0.07% compared to the second row, this complete model outperforms the base model without SMGAM and TCMD in all metrics. Specifically, the value of mIoU improved by 3.72% and 15.23% on the two datasets, respectively. These results demonstrate that our proposed SMGAM and TCMD are effective in improving the overall segmentation capability and play a crucial role in handling complex scenes.

2) SMGAM Analysis: To comprehensively assess the impact of two submodules in SMGAM, we conducted ablation studies on the test set of the RRSIS-HR dataset. As shown in Table V, the base model lacking VGLVA and LGVLA submodules exhibits significant performance degradation, achieving only 37.17% mIoU. When only the VGLVA submodule is used, the mIoU metric improves to 41.48%. The results indicate that the introduction of VGLVA brings 4.31% gains for the base model in mIoU. In contrast, the LGVLA submodule alone yields a more substantial improvement, boosting the mIoU by 12.06%. These results indicate that while the LGVLA module plays a more prominent role, the VGLVA module also contributes

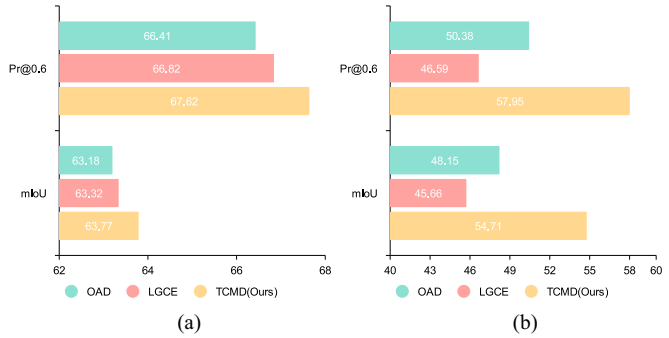

Fig. 8. TCMD module analysis, where OAD denotes the OAD, and LCGE represents that the model uses the standard decoder in LGCE. (a) Comparison on the RRSIS-D test (b) Comparison on the RRSIS-HR test.

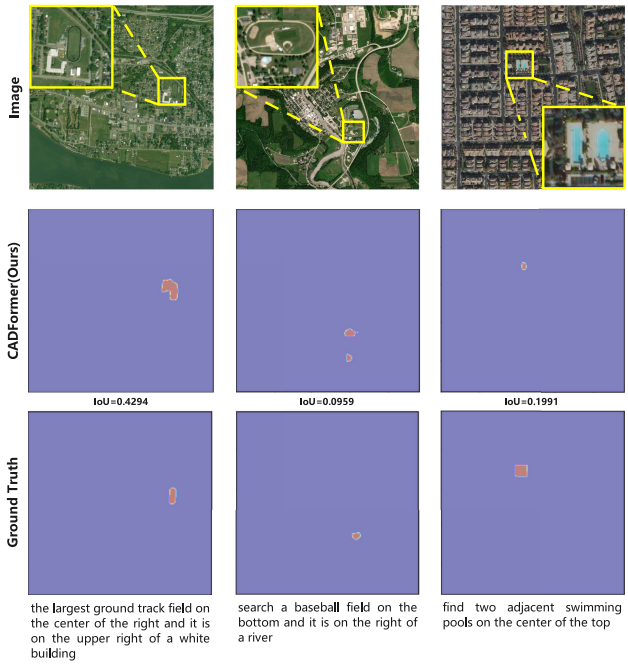

Fig. 9. Failure cases of our method on the RRSIS-HR test set. The approximate area around the target object is enlarged and indicated with yellow boxes in the first row.

critically to the cross-modal alignment process. The synergistic interaction between the two modules leads to optimal model performance, validating the importance of the semantic mutual guidance alignment approach.

3) TCMD Module Analysis: To further demonstrate the effectiveness of the proposed TCMD, we compared it with two existing decoders: the oriented-aware decoder (OAD) in RMSIN [19] and the standard segmentation decoder in LGCE [18]. We utilize the two decoders to replace the TCMD, respectively, maintaining the integrity of the remaining components within the CADFormer model. Fig. 8 presents the comparative experimental results of using different decoders, where Fig. 8(a) and (b) demonstrate the performance of the model on the test sets of the RRSIS-D and RRSIS-HR datasets, respectively. The experimental results indicate that our proposed TCMD method outperforms the other two decoders, which not only validates the importance of incorporating language features at the decoding stage, but also confirms the effectiveness of our proposed TCMD module.

E. Discussion and Analysis

1) Limitations: Although our CADFormer can effectively model and reason based on language expressions involving complex relationships, it still exhibits certain failure cases. There are primarily three types of failure cases, as illustrated in Fig. 9. The first type is shown in the first column. When the target object shares similar visual features with the background, it is challenging to accurately delineate the boundary of the target object. The second column presents the second type. The ambiguity, imprecision, and complexity of the language expressions may cause confusion. For instance, the phrase "a baseball field on the bottom" is ambiguous. In the given image, multiple baseball fields are present, and a distracting baseball field located at the bottom better matches the phrase than the intended target, making it easier for the model to misidentify the object. As shown in the third column, when the language expression refers to multiple targets, the model occasionally segments only a single instance. We argue that specialized methods are necessary for multiobject segmentation tasks, similar to the generalized RIS [39] task. We believe that multiobject segmentation based on language expressions in the remote sensing domain will become one of the major research directions in the future.

TABLE VI
COMPUTATIONAL COMPLEXITY COMPARISON OF DIFFERENT METHODS

Methods	FLOPs ↓	Params ↓	Inference Time ↓
LAVT [24]	383.84G	118.85M	0.0878 s
LGCE [18]	401.88G	167.37M	0.0930 s
RMSIN [19]	433.02G	240.04M	0.1264 s
FIANet [27]	435.87G	256.17M	0.1365 s
CADFormer (Ours)	468.14G	272.39M	0.1446 s

2) Efficiency and Complexity: To evaluate the computational efficiency and complexity, we report the floating point operations (FLOPs), the number of parameters (Params), and inference time (in seconds) per image on the test set of RRSIS-HR, as shown in Table VI. From the table, it can be observed that CADFormer has a higher computational cost in terms of FLOPs and Params compared to the other models. This is due to the additional complexity introduced by the fine-grained cross-modal alignment mechanism and the TCMD. However, this higher computational complexity also leads to superior model performance, as demonstrated in other experimental sections where CADFormer outperforms other models in multiple performance metrics. In addition, we demonstrate through ablation studies that each newly introduced module contributes to the final performance of the model. Regarding inference time, CADFormer requires 0.1446 s, which is slightly higher than RMSIN [19] (0.1264 seconds) and FIANet [27] (0.1365 s), but the increase is minimal and falls within a reasonable and acceptable range, without causing significant impact on practical applications.

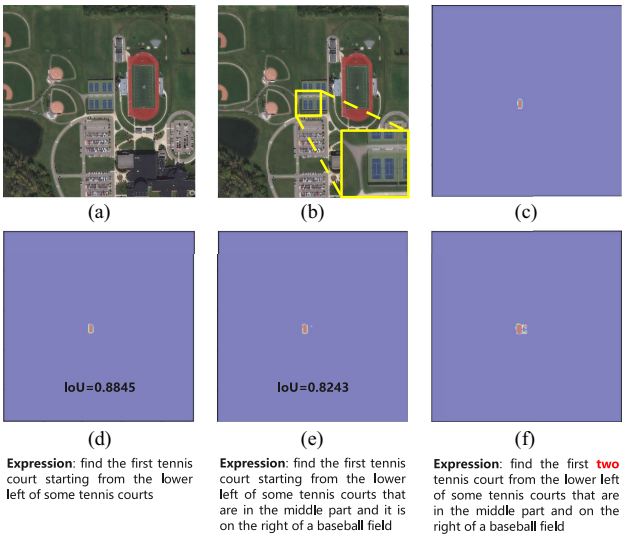

Fig. 10. Qualitative results of CADFormer on the same image with different language expressions. (a) Input image. (b) Input image with magnified detail. (c) Ground Truth for (d) and (e). (d) Result with short language expression. (e) Result with long language expression. (f) Result with multiobjective language expression.

3) Different Language Expressions: To explore the impact of different language expressions, we conduct additional experiments by feeding the same image with three types of language expressions, i.e., short language expression, long language expression, and multiobjective language expression, to our trained model, as shown in Fig. 10. For the short language expression, the model achieves strong segmentation performance with an IoU score of 0.8845, as shown in Fig. 10(d). Short descriptions are simple and direct, allowing the model to focus more effectively on key image regions, resulting in more accurate segmentation. When using the long language expression, the segmentation performance of the model slightly decreases, with an IoU score of 0.8243, as shown in Fig. 10(e). Long descriptions provide richer contextual details, which can help disambiguate target objects in complex scenes. However, they can also introduce irrelevant information, hindering segmentation accuracy. As shown in Fig. 10(f), for the multiobjective language expression, the model attempts to segment multiple targets but struggles with precision. This limitation arises because our dataset contains few multiobjective annotations and the model architecture is not optimized for such tasks. Future work should explore dedicated datasets and model designs for multiobject segmentation.

VI. Conclusion

In this article, we propose CADFormer, a novel RRSIS method based on semantic mutual guidance alignment and a TCMD, which excels at segmenting specific target objects in complex remote sensing scenes. Specifically, SMGAM aims to enhance the semantic correlation between visual and language features and achieve fine-grained cross-modal alignment, generating refined multiscale visual and refined language features based on semantic mutual guidance. In TCMD, we use the refined language features as contextual information to retrieve and aggregate referential object information from refined multiscale visual features, achieving precise segmentation. Besides, a new RRSIS dataset based on very high-resolution remote sensing images with longer, semantically richer language expressions is constructed to evaluate the performance of the existing RRSIS methods and our proposed methods in complex remote sensing scene understanding. Experimental results on two RRSIS datasets demonstrate that our CADFormer outperforms the majority of existing RRSIS methods. In addition, when dealing with complex scenes and language expressions, our method can generate fine-grained segmentation results. From the experiments, we find that semantic mutual guidance alignment facilitates fine-grained cross-modal alignment, while incorporating text features during the decoding process effectively improves segmentation accuracy. Future work could focus on developing a generalized RIS approach for remote sensing images, which can match multiple targets or no target based on language expressions.

Acknowledgment

The authors would like to thank the editors and reviewers for their instructive comments, which helped to improve this article.

References

[1] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.
[2] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1688.
[3] 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 Sens.*, vol. 60, 2022, Art. no. 5629419.
[4] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5606514.
[5] X. Ma, X. Zhang, X. Ding, M.-O. Pun, and S. Ma, "Decomposition-based unsupervised domain adaptation for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5645118.
[6] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403215.
[7] L. Wang et al., "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.
[8] X. Zhang, W. Wu, M. Zhang, W. Yu, and P. Ghamisi, "Prototypical unknown-aware multiview consistency learning for open-set cross-domain remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5643616.
[9] Y. Zhang, S. Yan, L. Zhang, and B. Du, "Fast projected fuzzy clustering with anchor guidance for multimodal remote sensing imagery," *IEEE Trans. Image Process.*, vol. 33, pp. 4640–4653, 2024.
[10] Y. Zhang, G. Jiang, Z. Cai, and Y. Zhou, "Bipartite graph-based projected clustering with local region guidance for hyperspectral imagery," *IEEE Trans. Multimedia*, vol. 26, pp. 9551–9563, 2024.
[11] K. Li, D. Wang, H. Xu, H. Zhong, and C. Wang, "Language-guided progressive attention for visual grounding in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5631413.
[12] K. Chen et al., "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701117.

[13] Z. Zhu, C. E. Woodcock, J. Rogan, and J. Kellndorfer, "Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using landsat and sar data," *Remote Sens. Environ.*, vol. 117, pp. 72–82, 2012.
[14] X. Zhang, W. Yu, M.-O. Pun, and W. Shi, "Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 1–17, 2023.
[15] B. Zhang et al., "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.
[16] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.
[17] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, 2020, Art. no. 111402.
[18] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "Rrsis: Referring remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5613312.
[19] S. Liu et al., "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26658–26668.
[20] Y. Zhan, Z. Xiong, and Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604513.
[21] M. Lan, F. Rong, H. Jiao, Z. Gao, and L. Zhang, "Language query based transformer with multi-scale cross-modal alignment for visual grounding on remote sensing images," *IEEE Trans. Geosci Remote Sens.*, vol. 62, 2024, Art. no. 5626513.
[22] L. Wang, S. Dong, Y. Chen, X. Meng, S. Fang, and S. Fei, "Metaseg-net: Metadata-collaborative vision-language representation learning for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5644211.
[23] S. Dong, L. Wang, B. Du, and X. Meng, "Changeclip: Remote sensing change detection with multimodal vision-language representation learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 208, pp. 53–69, 2024.
[24] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18155–18165.
[25] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, vol. 1, pp. 4171–4186.
[26] Y. X. Chng, H. Zheng, Y. Han, X. Qiu, and G. Huang, "Mask grounding for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26573–26583.
[27] S. Lei, X. Xiao, T. Zhang, H.-C. Li, Z. Shi, and Q. Zhu, "Exploring fine-grained image-text alignment for referring remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2024, Art. no. 5604611.
[28] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 108–124.
[29] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 1271–1280.
[30] R. Li et al., "Referring image segmentation via recurrent refinement networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5745–5753.
[31] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.
[32] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process Syst.*, 2017.
[33] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10502–10511.
[34] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4424–4433.
[35] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16321–16330.
[36] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 5729–5739.
[37] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao, "Towards robust referring image segmentation," *IEEE Trans. Image Process.*, vol. 33, pp. 1782–1794, 2024.
[38] C. Liu, H. Ding, and X. Jiang, "GRES: Generalized referring expression segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23592–23601.
[39] Y. Hu et al., "Beyond one-to-one: Rethinking the referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4067–4077.
[40] Y. Pan, R. Sun, Y. Wang, T. Zhang, and Y. Zhang, "Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 2031–2040.
[41] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
[42] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[43] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang, and X. Huang, "Visual grounding in remote sensing images," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 404–412.
[44] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[45] T. Wolf, "Transformers: State-of-The-Art Natural Language Processing," *arXiv:1910.03771*, 2020.
[46] T. Hui et al., "Linguistic structure guided context modeling for referring image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 59–75.
[47] S. Huang et al., "Referring image segmentation via cross-modal progressive comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10488–10497.
[48] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4761–4775, May 2021.

Maofu Liu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Wuhan University, Wuhan, China, in 2005, 2002, and 1998, respectively.

He is currently a Professor with the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China. He is the author of five books and more than 120 articles. His main research interests include natural language processing, and multimodal semantic analysis and large language model.

Prof. Liu is the Senior Member of CCF (China Computer Federation) and members of ACM.

Xin Jiang received the B.E. degree in software engineering from Wuhan University of Science and Technology, Wuhan, China, in 2024. He is currently working toward the M.S. degree in computer science and technology with the School of Wuhan University of Science and Technology.

His research interests include natural language processing and multimodal semantic analysis.

Xiaokang Zhang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018.

From 2019 to 2022, he was a Postdoctoral Research Associate with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong, Shenzhen, China. He is currently a specially appointed Professor with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China. He has authored or coauthored more than 50 scientific publications in international journals and conferences. He is currently a Reviewer for more than 40 renowned international journals and conferences. His research interests include remote sensing image analysis, computer vision and machine learning.