CLIP-Driven with Dynamic Feature Selection and Alignment Network for Referring Remote Sensing Image Segmentation

Qianqi Lu , Yuxiang Xie * , Jing Zhang , Yanming Guo , Yingmei Wei , Jie Jiang , Xidao Luan

Posted Date: 17 September 2025

doi: 10.20944/preprints202509.1502.v1

Keywords: Remote sensing images; referring image segmentation; Vision and Language Alignment; CLIP

Article

CLIP-Driven with Dynamic Feature Selection and Alignment Network for Referring Remote Sensing Image Segmentation

Qianqi Lu [1], Yuxiang Xie [1,*], Jing Zhang [1], Yanming Guo [1], Yingmei Wei [1], Jie Jiang [1] and Xidao Luan [2]

[1] College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
[2] College of Computer Science and Engineering, Changsha University, Changsha 410000, China
* Correspondence: yxxie@nudt.edu.cn

Abstract

Referring Remote Sensing Image Segmentation (RRSIS) aims to accurately locate and segment target objects in high-resolution aerial imagery based on natural language descriptions. The current RRSIS model faces numerous gaps due to the significant differences between remote sensing images and natural images, including scale variations, object rotation, and the difficulty of matching complex linguistic queries with spatially variable targets. Existing methods often rely on high-level semantic features or multi-stage cross-modal alignment, resulting in long training times and inefficiencies with complex queries. In this context, we propose CLIP-Driven with Dynamic Feature Selection and Alignment Network(CD2FSAN) , a novel framework that includes information-theoretic feature selection, adaptive multi-scale aggregation and alignment , and dynamic rotation correction decoder to better align remote sensing visual features with textual descriptions. CD2FSAN dynamically selects visual features that best match the language description based on the principle of maximizing cross-modal information, alleviating domain shift from the pretraining of CLIP on natural images, and integrates language information during encoding. The framework also incorporates a multi-scale feature aggregation and alignment mechanism, ensuring precise cross-modal alignment, particularly for small targets. Additionally, CD2FSAN introduces a differentiable affine transformation-based dynamic rotation correction mechanism, enabling the network to adaptively adjust object orientations, improving segmentation accuracy. Experiments on three standard datasets, RefSegRS, RRSIS-D, and RISBench, demonstrate CD2FSAN's superior performance in terms of oIoU, mIoU, and precision. Ablation studies and qualitative visualizations validate the efficacy of each module, confirming the framework's robustness in handling spatial variations, rotation, and cross-modal alignment, significantly reducing the cross modality gap in CLIP-based single-stage RRSIS tasks.

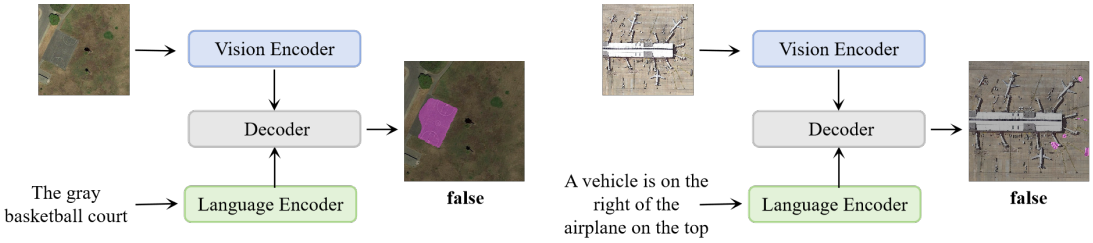
Keywords: Remote sensing images; referring image segmentation; Vision and Language Alignment; CLIP

1. Introduction

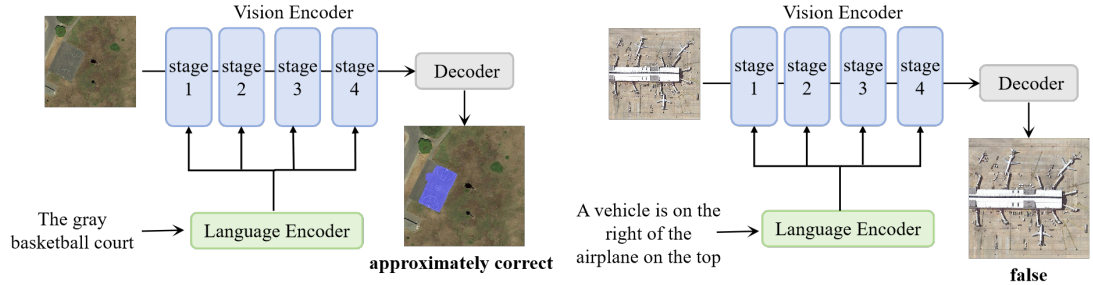
Referring Remote Sensing Image Segmentation (RRSIS) [1,2] is a novel task that combines remote sensing image analysis [3] with language expression. Unlike traditional segmentation methods [4,5] with known and fixed category labels, RRSIS performs open-domain segmentation based on free-form textual descriptions. This enables precise identification of specific targets in complex aerial scenes, making it valuable for applications such as urban planning [6], disaster assessment [7], environmental monitoring [8], precision agriculture [9] and land cover classification [10]. Recent advances in referring remote sensing image segmentation (RRSIS) have demonstrated promising results by extending techniques from Referring image segmentation(RIS), but the unique characteristics of aerial imagery, such as diverse spatial scales and arbitrary object orientations, pose additional challenges.

 doi:10.20944/preprints202509.1502.v1

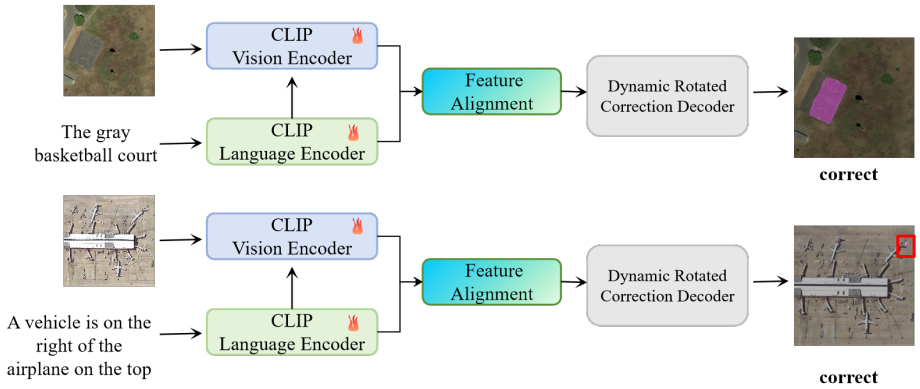
Current RRSIS methods primarily follow two representative cross-modal alignment paradigms. The first is the encode-then-decode paradigm [11,12], illustrated in Figure 1(a), where visual and textual features are independently encoded and fused during decoding. Although this approach is modular, it often suffers from weak cross-modal interaction, leading to suboptimal segmentation—particularly for remote sensing scenes with small, rotated, or scale-diverse targets. As shown in Figure 1(a), it fails to accurately delineate arbitrarily oriented objects and frequently produces false positives or incomplete masks for small targets, reflecting its limited adaptability to the complexities of remote sensing imagery. The second is the language-aware vision encoder paradigm [13–15], as shown in Figure 1(b), which embeds linguistic features into visual representations during encoding via cross-attention. This enables early vision-language interaction and supports a lightweight convolutional decoder design, improving segmentation accuracy. However, repeated cross-attention layers incur high computational cost and hinder parallelization, especially on large-scale, high-resolution datasets. Moreover, the model's limited capacity to capture fine-grained multi-scale features makes it less effective for small or densely distributed targets.



(a) Encode-then-decode cross-modal alignment paradigm



(b) Language-aware vision Encoder paradigm



(c)Our CLIP-driven dynamic feature selection and alignment paradigm

Figure 1. Illustration of three cross-modal alignment paradigms for RRSIS

Furthermore, both existing paradigms typically adopt isolated visual and textual encoders, preventing them from leveraging the shared multimodal priors learned by vision-language models such as CLIP [16]. While recent works like CRIS [17] have shown the benefits of integrating CLIP into natural image segmentation, directly applying CLIP-based models to remote sensing scenarios remains challenging. This difficulty arises not from inherent flaws in CLIP itself, but from a significant domain gap: CLIP is pre-trained on natural image–text pairs, which differ substantially from remote

 doi:10.20944/preprints202509.1502.v1

sensing data in both visual appearance and semantic structure. Remote sensing images exhibit abstract patterns, top-down views, and domain-specific object categories that rarely appear in CLIP's pretraining distribution. As a result, cross-modal representations extracted by CLIP often suffer from distributional mismatch, degrading alignment accuracy, particularly for small, rotated, or densely distributed targets. Therefore, beyond fine-tuning CLIP, it is crucial to design dedicated modules that explicitly adapt its representations to the unique characteristics of remote sensing imagery.

To address the limitations of existing methods, we propose CLIP-Driven Dynamic Feature Selection and Alignment Network (CD2FSAN), a framework tailored for RRSIS. CD2FSAN integrates three task-aware modules to enhance cross-modal alignment, spatial adaptability, and geometric robustness in complex aerial scenes, within a single-stage CLIP-based RRSIS pipeline. We improve the CLIP visual encoder with a mutual-information-inspired dynamic feature selection mechanism. Rather than relying solely on the final high-level layer, we rank cross-modal similarity scores to pick language-consistent features from CLIP's full hierarchy. This leverages multi-level cues to enable earlier and more informative text-vision coupling. To improve multi-scale perception and enhance small object detection, the Multi-scale Aggregation and Alignment Module (MAAM) aligns features using asymmetric, dilated, and depthwise separable convolutions. This design captures fine-grained targets efficiently and supports multiscale representation with low computational cost. To address arbitrary object orientations and positions in remote sensing images, we propose the Dynamic Rotation Correction Decoder (DRCD) employing a Dynamic Rotation Correction (DRC) mechanism. When targets undergo arbitrary rotation, the convolution kernel becomes misaligned with the target's edge direction, making it difficult to fully leverage local spatial correlation. This prevents the model from aggregating similar pixels along edge directions and instead averages pixels on both sides of the boundary, leading to blurred edges and missed detections. The mechanism predicts the rotation angle for each sample and re-parameterizes the convolution kernel via variable affine transformations. This "kernel steering" aligns the receptive field with the estimated object orientation, enabling the network to aggregate local spatial autocorrelation along object boundaries rather than across them. As a result, it achieves precise segmentation of targets with variable orientations and positions. Together, these three innovations form a cohesive network that explicitly addresses the unique challenges of RRSIS-scale variation and orientation diversity, leading to more accurate and robust segmentation across complex aerial scenes.

The main contributions of this paper are as follows:

- We propose CD2FSAN (CLIP-Driven Dynamic Feature Selection and Alignment Network), a CLIP-tailored, single-stage framework for RRSIS that is jointly optimized to improve cross-modal alignment, small object localization, and geometry-aware decoding in complex aerial scenes.
- To enhance cross-modal alignment and fine-grained segmentation, we introduce an integrated visual–language alignment and geometry-aware decoding design. This design incorporates a mutual-information–inspired dynamic feature selection to pick language-consistent features across CLIP's hierarchy, and a Multi-scale Aggregation and Alignment mechanism (MAAM) that establishes scale-consistent representations to sharpen small-object localization. Additionally, a Dynamic Rotation Correction Decoder (DRCD) predicts per-sample rotation angles and re-parameterizes convolutional kernels via differentiable affine transformations, aligning receptive fields with object orientations and improving the delineation of rotated targets.
- Extensive experiments on three public RRSIS benchmarks show that CD2FSAN achieves state-of-the-art performance in both accuracy and efficiency. Ablation and visualization further confirm the effectiveness and interpretability of each component.

2. Related Work

2.1. Referring Image Segmentation

Referring Image Segmentation (RIS) aims to localize and segment the region in an image described by a natural language expression. This task requires fine-grained semantic alignment between vision

 doi:10.20944/preprints202509.1502.v1

and language, and has attracted increasing attention. Early approaches typically used convolutional and recurrent networks to independently encode visual and textual inputs, followed by simple fusion techniques such as multiplication or concatenation [18–20]. However, these naive combinations often led to weak cross-modal interaction and poor segmentation quality. To address this, subsequent methods introduced more advanced fusion schemes, including attention-based alignment, gated integration, and hierarchical reasoning modules, to enhance visual-language integration [15,21–23]. With the rise of Transformers, RIS methods began leveraging joint multimodal encoding to improve feature interaction. MDETR [24] and VLT [25] proposed unified decoder-based fusion strategies. Among them, LAVT [13] has become a widely used baseline, particularly for remote sensing tasks. It employs a Swin Transformer backbone [26] and injects language-guided attention at multiple encoder stages, which enhances the alignment between linguistic expressions and salient visual regions. While its hierarchical design facilitates cross-scale fusion, LAVT tends to focus on prominent or contextually obvious targets, and often struggles with small, cluttered, or densely distributed objects due to the absence of explicit spatial modeling or geometry-aware mechanisms. ReSTR [27] and CRIS [17] adopt dual-branch Transformer encoders followed by multimodal fusion, while PolyFormer [28] and SeqTR [29] redefine RIS as sequence prediction over boundary points. Other methods like GRES [30] and CGFormer [31] apply query-based proposal matching for region grounding. Collectively, these works demonstrate how Transformer-based architectures have reshaped RIS through more expressive and flexible cross-modal reasoning.

Despite this progress, most existing RIS models are tailored to natural images, where targets are typically salient, well-aligned, and semantically coherent with human language. In contrast, remote sensing imagery presents unique challenges: varied spatial resolution, cluttered backgrounds, and frequent presence of small or rotated targets. Such characteristics often degrade the performance of standard RIS models when applied to Referring Remote Sensing Image Segmentation (RRSIS). In this work, we build upon RIS advancements and propose a CLIP-guided framework with dynamic feature selection based on cross-modal information maximization and rotation-aware decoding, aiming to improve cross-modal alignment and segmentation performance under remote sensing conditions.

2.2. Referring Remote Sensing Image Segmentation

Referring Remote Sensing Image Segmentation (RRSIS) [32] is a newly emerged multimodal task that segments specific regions in aerial imagery based on natural language expressions. Compared with traditional semantic segmentation, RRSIS enables more flexible human-computer interaction but poses unique challenges due to the high resolution, rotation, and scale diversity in remote sensing scenes [3]. To advance research in this domain, Yuan et al. introduced the RRSIS task alongside the RefSegRS dataset [1], which comprises over 4,000 image-text-mask triplets. They further proposed the LGCE module, built upon the LAVT framework, to enhance multi-scale visual-linguistic fusion. Despite its foundational contributions, RefSegRS exhibits notable limitations, including ambiguous boundaries between instances and classes, as well as a lack of linguistic diversity. For example, expressions such as "road" are often annotated with masks that encompass all road regions in the image, thereby blurring the distinction between referring segmentation and open-vocabulary semantic segmentation [33] or GRES [30]. To address these limitations, Liu et al. proposed RRSIS-D [14], a large-scale benchmark with 17,402 triplets across 20 categories and 7 attributes. Featuring fine-grained, rotated, and small objects with high-quality semi-automatic annotations, it has become the most widely adopted dataset in RRSIS research. Built on this dataset, RMSIN was proposed to improve hierarchical alignment through spatial- and scale-aware interaction. More recently, RISBench [34] extended dataset diversity by introducing over 52,000 samples with more varied object categories and complex language structures. CroBIM was introduced as a strong baseline, emphasizing bidirectional interaction and spatial reasoning. However, as RISBench is newly released and currently in preprint, its adoption remains limited. Each dataset presents distinct strengths and trade-offs. RSRefSeg provides basic coverage with coarse annotations, RRSIS-D emphasizes precision and rotation awareness, while RISBench focuses on linguistic diversity

 doi:10.20944/preprints202509.1502.v1

and generalization. We conduct comprehensive experiments across all three to ensure a robust and multidimensional evaluation of our proposed model.

Alongside these benchmarks, recent methods such as FIANet [35], MAFN [36], and RSRefSeg [37] have pushed the field forward. FIANet enhances vision-language alignment by disentangling object and positional cues; MAFN adopts correlation-guided multi-scale fusion to handle rotation and scale variance; RSRefSeg leverages prompt-driven segmentation via SAM, showcasing the potential of foundation models in remote sensing.

Nevertheless, most existing methods focus on only one aspect, alignment, scale variance, or geometric distortion, while leaving others unaddressed. Additionally, few approaches leverage the powerful semantic priors embedded in CLIP-like models, due to the vision-language domain gap and lack of adaptive alignment mechanisms. In order to bridge these gaps, we build on RIS advances and present CD2FSAN, a CLIP-tailored single-stage framework that unifies mutual-information–inspired dynamic feature selection (using cross-modal similarity as a surrogate), a Multi-scale Aggregation and Alignment mechanism (MAAM), and a Dynamic Rotation Correction Decoder (DRCD) with differentiable affine transformations, thereby strengthening cross-modal alignment, small-object localization, and rotation-robust segmentation under remote sensing conditions.

3. Materials and Methods

3.1. Overview

As illustrated in Figure 2, we propose CD2FSAN, a CLIP-driven segmentation framework tailored for Referring Remote Sensing Image Segmentation (RRSIS) to address the challenges of semantic misalignment, spatial heterogeneity, and geometric distortion in aerial imagery.

At the front end of the architecture, the CLIP visual encoder is enhanced through a dynamic feature selection mechanism based on mutual information maximization. During the visual encoding stage, cross-modal similarity is computed between sentence-level text features F_T and CLIP's hierarchical intermediate visual embeddings for subsequent cross-modal aggregation and alignment. This process yields a adaptive feature pyramid composed of low-level F_{V1}, mid-level F_{V2}, and high-level F_{V3} representations, facilitating early-stage alignment between language and vision cues while preserving both fine-grained and abstract semantics. To enhance spatial awareness and facilitate effective multi-scale fusion, the Multi-scale Aggregation and Alignment Module (MAAM) consolidates the hierarchical visual features F_{V1}, F_{V2}, F_{V3} into an intermediate representation F_F firstly. Then this module introduces a hybrid alignment strategy that integrates Image Multi-scale Convolution (IMC) and Text Multi-scale Convolution (TMC). IMC employs directional and dilated convolutions to capture diverse spatial patterns within remote sensing imagery, while TMC applies scale-adaptive 1D convolutions to extract hierarchical linguistic structures from word-level embeddings F_W. Through joint self-attention and cross-attention mechanisms, MAAM produces an aligned representation F_A that encodes both spatial detail and semantic consistency across modalities. The final prediction is generated by the Dynamic Rotated Correction Decoder (DRCD), which is designed to address orientation variability, a key obstacle in remote sensing segmentation. Based on the input F_A, the decoder predicts sample-specific rotation angles θ and dynamically transforms convolutional kernels via differentiable affine operations. This rotation-aware process aligns the convolutional receptive fields with the dominant object orientations, resulting in a robust pose-adaptive decoding pipeline. The output features are progressively refined through a top-down pathway to produce the final segmentation mask M, with particular efficacy in capturing arbitrarily rotated or densely packed targets.

 doi:10.20944/preprints202509.1502.v1

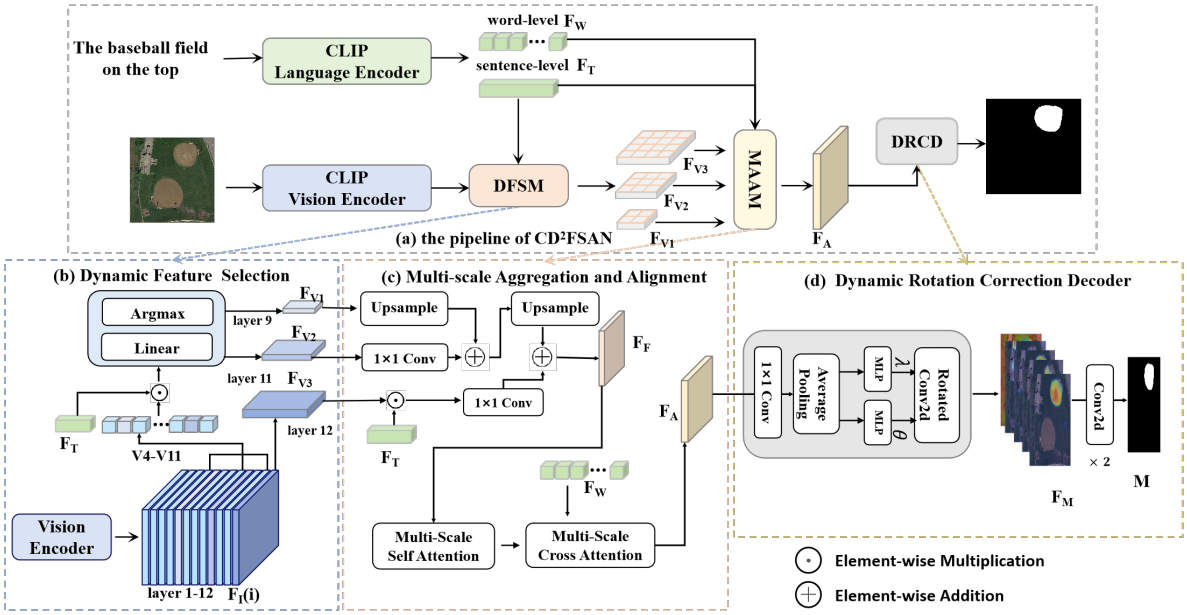


Figure 2. The overall architecture of the proposed CD2FSAN framework and detailed structures of its three core modules. (a) The end-to-end pipeline of CD2FSAN, which integrates cross-modal semantic priors from CLIP encoders with specialized modules for remote sensing segmentation. (b) Dynamic Feature Selection Mechanism Based on Cross-Modal Information Maximization, which adaptively selects and organizes hierarchical visual features from multiple levels of the CLIP vision encoder based on their semantic relevance to the input expression, thereby enhancing the expressiveness and discriminability of visual representations for downstream fusion. (c) The Multi-scale Aggregation and Alignment Module (MAAM), which aggregates hierarchical features and performs cross-modal alignment through self-attention and scale-aware convolutional operations on both image and text modalities. (d) The Dynamic Rotated Correction Decoder (DRCD), which employs learnable rotated correction to generate orientation-adaptive features for accurate segmentation of arbitrarily oriented targets.

3.2. Image and Text Feature Encoding

Image Encoder. We adopt the CLIP-pretrained Vision Transformer (ViT-B) as our image encoder to extract hierarchical visual representations. For an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we first split it into small, non-overlapping pieces of a set size $P \times P$, and each piece is turned into a token embedding. This results in a patch sequence $\mathbf{I}_p \in \mathbb{R}^{h\times w\times c}$, where $(h, w) = (H/P, W/P)$ and c denotes the embedding dimension. The patch embeddings are then processed through a 12-layer Transformer encoder following the ViT-B architecture. We denote the output of the i-th Transformer block as $\mathbf{F}_I(i) \in \mathbb{R}^{h\times w\times c}$, for $i = 1, \ldots, 12$, capturing progressively enriched visual features across layers.

Text Encoder. For language representation, we utilize the CLIP text encoder to process the input referring expression. The expression is tokenized and prepended with a special [SOS] token and appended with an [EOS] token to indicate sequence boundaries. The encoder outputs a sequence of contextualized word embeddings denoted as F_W, representing fine-grained linguistic cues. In addition, the final embedding corresponding to the [EOS] token is extracted as the sentence-level representation F_T, which encapsulates the global semantics of the entire expression.

3.3. Dynamic Feature Selection Mechanism

As illustrated in Figure 3, different layers of the CLIP image encoder exhibit distinct attention distributions across semantic regions in remote sensing images, reflecting varying levels of abstraction and spatial granularity. Despite this, existing CLIP-based referring segmentation models commonly rely solely on the final-layer visual features for cross-modal interaction, thereby overlooking the rich and complementary semantic cues embedded in intermediate layers. Motivated by insights from SegFormer [38] and related hierarchical fusion frameworks [39], we initially explored a simple approach by randomly selecting two intermediate layers and fusing them with final-layer features through a pyramid fusion strategy. While this method occasionally yielded marginal improvements, experiment

 doi:10.20944/preprints202509.1502.v1

results showed that the performance gains were inconsistent across samples and configurations. In particular, we observed that the segmentation mIoU did not exhibit a clear advantage over using the final layer alone, and the results were highly sensitive to the number of the randomly chosen layers. To systematically study this phenomenon, we conducted comprehensive experiments and visualization analyses on the CLIP visual layer under the same image-text pairings. These findings suggest that layer choice should not be random but language query-adaptive. We therefore replace heuristic fusion with a mutual-information-inspired dynamic selection that casts layer selection as cross-modal information maximization. Concretely, given the sentence-level embedding and CLIP's hierarchical visual features, we compute a cross-modal cosine similarity between and each as a tractable surrogate for cross-modal information, rank the layers, and forward the top-ranked language-consistent features to subsequent alignment. This per-sample selection exposes textual cues early in the encoder and reduces the natural-to-remote-sensing domain bias observed when relying solely on the final CLIP layer, yielding more stable alignment and stronger downstream segmentation.

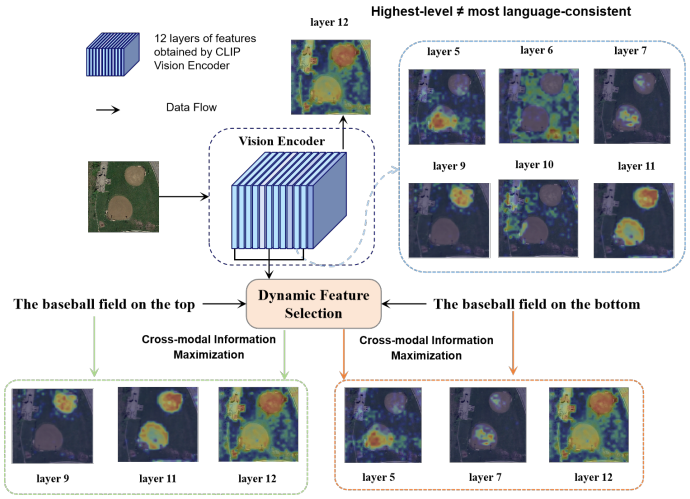


Figure 3. Illustration of the dynamic feature selection behavior of the proposed DFSM. Given the same remote sensing image but different referring expressions, DFSM dynamically selects different CLIP visual encoder layers for cross-modal fusion. The selected layers exhibit higher semantic alignment with the linguistic input, enabling more context-aware and accurate segmentation. This observation supports the motivation for dynamic, expression-dependent feature selection rather than relying solely on final-layer feature.

Specifically, we extract visual features from layers 4 to 12 of the CLIP image encoder, denoted as $\mathbf{F}_I(i) \in \mathbb{R}^{h\times w\times c}$, where $i = 4, 5, \ldots, 12$. For each layer, we use the global visual feature $\mathbf{F}'_\mathbf{I}(i)$ from the CLIP encoder, which is a pooled representation of the entire image. The global visual feature for the i-th layer is mapped to the same dimension as the text features using a shared linear transformation to obtain $\mathbf{V}(i)$, where $i = 4, 5, \ldots, 11$:

$$V_i = \text{Linear}(F'_I(i)) \in \mathbb{R}^d \tag{1}$$

The similarity score between the i visual layer and the text is defined as:

$$\text{Score}_i = \mathbf{V}_i \odot \mathbf{F}_T \tag{2}$$

where $\odot$ denotes element-wise multiplication. These similarity scores reflect the alignment between each layer's global visual semantics and the textual description. All Score_i values are concatenated and passed through a lightweight selection network to produce a refined score vector:

$$\mathbf{S} = \Phi_{\text{sel}}([\text{Score}_1, \ldots, \text{Score}_{11}]) \in \mathbb{R}^{11} \tag{3}$$

 doi:10.20944/preprints202509.1502.v1

The Φ_{sel} selection network plays a crucial role in identifying the most relevant feature layers based on semantic similarity scores. In this work, we employ a simple yet effective combination of linear layers followed by softmax activation as the selection mechanism. We select the values with the highest and second-highest semantic similarity scores from the vector $\mathbf{S}$ and record the corresponding indices k and m, with the constraint that $k < m$.

$$k, m = \arg\max_{i \neq j} (\mathbf{S}_i, \mathbf{S}_j), \quad \text{with} \quad k < m \tag{4}$$

These indices k and m are then used to directly select the corresponding feature layers F_{I_k} and F_{I_m} from the visual feature layers:

$$F_{V_1} = F_{I_k}, \quad F_{V_2} = F_{I_m}, \quad k < m \tag{5}$$

Here, F_{I_k} and F_{I_m} represent the visual feature layers selected based on the indices k and m, which correspond to the highest similarity scores in $\mathbf{S}$. We further include the final-layer output of the CLIP encoder as the high-level visual feature:

$$F_{V_3} = F_{I_{12}} \tag{6}$$

This feature F_{V_3} captures high-level global semantics of the image. Initially, all three features F_{V_1}, F_{V_2}, and F_{V_3} have the same spatial resolution, but to enhance feature expressiveness, we apply downsampling to introduce different spatial resolutions. As a result, these features share the same channel dimension C but differ in spatial resolution.

$$F_{V_3} \in \mathbb{R}^{H \times W \times C}, \quad F_{V_2} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}, \quad F_{V_1} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C} \tag{7}$$

This multi-scale representation enhances spatial granularity and contextual diversity, benefiting subsequent cross-modal alignment.

3.4. Multi-Scale Aggregation and Alignment Module

In the visual-linguistic fusion and alignment stage, we begin by fusing the visual features F_{V_1}, F_{V_2}, and F_{V_3}, obtained in the previous stage, with the global sentence-level textual features F_T through multi-level feature aggregation. This aggregation produces the initial combined representation $F_F \in \mathbb{R}^{H \times W \times C}$. To improve cross-modal alignment between the visual and textual modalities, we introduce a dual-attention mechanism, which includes both a multi-scale self-attention module and a cross-attention module. This dual-attention mechanism refines F_F, enhancing the alignment accuracy.

Unlike natural images, remote sensing imagery exhibits significant scale variations and often contains numerous small objects, which are frequently overshadowed by larger, more dominant features. This phenomenon makes it difficult for traditional models to accurately align these small objects with their corresponding textual descriptions. Moreover, the complexity of remote sensing scenes can lead to false positives during small object recognition, especially when background elements share similar visual features. To address these challenges, we suggest a multi-scale self-attention and cross-attention mechanism, designed to capture cross-scale correlations within the visual features and improve the representation of fine-grained objects. This mechanism, inspired by recent advances in remote sensing segmentation and referring image comprehension, allows the fused features F_F to interact with the textual embedding F_W across multiple scales. This interaction facilitates robust semantic alignment between the visual and textual modalities. The computational architecture for this multi-scale self-attention and cross-attention mechanism is illustrated in Figure 4(a).

 doi:10.20944/preprints202509.1502.v1

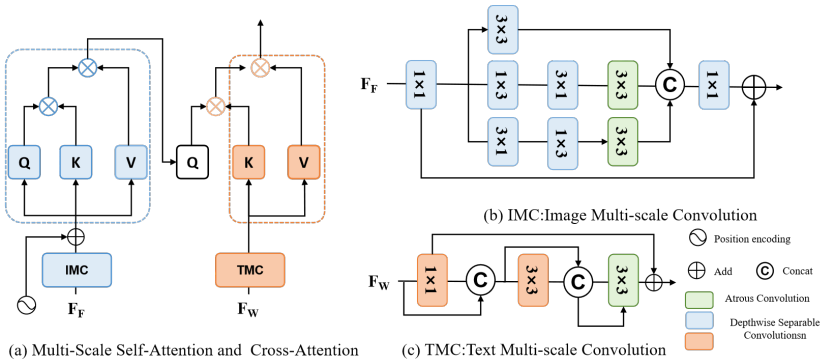


Figure 4. The computational architecture design of the proposed multi-scale alignment module. The module integrates both multi-scale self-attention and cross-attention mechanisms to enhance the semantic alignment between visual features and linguistic embeddings.

Specifically, given the three levels of fused multimodal features F_F, we apply multi-scale self-attention through the Image Multi-scale Convolution (IMC) module to further enhance the visual encoding. The IMC module captures discriminative semantics by following four main design principles: (1) multi-branch convolutional pathways using kernels of different sizes to enrich feature diversity; (2) dilated convolutions to expand the receptive field and capture a broader range of contextual dependencies; (3) depthwise separable convolutions to improve computational efficiency; (4) asymmetrically shaped convolutional kernels to capture horizontal and vertical features, respectively. The specific design of the IMC module is illustrated in Figure 4(b). Our proposed IMC module adopts a notably lightweight yet effective architecture, consisting of only three parallel branches. To improve computational efficiency and feature discriminability, we leverage depthwise separable convolutions, which significantly reduce the computational cost without sacrificing representational capacity. Moreover, to expand the receptive field and capture broader contextual information, we employ expansion convolutions in place of conventional convolution kernels. Specifically, each input feature map F_F is first processed by a 1×1 convolution for channel adjustment, and then passed through multiple convolutional branches, each designed to extract complementary semantic patterns. The outputs of these branches are subsequently aggregated to form the final representation. The overall computation of the IMC module is formalized as follows:

$$F'_F = \Phi_{\text{Conv}_{1\times1}}(F_F) \tag{8}$$

$$branch_1 = \Phi_{\text{Conv}_{3\times3}}(F'_F) \tag{9}$$

$$branch_2 = \Phi'_{\text{Conv}_{3\times3}}(\Phi_{\text{Conv}_{3\times1}}(\Phi_{\text{Conv}_{1\times3}}(F'_F))) \tag{10}$$

$$branch_3 = \Phi'_{\text{Conv}_{3\times3}}(\Phi_{\text{Conv}_{1\times3}}(\Phi_{\text{Conv}_{3\times1}}(F'_F))) \tag{11}$$

$$F_{MSA} = \Phi_{\text{Conv}_{1\times1}}[\text{Cat}(branch_1, branch_2, branch_3)] \oplus F'_F \tag{12}$$

where branch_1, branch_2, and branch_3 are the outputs of the three branches, each consisting of different convolution operations. Specifically, $\Phi_{\text{Conv}_{1\times1}}$ and $\Phi_{\text{Conv}_{3\times3}}$ are standard depthwise separable convolutions, while $\Phi_{\text{Conv}_{3\times1}}$ and $\Phi_{\text{Conv}_{1\times3}}$ refer to directional convolutions with different kernel shapes. Additionally, $\Phi'_{\text{Conv}_{3\times3}}$ represents a dilated convolution with a dilation rate of 5. The operator $\text{Cat}(\cdot)$ indicates channel-wise concatenation, and $\oplus$ denotes element-wise addition for residual enhancement. The use of directional convolutions (e.g., 1×3 and 3×1) improves the extraction of structured edge information, which is particularly beneficial for detecting small or elongated objects in remote sensing imagery. Moreover, the combination of multi-branch design and dilation allows the module to capture both fine-grained details and broader spatial context, thereby significantly improving the quality and discriminability of the visual representations.

 doi:10.20944/preprints202509.1502.v1

The enhanced visual features are subsequently aligned with their corresponding text embeddings through a cross-attention mechanism. To strengthen the multi-scale representation of language features, we introduce a novel Text Multi-scale Convolution (TMC) module. The architecture of TMC is shown in Figure 4(c). Building upon the principles of the Image Multi-scale Convolution (IMC), TMC employs convolutions with varying receptive fields to capture textual information across multiple semantic scales. However, in contrast to IMC, which utilises 2D convolutions to process spatial visual features, TMC leverages 1D convolutions specifically designed for sequential data, ensuring seamless compatibility with the textual feature structure. This innovative design enhances sequence-level encoding, while simultaneously maintaining tight alignment with the visual modality, facilitating richer cross-modal interaction and improving the overall feature fusion. Given the input textual embedding denoted as F_W, TMC generates a multi-scale enriched textual feature represented as:

$$F_{W_1} = \Phi_{\text{Conv}_{1\times1}}(F_W) \tag{13}$$

$$F_{W_2} = \Phi_{\text{Conv}_{3\times3}}[cat(F_W, F_{W1})] \tag{14}$$

$$F'_W = \Phi'_{\text{Conv}_{3\times3}}[cat(F_{W1}, F_{W2})] \oplus F_{W_1} \tag{15}$$

Finally, the refined visual features F'_F and the enhanced multi-scale text features F'_W are integrated through a cross-attention module, producing the aligned multimodal representation F_A. This process facilitates stronger semantic interaction and improves the localisation of fine-grained objects based on referring expressions.

3.5. *Dynamic Rotation Correction Decoder*

To address the challenges posed by the wide range of object orientations in remote sensing imagery, we propose a novel decoder module, termed the Dynamic Rotated Correction Decoder (DRCD). Unlike traditional convolutional decoders that utilise spatially fixed kernels, DRCD explicitly models rotation variance by dynamically generating orientation-aligned filters for each input sample. This allows the decoder to better adapt to pose diversity and improve segmentation accuracy for rotated or skewed targets. The decoder design is presented in Figure 2(d).

The core component of DRCD is the Dynamic Rotated Correction (DRC) mechanism. For each input sample b, the aligned multimodal representation $\mathbf{F}_A^{(b)} \in \mathbb{R}^{C\times H\times W}$, a lightweight routing network predicts a set of sample-specific rotation angles $\{\theta_i^{(b)}\}_{i=1}^n$ which specify the orientation for each of the n rotation bases, and gating weights $\{\lambda_i^{(b)}\}_{i=1}^n$, which control the contribution of each rotated base kernel to the final output. Here, n denotes the number of rotation bases used for the transformation. The predicted rotation angles are used to rotate each base kernel $\mathbf{W}_i \in \mathbb{R}^{C\times k\times k}$ using a differentiable affine transformation:

$$\mathbf{W}_i^{(b)} = \text{Rot}_\theta(\mathbf{W}_i, \theta_i^{(b)}), \quad i = 1, \ldots, n, \tag{16}$$

where $\text{Rot}_\theta(\cdot)$ denotes a bilinear grid-sampling rotation operator. The resulting rotated kernels are aggregated via a weighted summation, with the gating weights $\lambda_i^{(b)}$ determining the contribution of each rotated kernel to the final filter:

$$\mathbf{W}^{(b)} = \sum_{i=1}^{n} \lambda_i^{(b)} \cdot \mathbf{W}_i^{(b)}. \tag{17}$$

Finally, the reparameterized filters $\mathbf{W}^{(b)}$ are applied through grouped convolution to produce the segmentation-enhanced feature maps:

$$\mathbf{F}_M^{(b)} = \mathbf{W}^{(b)} * \mathbf{F}_A^{(b)}. \tag{18}$$

 doi:10.20944/preprints202509.1502.v1

This design effectively enhances the model's ability to handle the multi-orientation characteristics of targets in remote sensing images. By dynamically adjusting the convolution kernel based on the predicted angles, DRC can better align with the actual orientation of objects, leading to more accurate feature extraction and mask generation. The structural guidance in DRC ensures that the model focuses precisely on the target region, reducing the dispersion of attention and improving the clarity and accuracy of boundary details. This results in more precise segmentation masks, particularly for objects with complex shapes and orientations. Additionally, the replacement of a portion of the convolutional layers with DRC helps to reduce redundant feature learning, making the model more efficient and effective in capturing the essential features of remote sensing images.

4. Results

4.1. Dataset and Implementation Details

In this study, we evaluate the effectiveness of the proposed method using three publicly available remote sensing datasets: RefSegRS [1], RRSIS-D [14], and RISBench [34]. These datasets, which were recently introduced, significantly contribute to the progress of the Remote Sensing Image Segmentation (RRSIS) task.

- **RefSegRS**. The dataset comprises 4,420 image–text–label triplets drawn from 285 scenes. The training, validation, and test splits contain 2,172, 431, and 1,817 triplets, respectively, corresponding to 151, 31, and 103 scenes. Fourteen categories (for example, road, vehicle, car, van, and building) and five attributes are annotated. Images are 512×512 pixels at a ground sampling distance (GSD) of 0.13 m.
- **RRSIS-D**. This benchmark contains 17,402 triplets of images, segmentation masks, and referring expressions. The training, validation, and test sets include 12,181, 1,740, and 3,481 triplets, respectively. It covers 20 semantic categories (for example, airplane, golf field, expressway service area, baseball field, and stadium) and seven attributes. Images are 800×800 pixels, with GSD ranging from 0.5 m to 30 m.
- **RISBench**. The dataset includes 52,472 image–language–label triplets. The training, validation, and test partitions contain 26,300, 10,013, and 16,159 triplets, respectively. It features 26 categories with eight attributes. All images are resized to 512×512 pixels, with GSD ranging from 0.1 m to 30 m.

We evaluate with a CLIP-initialized ViT-B as the visual encoder and a Transformer as the language encoder. Images are resized to 480×480 pixels, the expression length is capped at 22 tokens (including [SOS] and [EOS]) on RefSegRS, RRSIS-D, and RISBench, and training uses Adam for 40 epochs with an initial learning rate of 5×10^{-5}, a batch size of 8, and two RTX 4090 GPUs.

Following prior studies, we use the following evaluation metrics:

- **Overall Intersection over Union (oIoU)**: This metric is calculated as the ratio of the cumulative intersection area to the cumulative union area across all test samples, with an emphasis on larger objects.
- **Mean Intersection over Union (mIoU)**: mIoU is computed by averaging the IoU values between predicted masks and ground truth annotations for each test sample, treating both small and large objects equally.
- **Precision@X**: Precision@X measures the percentage of test samples for which the IoU between the predicted result and the ground truth exceeds a threshold $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. This metric evaluates the model's accuracy at specific IoU thresholds, reflecting its performance in object localization.

4.2. Comparisons with Other Methods

To ensure a fair and comprehensive comparison, we adopt the experimental results reported in the original publications, supplemented by those reproduced in subsequent studies and publicly available benchmark papers. Our evaluation includes methods specifically tailored for remote sensing

 doi:10.20944/preprints202509.1502.v1

imagery, such as LGCE [1], RMSIN [14], FIANet [35], and CroBIM [34], as well as general-purpose referring image segmentation approaches like LAVT [13], CrossVLT [40], and CRIS [17]. Although the latter are not explicitly designed for remote sensing scenarios, they represent the state of the art in referring image segmentation and thus serve as valuable baselines.

As summarized in Table 1, referring remote sensing image segmentation remains at an early stage because existing datasets differ markedly in annotation volume, image sources, target object categories, object size distributions, and linguistic annotation style. Unlike referring natural image segmentation, which benefits from unified benchmark datasets such as RefCOCO and RefCOCOg, remote sensing lacks standardized benchmarks, leading to variability in model performance across datasets. Despite these challenges, CD2FSAN shows consistent and strong performance on three public benchmarks. It achieves state-of-the-art oIoU on both the validation and test sets, improving prior scores by 1.53%, 0.77%, and 0.81% on the RefSegRS, RRSIS-D, and RISBench validation sets, and by 1.60%, 0.64%, and 0.69% on the corresponding test sets. It also attains the highest mIoU on RRSIS-D and RISBench. These results confirm the effectiveness and robustness of the proposed architecture for diverse referring scenarios in remote-sensing imagery. The superior performance arises from three components acting synergistically: a dynamic feature selection mechanism for semantically aligned visual extraction, a multi-scale aggregation and alignment mechanism for robust cross-scale modelling, and a dynamic rotated correction decoder for enhanced orientation-aware decoding. Together, these components better address cross-modal alignment, geometric deformation, and fine-grained object delineation.

Table 1. Comparison of mIoU (%) and oIoU (%) for various methods on RefSegRS, RRSIS-D, and RISBench datasets.

Metric	Model	Publication	Visual Encoder	Text Encoder	RefSegRS		RRSIS-D		RISBench	
					Val	Test	Val	Test	Val	Test
oIoU	RRN [41]	CVPR-18	ResNet-101	LSTM	69.24	65.06	66.53	66.43	47.28	49.67
	BRINet [42]	CVPR-20	ResNet-101	LSTM	61.59	58.22	70.73	69.68	46.27	48.73
	ETRIS [43]	ICCV-23	ResNet-101	CLIP	72.89	65.96	72.75	71.06	64.09	67.61
	CRIS [17]	CVPR-22	CLIP	CLIP	72.14	65.87	70.98	70.46	66.26	69.11
	CrossVLT [40]	TMM-23	Swin-B	BERT	76.12	69.73	76.25	75.48	69.77	74.33
	LAVT [13]	CVPR-22	Swin-B	BERT	78.50	71.86	76.27	76.16	69.39	74.15
	LGCE [1]	TGRS-24	Swin-B	BERT	83.56	76.81	76.68	76.34	68.81	73.87
	RMSIN [14]	CVPR-24	Swin-B	BERT	74.40	68.31	78.27	77.79	69.51	74.09
	FIANet [35]	TGRS-25	Swin-B	BERT	85.51	78.28	76.77	76.05	-	-
	CroBIM [34]	Arxiv-25	Swin-B	BERT	78.85	72.30	76.24	76.37	69.08	73.61
	CD2FSAN (ours)	-	CLIP	CLIP	87.04	79.88	79.04	78.43	70.32	74.84
mIoU	RRN [41]	CVPR-18	ResNet-101	LSTM	50.81	41.88	46.06	45.64	42.65	43.18
	BRINet [42]	CVPR-20	ResNet-101	LSTM	38.73	31.51	51.41	49.45	41.54	42.91
	ETRIS [43]	ICCV-23	ResNet-101	CLIP	54.03	43.11	55.21	54.21	51.13	53.06
	CRIS [17]	CVPR-22	CLIP	CLIP	53.74	43.26	50.75	49.69	53.64	55.18
	CrossVLT [40]	TMM-23	Swin-B	BERT	55.27	42.81	59.87	58.48	61.54	62.84
	LAVT [13]	CVPR-22	Swin-B	BERT	61.53	47.40	57.72	56.82	60.45	61.93
	LGCE [1]	TGRS-24	Swin-B	BERT	72.51	59.96	60.16	59.37	60.44	62.13
	RMSIN [14]	CVPR-24	Swin-B	BERT	54.24	42.63	65.10	64.20	61.78	63.07
	FIANet [35]	TGRS-25	Swin-B	BERT	80.61	68.63	62.99	63.64	-	-
	CroBIM [34]	Arxiv-25	Swin-B	BERT	65.79	52.69	63.99	64.24	67.52	67.32
	CD2FSAN (ours)	-	CLIP	CLIP	76.95	66.96	66.47	65.37	68.36	69.74

On RefSegRS, CD2FSAN is competitive but slightly trails FIANet in mIoU because many annotations are class-level (e.g., all road segments labeled "road"), yielding large and coarse masks that reward broad semantic coverage over precise localization. FIANet benefits from this setting by capturing spatially extensive objects, and its Swin Transformer with fewer parameters may further aid adaptation to small datasets. By contrast, CD2FSAN targets small and arbitrarily oriented objects via multi-scale fusion and a dynamic rotated convolutional decoder, making it better suited to instance-level referring segmentation that requires precise localization. Consistent with this, CD2FSAN surpasses FIANet in both oIoU and mIoU across the other datasets, indicating stronger generalization to small, rotated, and structurally complex targets. Future work could integrate region-aware linguistic cues and stronger visual pretraining to improve mIoU under class-level supervision without compromising instance-level strengths.

To comprehensively evaluate the effectiveness of our method, we conducted more experiments on the RRSIS-D validation set and presented additional evaluation metrics. The specific results are

 doi:10.20944/preprints202509.1502.v1

shown in Table 2. CD2FSAN achieves state-of-the-art performance in both overall Intersection oIoU and mIoU, reaching 79.04% and 66.47% respectively outperforming all existing baselines. These results confirm the model's strong ability to balance instance-level localization and category-level segmentation consistency. In terms of Precision@X, our method also achieves the best results at X=0.5, X=0.6, and X=0.7, demonstrating robust localization under moderate confidence thresholds. Nevertheless, at higher thresholds (X = 0.8 and X = 0.9), the performance of our model is slightly inferior to that of early-fusion methods such as LAVT and LGCE. This suggests that CD2FSAN, which employs single-stage cross-modal feature alignment and a CLIP-based visual encoder, is less confident in assigning extremely high-probability predictions to foreground pixels. In contrast, models like LAVT integrate language features into the visual backbone across multiple stages and benefit from hierarchical token-wise refinement via Swin Transformer encoders. Such architectures may retain more precise spatial structure and support stronger confidence calibration.

Table 2. Performance comparison on the RRSIS-D validation set.

Method	oIoU	mIoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
LAVT [13]	76.27	57.72	65.23	58.79	50.29	40.11	23.05
CRIS [17]	70.98	50.75	56.44	47.87	39.77	29.31	11.84
LGCE [1]	76.68	60.16	68.10	60.61	51.45	**42.34**	**23.85**
FIANet [35]	76.77	62.99	74.20	66.15	54.08	41.27	22.30
RMSIN [14]	78.27	65.10	68.39	61.72	52.24	41.44	23.16
CroBIM [34]	76.24	63.99	74.20	66.15	54.08	41.38	22.30
CD2FSAN (Ours)	**79.04**	**66.47**	**78.28**	**70.11**	**56.78**	41.38	20.57

Nevertheless, CD2FSAN markedly outperforms CRIS, which also uses CLIP-based visual encoding, highlighting the benefits of our dynamic feature selection, multi-scale feature alignment, and dynamic rotated correction decoder. The remaining gap at high thresholds likely stems from the limited dense spatial granularity of CLIP features trained for global alignment. Recent advances mitigate this via segmentation-oriented prompts or pixel-aware backbones (for example, SAM decoders). We will explore incorporating auxiliary pixel-level prompts into the decoder to sharpen confidence discrimination for fine-scale foreground prediction. Importantly, CD2FSAN remains superior on the most representative metrics, oIoU and mIoU, consistently producing masks with accurate contours and spatial coverage, which underscores its overall effectiveness for remote-sensing image-text segmentation.

Table 3 presents a class-wise analysis on RRSIS-D. Beyond dataset-level metrics, CD2FSAN attains the best score in most land-cover categories (e.g., golf fields, baseball fields, vehicles, and basketball courts) and achieves the highest average mIoU of 69.98%, exceeding the second-best FIANet by 3.20 percentage points. As shown in Table 3, the model's performance varies by category. Objects with sparse boundaries and heavy clutter (bridge and golf field) remain challenging, whereas more homogeneous regions (building) are easier. Across most categories, CD2FSAN surpasses LGCE, RMSIN, and LAVT, with the largest gains on small-object classes such as vehicles and bridges; MAAM enhances multi-scale perception and fine-grained discrimination, improving localization of small, intricate targets. DFSM further strengthens the selection of semantically relevant visual features, and DRCD adapts convolutional kernels to object orientation, benefiting rotated targets such as tennis courts and ships where conventional decoders struggle to maintain alignment. Taken together, the consistently superior per-class mIoU indicates that CD2FSAN adapts well to diverse appearances and scales, from small, cluttered objects to large, rotated structures, while accurately capturing both fine details and overall structure.

 doi:10.20944/preprints202509.1502.v1

Table 3. Per-class mIoU (%) on the RRSIS-D validation set. Average is the unweighted mean across listed classes.

Category	LAVT [13]	RMSIN [14]	LGCE [1]	FIANet [35]	CD2FSAN (Ours)
Airport	66.44	68.08	68.11	**68.66**	68.61
Golf field	56.53	56.11	56.43	57.07	**64.22**
Expressway service area	76.08	76.68	71.19	**77.35**	72.31
Baseball field	68.56	66.93	70.93	70.44	**88.43**
Stadium	81.77	83.09	**84.90**	84.87	84.43
Ground track field	81.84	81.91	82.54	82.00	**83.06**
Storage tank	71.33	73.65	73.33	**76.99**	74.89
Basketball court	70.71	72.26	74.37	74.86	**88.43**
Chimney	65.54	68.42	68.44	68.41	**79.85**
Tennis court	74.98	76.68	75.63	**78.48**	72.88
Overpass	66.17	**70.14**	67.67	70.01	65.63
Train station	57.02	62.67	58.19	61.30	**68.32**
Ship	63.47	64.64	63.48	65.96	**68.32**
Expressway toll station	63.01	65.71	61.63	64.82	**72.32**
Dam	61.61	68.70	64.54	**71.31**	66.11
Harbor	60.05	60.40	60.47	**62.03**	57.77
Bridge	30.48	36.74	34.24	37.94	**43.53**
Vehicle	42.60	47.63	43.12	49.66	**49.72**
Windmill	35.32	41.99	40.76	46.72	**63.76**
Average	62.82	65.39	64.21	66.78	**69.98**

We also evaluate performance as a function of object size, using the classification standard defined by the RRSIS-D dataset, where instances are categorized based on their mask coverage.

$$\theta = \frac{|M|}{H \times W} \tag{19}$$

with Small defined as $\theta \leq 0.20$, Medium as $0.20 < \theta \leq 0.60$, and Large as $\theta > 0.60$. Under this classification, CD2FSAN outperforms all other methods across all size bins, achieving oIoU scores of 49.52, 73.56, and 83.80 for small, medium, and large objects, respectively, surpassing the strongest baselines by 1.77, 1.70, and 1.47 points, as shown in Figure 5. The improvement on small objects is primarily attributed to the MAAM, which uses asymmetric, dilated, and depthwise separable convolutions to capture fine details at low computational cost while focusing alignment on language-relevant features. Additionally, DFSM enhances early language-guided grounding by selecting semantically aligned layers from CLIP rather than relying solely on the final layer, thereby improving localisation without compromising performance on larger objects.

 doi:10.20944/preprints202509.1502.v1

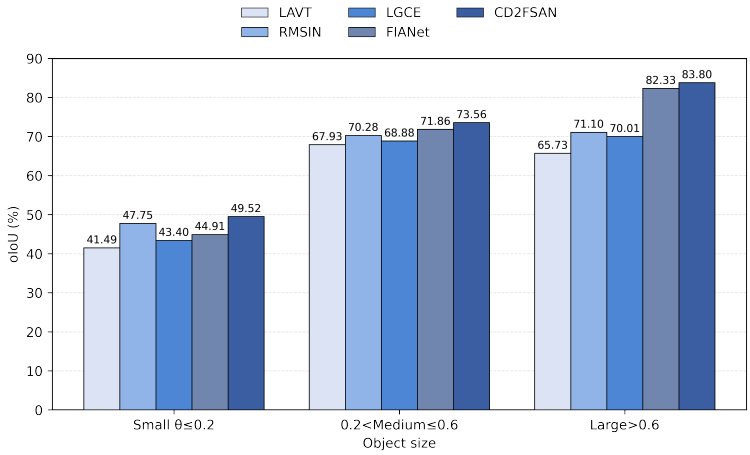


Figure 5. Size-wise oIoU on RRSIS-D. Instances are binned by mask coverage θ following the dataset paper: Small ($\theta \leq 0.20$), Medium ($0.20 < \theta \leq 0.60$), and Large ($\theta > 0.60$). Bars compare LAVT, RMSIN, LGCE, FIANet, and CD2FSAN.

4.3. Ablation Study

We perform ablation studies on the RRSIS-D validation subset and report Pr@0.5-0.9 and mIoU over seven variants grouped into three settings (Table 4a–g).

Baseline model. As shown in Table 4(a), the baseline uses the visual characteristics of the CLIP last layer, fused with the characteristics of the global language by multiplication by elements, and decodes the mask.

Effect of the dynamic feature selection mechanism. We ablate layer choice by randomly selecting two intermediate CLIP layers (from layers 4-11) and fusing them with the twelfth (final) layer, instead of using only the final-layer features as in the baseline (Table 4b-c). This strategy does not consistently outperform the final layer alone because the intermediate layers contribute unevenly to the task. When the sampled layers are not semantically aligned with the expression, they inject noise and degrade multimodal fusion (Table 4b). By contrast, DFSM (Table 4d) automatically identifies the visual layers containing the highest linguistic referential semantic information and retains only those for fusion. This data-driven selection yields consistent gains across metrics and improves both cross-modal alignment and segmentation accuracy, validating DFSM's effectiveness in selectively leveraging informative CLIP layers.

Effect of the multi-scale aggregat and alignment module. After DFSM (Table 4d), we compare two alignment strategies on the selected multilevel visual features: (e) a conventional transformer using self-attention and cross-attention, and (f) the proposed MAAM, which augments this design with hierarchical multi-scale self-attention and cross-attention (Table 4e and Table 4f). Both improve segmentation over (d); the standard Transformer raises mIoU by 9.69% on the validation set and 10.96% on the test set. MAAM yields further gains of 4.56% on validation and 3.37% on test relative to the standard Transformer, demonstrating stronger cross-scale semantic modeling and spatial-detail alignment. The advantage of MAAM comes from two components: an Image Multi-scale Convolution (IMC) block that uses multibranch convolutions with varied kernel sizes, dilated convolutions for expanded receptive fields, and asymmetric kernels for directional cues to better represent small and scale-varying objects; and a Text Multi-scale Convolution (TMC) block that enriches multi-scale language representations to align with visual features across semantic levels. Their integration strengthens pixel language correspondence and yields superior mIoU and related metrics, addressing remote sensing challenges such as large-scale variation and numerous small targets.

Impact of the dynamic rotation correction decoder.As shown in Table 4(g), integrating DRC into the decoder increases mIoU by 1.60% on the validation set and 2.82% on the test set relative to the preceding configuration. DRC dynamically orients convolutional kernels using predicted angles, improving feature extraction and mask generation, aligning the operations with actual object

 doi:10.20944/preprints202509.1502.v1

orientations, and suppressing redundant responses. These effects produce sharper, more accurate boundaries and boost overall segmentation accuracy, particularly for multi-oriented and complex-shaped targets common in remote-sensing imagery.

Table 4. Ablation study results on RRSIS-D (Validation/Test).

#	Method	Validation						Test					
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	mIoU	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	mIoU
(a)	Baseline	45.98	35.92	26.26	15.86	4.48	42.59	39.73	29.93	22.01	12.41	4.14	38.58
(b)	(a) + RFS(4,7,12)	44.19	34.02	24.71	14.65	3.67	41.75	38.67	28.51	18.10	10.34	2.12	37.39
(c)	(a) + RFS(6,9,12)	50.97	40.45	31.43	19.31	6.14	46.21	45.51	36.09	26.61	15.86	4.31	43.66
(d)	(a) + DFSM	57.82	51.32	42.53	32.30	15.40	50.62	52.24	42.53	31.15	12.53	5.74	48.22
(e)	(d) + Transformer	71.84	62.29	51.21	37.59	17.30	60.31	71.65	61.96	50.47	35.94	18.04	59.18
(f)	(d) + MAAM	76.32	66.60	54.82	40.80	19.71	64.87	72.39	62.37	50.83	36.11	17.81	62.55
(g)	(f) + DRCD (full)	**78.28**	**70.11**	**56.78**	**41.38**	**20.57**	**66.47**	**73.14**	**63.46**	**51.19**	**36.37**	**19.42**	**65.37**

4.4. Visualization and Qualitative Analysis

Figure 6 compares CD2FSAN with two representative baselines, RMSIN and CRIS, on RRSIS-D. RMSIN is a state-of-the-art method tailored to referring remote-sensing image segmentation with public code and pretrained weights, and CRIS is a CLIP-based framework for general referring image segmentation. As the field's most widely used benchmark, RRSIS-D spans varied object sizes and spatial contexts, making it a challenging and representative testbed. Across diverse image-text pairs, CD2FSAN produces more accurate and structurally coherent masks than both baselines, with sharper edges and cleaner boundaries, and gains are especially pronounced for small objects, rotated targets, and densely cluttered scenes typical of aerial imagery. Relative to RMSIN and CRIS, these improvements reflect complementary module design. RMSIN relies on task-specific modules without strong pretrained vision-language alignment and thus struggles with fine-grained distinctions and small or subtly referred objects. CRIS, although CLIP-based, is designed for natural images and lacks rotation-aware mechanisms and explicit multi-scale spatial reasoning, leading to coarse masks under diverse orientations or crowded layouts. CD2FSAN addresses these issues with three specialized designs. DFSM filters intermediate visual features according to the referring expression to suppress irrelevant background, the MAAM strengthens cross-scale correspondence, and the DRCD improves orientation-aware decoding.

As illustrated in Figure 7, these components enable higher-fidelity segmentation of small and rotated objects and cleaner boundaries in complex scenes, improving semantic grounding, spatial precision, and generalization to real-world remote sensing data.

Figure 7 presents qualitative results on the test set in three representative scenarios: (1) salient and visually dominant targets, (2) small and spatially compact objects, and (3) rotated objects, which reflect common challenges in referring remote-sensing segmentation and isolate the contribution of each module. (1) Without the Dynamic Feature Selection mechanism based on cross-modal information maximization, baselines can roughly localize salient targets but exhibit diffuse attention and unclear boundaries, indicating imprecise visual grounding. DFSM selects intermediate visual features that align with the referring expression, sharpening attention and boundaries and yielding more accurate masks. In scenes with small objects or many similar candidates, its effect is less consistent, but it still narrows attention to semantically plausible regions and provides a strong basis for later refinement. (2) For small or fine-grained targets, accuracy is limited by downsampling and loss of spatial detail. The Multi-scale Aggregation and Alignment Module (MAAM) aggregates features across resolutions and aligns them with hierarchical attention, increasing sensitivity to subtle structures and boundaries. This improves segmentation of small buildings, aircraft, and scattered infrastructure and helps concentrate attention when many similar objects co-occur. (3) With rotated instances the spatial reasoning is critical.

 doi:10.20944/preprints202509.1502.v1

The Dynamic Rotated Correction Decoder (DRCD) introduces rotation-aware context modeling. It uses adaptive receptive fields to focus on the correctly oriented target, suppress background noise, and produce sharper masks with cleaner boundaries, especially in dense or overlapping layouts. Overall, these examples show each added component yields more focused attention, crisper boundaries, and stronger spatial-linguistic grounding, supporting the generalization and practical effectiveness of CD2FSAN.

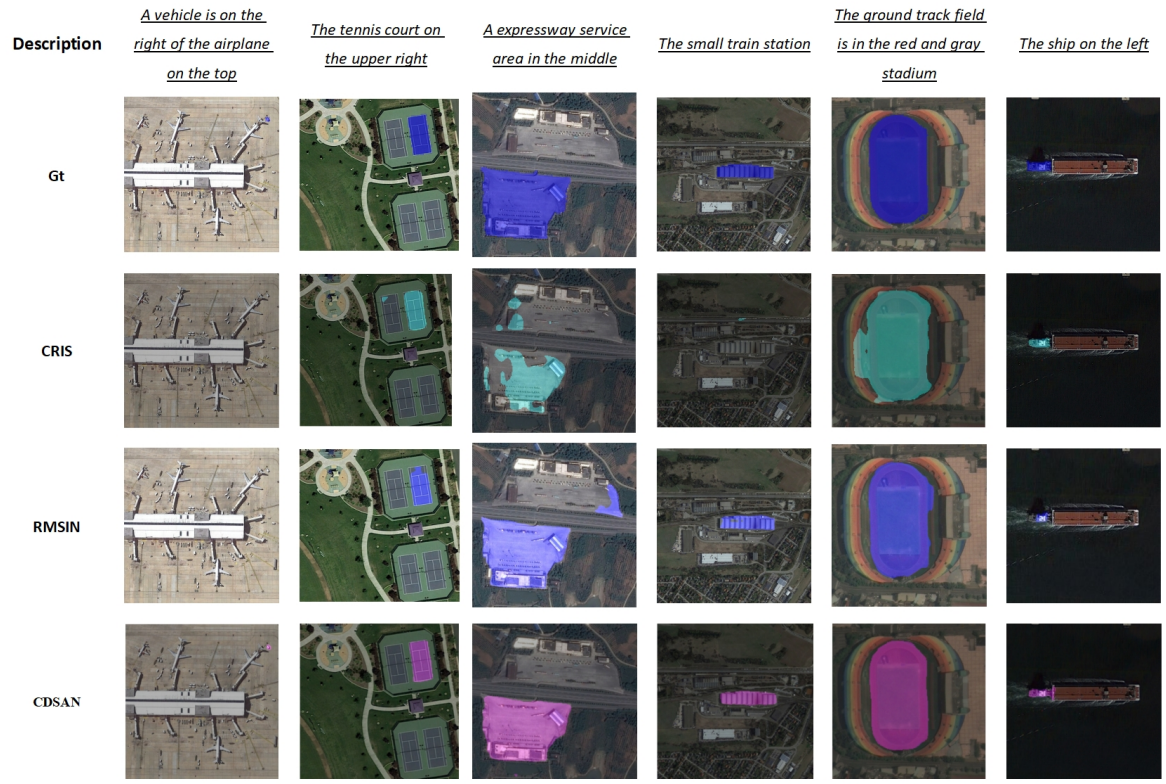


Figure 6. Visualization results of our CD2FSAN and the model RMSIN and CRIS. Our model is able to predict more accurate masks.

Despite the strong performance of the CD2FSAN model across various datasets, there are still certain failure cases that reveal its limitations. Figure 8 illustrates these failure cases, where the model's segmentation results are compared to the ground truth annotations. For instance, in the first case, the phrase "An overpass is on the upper right of the train station on the upper left" caused confusion due to the similarity between the overpass and the surrounding urban features. The model struggled to distinguish the target from the complex, texture-rich background, which highlights the challenge of segmenting targets in densely structured urban environments. A potential solution to this issue is enhancing context awareness and multi-scale feature fusion to better differentiate targets from similar background elements. In another case, the problem stems from annotation and instruction ambiguity rather than model deficiency. The expression "a black vehicle" is underspecified because several black vehicles are present, yet the ground truth marks only one instance without disambiguating cues. Under such conditions, the correct system behavior is to flag the instruction as ambiguous or abstain from a unique prediction, rather than returning an arbitrary mask. Our current pipeline selects the highest confidence candidate, which is then penalized by the single-instance annotation. Future work will incorporate an ambiguity detector and a no-unique-referent option, and we recommend dataset revisions that tag ambiguous expressions or provide multi-instance masks. Additionally, the phrase "The vehicle is at the bottom on the left" resulted in the model segmenting the wrong vehicle. The misunderstanding arose because the model interpreted the phrase as "left bottom" instead of the intended "bottom left," demonstrating the model's limitation in accurately processing complex spatial

 doi:10.20944/preprints202509.1502.v1

relationships and multiple positional references. A possible improvement would be to incorporate graph-based models or enhanced spatial reasoning techniques to better understand and differentiate spatial positions in referring expressions. Lastly, in the case of the phrase "A green and brown golf field," the model segmented the golf field, but due to inaccurate ground truth annotations, the segmentation was suboptimal. The significant surface variation of the golf field led to inconsistent labeling, which affected the model's ability to accurately segment the target. This issue stems from the dataset's labeling inconsistencies, especially for objects with varying surface textures. Future improvements should focus on refining and standardizing the annotations, particularly for objects with heterogeneous surface features, to ensure more accurate segmentation.

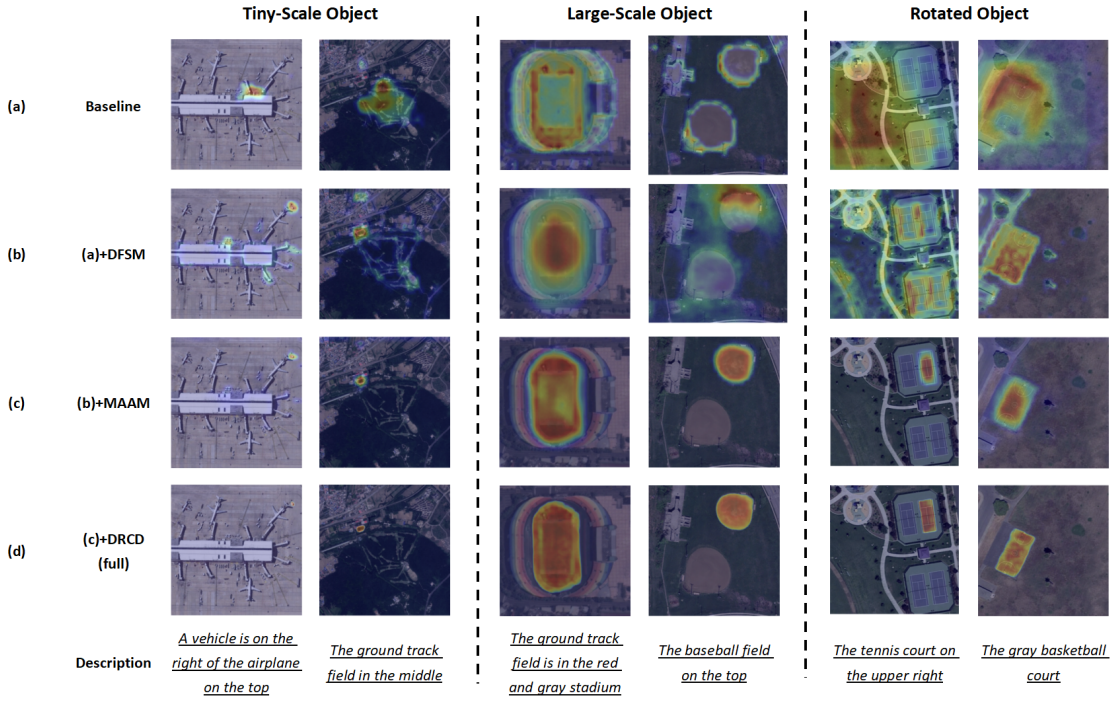


Figure 7. Qualitative examples of the proposed components. (a) The baseline model. (b) The baseline model with DFSM. (d) The baseline model with DFSM+MAAM. (e) The CD2FSAN (with DFSM+MAAM+DRCD).

 doi:10.20944/preprints202509.1502.v1

Figure 8. Failure cases of CD2FSAN in remote sensing image segmentation.

4.5. Efficiency and Complexity Analysis

To evaluate the computational efficiency of CD2FSAN, we report floating-point operations (FLOPs), parameter counts (Params), and throughput (FPS) for leading methods on RRSIS-D. Table 5 reports model complexity, FPS, and the associated oIoU, and Figure 9 depicts the accuracy–efficiency trade-off (x-axis: FPS; y-axis: oIoU), with bubble radius proportional to FLOPs and bubble colour encoding Params. CD2FSAN lies in the top-right region, delivering the fastest runtime (14.96 FPS) and the highest oIoU (78.43%) among peers. In terms of model size, it remains mid-range. The main computational cost arises from CLIP-based visual-linguistic similarity computation and the decoder's rotated correction. Despite performing multimodal interactions in both encoder and decoder, our cross-modal alignment introduces far fewer parameters than the multi-stage cross-attention stacks in CrossVLT and LAVT, enabling a moderate parameter budget while achieving state-of-the-art accuracy and speed.

Table 5. Model complexity and efficiency comparison on the RRSIS-D set.

Method	Params (M)	FLOPs (G)	oIoU (%)	FPS
LAVT	203.70	193.97	76.16	14.10
CrossVLT	212.68	200.39	75.48	14.04
RefSegFormer	195.00	**103.64**	76.39	11.95
RMSIN	196.65	149.51	77.79	13.71
MPBF	**159.12**	146.94	78.20	11.82
CD2FSAN (Ours)	197.54	148.20	**78.43**	**14.96**

 doi:10.20944/preprints202509.1502.v1

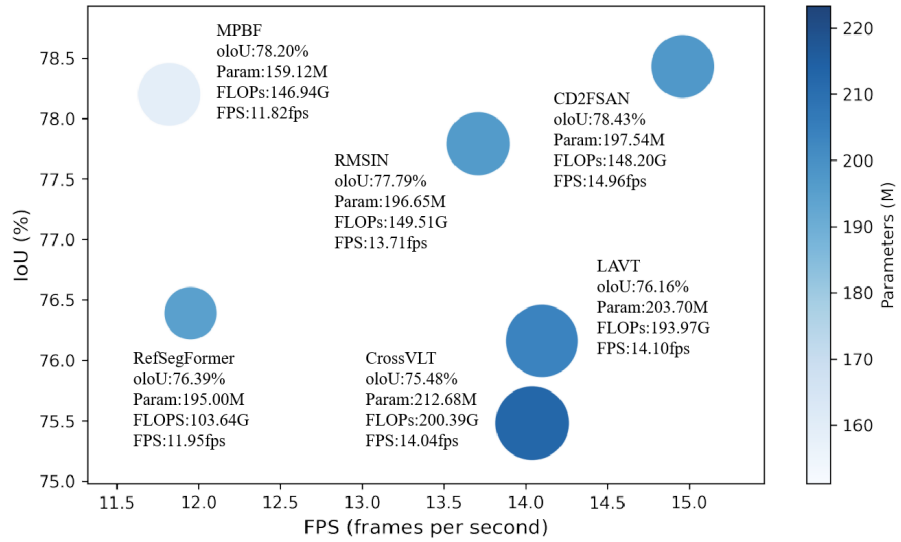


Figure 9. Efficiency–accuracy trade-off on RRSIS-D: models are positioned by FPS (abscissa) and oIoU (ordinate); bubble area scales with FLOPs and colour maps to parameter count.

5. Conclusion

This paper presents CD2FSAN, a novel framework for Referring Remote Sensing Image Segmentation (RRSIS) that improves segmentation accuracy. CD2FSAN comprises three key components: the Dynamic Feature Selection based on the principle of maximizing cross-modal information, which adaptively highlights the most informative cross-modal cues; the Multi-Scale Aggregation and Alignment Module (MAAM), which bridges semantic gaps across resolutions to better detect small and scale-varying targets; and the Dynamic Rotated Correction Decoder (DRCD), which refines segmentation masks for arbitrarily oriented objects. Extensive experiments show that CD2FSAN achieves competitive or superior performance to state-of-the-art methods across multiple benchmarks, particularly excelling with rotated and small objects. Ablation studies validate the contribution of each module, and visualizations illustrate their complementary roles in enhancing feature integration and boundary precision. Looking forward, we aim to extend CD2FSAN to high-precision localization tasks, such as fine-grained object detection under strict thresholds like Precision@0.8 and @0.9. We also plan to incorporate more advanced vision-language foundation models to enrich pixel-level visual features and capture deeper linguistic context, further strengthening cross-modal learning. These directions will enhance CD2FSAN's robustness and real-world applicability.

Author Contributions: Conceptualization, Qianqi Lu and Yuxiang Xie; methodology, Qianqi Lu and Jing Zhang; software, Qianqi Lu; validation, Qianqi Lu; formal analysis, Qianqi Lu; resources, Yuxiang Xie; data curation, Qianqi Lu; writing—original draft preparation, Qianqi Lu; writing—review and editing, Qianqi Lu, Yuxiang Xie, Jing Zhang, Yanming Guo, Yingmei Wei, Jie Jiang and Xidao Luan; supervision, Yuxiang Xie; project administration, Yuxiang Xie; funding acquisition, Yuxiang Xie. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Provincial Natural Science Foundation of Hunan grants 2023JJ30082.

Data Availability Statement: The datasets analyzed in this study are publicly available from third-party sources: RRSIS-D at https://drive.google.com/drive/folders/1Xqi3Am2Vgm4a5tHqiV9tfaqKNovcuK3A (accessed on 15 September 2025); RefSegRS at https://huggingface.co/datasets/JessicaYuan/RefSegRS (accessed on 15 September 2025); and RISBench at https://github.com/HIT-SIRS/CroBIM (accessed on 15 September 2025). No new data were created in this work. The code used to reproduce the experiments will be made publicly available upon acceptance at https://github.com/luqianqi/CD2FSAN.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202509.1502.v1

References

1. Yuan, Z.; Mou, L.; Hua, Y.; Zhu, X.X. RRSIS: Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–12. https://doi.org/10.1109/TGRS.2024.3369720.
2. Liu, S.; Ma, Y.; Zhang, X.; Wang, H.; Ji, J.; Sun, X.; Ji, R. Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26648–26658. https://doi.org/10.1109/CVPR52733.2024.02517.
3. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *117*, 11–28. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2016.03.014.
4. Cheng, J.; Deng, C.; Su, Y.; An, Z.; Wang, Q. Methods and datasets on semantic segmentation for Unmanned Aerial Vehicle remote sensing images: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* **2024**, *211*, 1–34. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2024.03.012.
5. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications* **2021**, *169*, 114417. https://doi.org/https://doi.org/10.1016/j.eswa.2020.114417.
6. Duan, L.; Lafarge, F. Towards Large-Scale City Reconstruction from Satellites. In Proceedings of the Computer Vision – ECCV 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds., Cham, 2016; pp. 89–104.
7. Abid, S.K.; Chan, S.W.; Sulaiman, N.; Bhatti, U.; Nazir, U. Present and Future of Artificial Intelligence in Disaster Management. In Proceedings of the 2023 International Conference on Engineering Management of Communication and Technology (EMCTECH), 2023, pp. 1–7. https://doi.org/10.1109/EMCTECH58502.2023.10296991.
8. Zhao, W.; Lyu, R.; Zhang, J.; Pang, J.; Zhang, J. A fast hybrid approach for continuous land cover change monitoring and semantic segmentation using satellite time series. *International Journal of Applied Earth Observation and Geoinformation* **2024**, *134*, 104222. https://doi.org/https://doi.org/10.1016/j.jag.2024.104222.
9. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment* **2020**, *236*, 111402. https://doi.org/https://doi.org/10.1016/j.rse.2019.111402.
10. Ji, R.; Tan, K.; Wang, X.; Tang, S.; Sun, J.; Niu, C.; Pan, C. PatchOut: A novel patch-free approach based on a transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images. *International Journal of Applied Earth Observation and Geoinformation* **2025**, *138*, 104457. https://doi.org/https://doi.org/10.1016/j.jag.2025.104457.
11. Ding, H.; Liu, C.; Wang, S.; Jiang, X. Vision-Language Transformer and Query Generation for Referring Segmentation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, pp. 16321–16330.
12. Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; Berg, T.L. MAttNet: Modular Attention Network for Referring Expression Comprehension. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
13. Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; Torr, P.H. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18134–18144. https://doi.org/10.1109/CVPR52688.2022.01762.
14. Liu, S.; Ma, Y.; Zhang, X.; Wang, H.; Ji, J.; Sun, X.; Ji, R. Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26648–26658. https://doi.org/10.1109/CVPR52733.2024.02517.
15. Jing, Y.; Kong, T.; Wang, W.; Wang, L.; Li, L.; Tan, T. Locate then Segment: A Strong Pipeline for Referring Image Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9853–9862. https://doi.org/10.1109/CVPR46437.2021.00973.
16. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning; Meila, M.; Zhang, T., Eds. PMLR, 18–24 Jul 2021, Vol. 139, *Proceedings of Machine Learning Research*, pp. 8748–8763.
17. Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; Liu, T. CRIS: CLIP-Driven Referring Image Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11676–11685. https://doi.org/10.1109/CVPR52688.2022.01139.
18. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; Murphy, K. Generation and Comprehension of Unambiguous Object Descriptions. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 11–20. https://doi.org/10.1109/CVPR.2016.9.

 doi:10.20944/preprints202509.1502.v1

19. Hu, R.; Rohrbach, M.; Darrell, T. Segmentation from Natural Language Expressions. In Proceedings of the Computer Vision – ECCV 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds., Cham, 2016; pp. 108–124.
20. Li, R.; Li, K.; Kuo, Y.C.; Shu, M.; Qi, X.; Shen, X.; Jia, J. Referring Image Segmentation via Recurrent Refinement Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5745–5753. https://doi.org/10.1109/CVPR.2018.00602.
21. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10494–10503. https://doi.org/10.1109/CVPR.2019.01075.
22. Hu, Z.; Feng, G.; Sun, J.; Zhang, L.; Lu, H. Bi-Directional Relationship Inferring Network for Referring Image Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4423–4432. https://doi.org/10.1109/CVPR42600.2020.00448.
23. Shi, H.; Li, H.; Meng, F.; Wu, Q. Key-Word-Aware Network for Referring Expression Image Segmentation. In Proceedings of the Computer Vision – ECCV 2018; Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y., Eds., Cham, 2018; pp. 38–54.
24. Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; Carion, N. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1760–1770. https://doi.org/10.1109/ICCV48922.2021.00180.
25. Ding, H.; Liu, C.; Wang, S.; Jiang, X. VLT: Vision-Language Transformer and Query Generation for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 7900–7916.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.
27. Kim, N.; Kim, D.; Kwak, S.; Lan, C.; Zeng, W. ReSTR: Convolution-free Referring Image Segmentation Using Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18124–18133. https://doi.org/10.1109/CVPR52688.2022.01761.
28. Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Kumar Satzoda, R.; Mahadevan, V.; Manmatha, R. PolyFormer: Referring Image Segmentation as Sequential Polygon Generation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18653–18663. https://doi.org/10.1109/CVPR52729.2023.01789.
29. Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; Ji, R. SeqTR: A Simple Yet Universal Network for Visual Grounding. In Proceedings of the Computer Vision – ECCV 2022; Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G.M.; Hassner, T., Eds., Cham, 2022; pp. 598–615.
30. Liu, C.; Ding, H.; Jiang, X. GRES: Generalized Referring Expression Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23592–23601. https://doi.org/10.1109/CVPR52729.2023.02259.
31. Quan, W.; Deng, P.; Wang, K.; Yan, D.M. CGFormer: ViT-Based Network for Identifying Computer-Generated Images With Token Labeling. *IEEE Transactions on Information Forensics and Security* **2024**, *19*, 235–250. https://doi.org/10.1109/TIFS.2023.3322083.
32. Yuan, Z.; Mou, L.; Hua, Y.; Zhu, X.X. RRSIS: Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–12. https://doi.org/10.1109/TGRS.2024.3369720.
33. Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; et al. Towards Open Vocabulary Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 5092–5113. https://doi.org/10.1109/TPAMI.2024.3361862.
34. Dong, Z.; Sun, Y.; Liu, T.; Zuo, W.; Gu, Y. Cross-Modal Bidirectional Interaction Model for Referring Remote Sensing Image Segmentation, 2025, [arXiv:cs.CV/2410.08613].
35. Lei, S.; Xiao, X.; Zhang, T.; Li, H.C.; Shi, Z.; Zhu, Q. Exploring Fine-Grained Image-Text Alignment for Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–11. https://doi.org/10.1109/TGRS.2024.3522293.
36. Shi, L.; Zhang, J. Multimodal-Aware Fusion Network for Referring Remote Sensing Image Segmentation. *IEEE Geoscience and Remote Sensing Letters* **2025**, *22*, 1–5. https://doi.org/10.1109/LGRS.2025.3527485.
37. Chen, K.; Zhang, J.; Liu, C.; Zou, Z.; Shi, Z. RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models, 2025, [arXiv:cs.CV/2501.06809].
38. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information

 doi:10.20944/preprints202509.1502.v1

Processing Systems; Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; Vaughan, J.W., Eds. Curran Associates, Inc., 2021, Vol. 34, pp. 12077–12090.

39. Li, Y.; Li, Z.Y.; Zeng, Q.; Hou, Q.; Cheng, M.M. Cascade-CLIP: cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML'24.
40. Cho, Y.; Yu, H.; Kang, S.J. Cross-Aware Early Fusion With Stage-Divided Vision and Language Transformer Encoders for Referring Image Segmentation. *IEEE Transactions on Multimedia* **2024**, *26*, 5823–5833. https://doi.org/10.1109/TMM.2023.3340062.
41. Li, R.; Li, K.; Kuo, Y.C.; Shu, M.; Qi, X.; Shen, X.; Jia, J. Referring Image Segmentation via Recurrent Refinement Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5745–5753. https://doi.org/10.1109/CVPR.2018.00602.
42. Hu, Z.; Feng, G.; Sun, J.; Zhang, L.; Lu, H. Bi-Directional Relationship Inferring Network for Referring Image Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4423–4432. https://doi.org/10.1109/CVPR42600.2020.00448.
43. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation . In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 2019; pp. 10494–10503. https://doi.org/10.1109/CVPR.2019.01075.