# Entity-Guided Attention Twisting Network for Referring Remote Sensing Image Segmentation

Yuyu Jia, Qing Zhou, Junyu Gao, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Referring Remote Sensing Image Segmentation (RRSIS) aims to establish pixel-level interpretation of specific regions queried by textual expressions, bridging textual semantics and intelligent analysis of remote sensing imagery. In contrast to natural scenarios, the intricate backgrounds in remote sensing scenarios result in low target-background contrast, often leading to semantic dispersion in segmented regions. Furthermore, conventional cross-attention-based referring image segmentation (RIS) methods struggle to bridge the modal gap, hindering fine-grained alignment between linguistic descriptions and geographical features. To overcome these challenges, we present a pioneering Entity-Guided Attention Twisting Network (Enti-TwistNet) for RRSIS. Our framework first introduces a SAM-inspired Entity Guidance (SEG) module that extracts spatially constrained entity prompts through a self-reasoning mask generation mechanism, constructing a comprehensive entity-visual-text tri-modal information cube. Subsequently, during cross-modal interaction, we propose a Dual-phase Attention-Twisting (DAT) mechanism: (1) initially sequential channel-wise scanning to facilitate cross-modal semantic propagation; (2) Subsequently, twist attention to the spatial dimension, integrating entity guidance to enhance the representation of irregular geographic boundaries. Extensive experiments on two widely used benchmarks, RefSegRS and RRSIS-D, demonstrate that Enti-TwistNet achieves significant performance improvements over existing state-of-the-art models.**

***Index Terms*—Remote Sensing, Referring Segmentation, Entity-aware Guidance, Attention Twisting.**


Fig. 1. The task of RRSIS faces two fundamental challenges. The lower part of the figure illustrates that the low target-background contrast in remote sensing images often leads to semantic dispersion during segmentation. The upper part highlights the significant modality gap between unstructured language descriptions and complex geospatial data.

## I. Introduction

REMOTE sensing image segmentation is a fundamental task in geospatial analysis, providing critical pixel-level information that underpins key applications such as environmental monitoring [1]–[3], urban planning [4], [5], and disaster management [6], [7]. Driven by the surge in large-scale remote sensing datasets and the transformative advances in deep learning, the field has evolved towards increasingly sophisticated segmentation paradigms [8], such as semi-supervised [9]–[11], few-shot [12]–[14], and zero-shot [15]–[17] methods—progressively advancing toward more intelligent and practical applications. However, these approaches typically aim to exhaustively segment all potential regions of interest based on predefined categories. As a result, they often fail to capture subtle semantic relationships between different regions within an image and, crucially, cannot support human-centric interpretations via natural language queries.

A more challenging task—Referring Image Segmentation (RIS)—has emerged to address these limitations. Guided by user-provided textual descriptions, RIS aims to localize accurately and segment specific target elements within a visual scene. Leveraging the powerful global modeling capabilities of attention mechanisms, recent advances in RIS have centered around unidirectional [18], [19] and bidirectional [20], [21] cross-attention structures to align the logical structure of text with the spatial relationships of visual objects. Despite remarkable progress in the natural image domain, directly transferring these techniques to Remote Sensing Referring Image Segmentation (RRSIS) remains suboptimal.

*From the perspective of inherent image characteristics,* remote sensing imagery fundamentally differs from natural images, which typically feature salient foreground objects against relatively clean backgrounds. Remote sensing scenes often exhibit low contrast between targets and background due to spectral similarity among land-cover types and complex background clutter. This results in ***semantic dispersion***,

This work was supported in part by the National Natural Science Foundation of China under Grant 62471394, and Grant U21B2041.

Yuyu Jia, Qing Zhou, Junyu Gao, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

E-mail: jyy2019@mail.nwpu.edu.cn, chautsing@gmail.com, gjy3035@gmail.com, crabwq@gmail.com.

Qi Wang is the corresponding author.

where segmented regions may include irrelevant background or adjacent features, and such dispersed semantics can easily disrupt the coarse-to-fine attention focus that natural-image RIS models rely on, causing target misidentification in RS scenes. *From a methodological standpoint,* most RIS frameworks rely on cross-attention mechanisms that flatten cross-modal interactions into holistic feature-level matching. However, they are often insufficient to bridge the ***significant modality gap*** between unstructured language and complex geospatial patterns, leading to reduced accuracy in both target localization and pixel-level interpretation.

To address the dual challenges of semantic dispersion and modality gaps in RRSIS, we propose two key insights. First, to alleviate semantic dispersion, we incorporate explicit spatial structural priors before multimodal feature fusion. Specifically, we are the first to integrate the Segment Anything Model (SAM) [22] into RRSIS in an entity-guidance paradigm—rather than using it merely as a segmentation decoder or fixed prompts. We further design a self-reasoning process that aggregates SAM's fragmented masks by jointly considering geometric overlap, appearance similarity, and text relevance, producing coherent and semantically aligned entity guidance that constrains the subsequent cross-modal interaction. Second, to overcome the limitations of conventional cross-attention under large modality gaps, we design a hierarchical cross-modal interaction strategy that decouples semantic alignment from spatial dependency modeling, enabling better alignment between unstructured language and complex geospatial patterns.

In this paper, we propose Enti-TwistNet, a novel framework that integrates the above insights. They are instantiated as a SAM-inspired Entity Guidance (SEG) module and a Dual-phase Attention-Twisting (DAT) mechanism. As illustrated in Fig. 2, both SEG and DAT are plug-and-play modules that can be flexibly integrated into multiple stages of a feature extractor to handle the multi-scale remote sensing targets. Specifically, SEG transforms multiple SAM-generated masks into spatially rich, entity-aware prompts. Given SAM's tendency to over-segment complete objects into fragmented masks, SEG applies a self-reasoning process to infer inter-prompt relationships, refining them into coherent and robust entity-guided cues. These cues are then combined with visual and textual embeddings to form a comprehensive tri-modal feature cube through channel-wise concatenation. Subsequently, DAT performs hierarchical cross-modal interaction through a dual-phase attention mechanism: (i) Cross-modal semantic propagation is realized via sequential scanning across the channel dimension, exploiting the linear complexity of the Mamba [23] model; (ii) Attention is then twisted toward the entity-aware spatial dimension to model fine-grained positional dependencies across modalities. While conceptually related to prior Mamba-based twisting methods such as ReMamber [24], our DAT differs in both design and modality: it operates on a *tri-modal* feature cube (visual–text–entity) rather than a dual-modality one, and replaces the second spatial Mamba scan with *entity-aware cross-attention* to explicitly integrate spatial priors with visual–textual features. This design is tailored to RRSIS, enabling robust, query-conditioned localization while bridging the large modality gap. Our main contributions are summarized as follows:

To summarize, the key contributions are as follows:

1) We present Enti-TwistNet, a novel framework addressing two core challenges in RRSIS: semantic dispersion from low target–background contrast, and the modality gap between language and dense geospatial features.
2) We design an entity-aware guidance strategy in SEG to impose spatial constraints on cross-modal interactions for semantic consistency, and a dual-phase mechanism enabling hierarchical cross-modal interaction by separating semantic alignment from spatial relation modeling.
3) Experiments on three public benchmarks show that Enti-TwistNet outperforms state-of-the-art methods, validating the effectiveness of our entity-guided and hierarchical interaction designs.

## II. Related Works

### A. Referring Image Segmentation in Natural Images

Referring Image Segmentation (RIS) aims to localize and segment image regions described by a natural language query, enabling fine-grained vision–language interaction [25].

Benefiting from attention mechanisms, cross-modal attention has become the core of RIS [26]–[28], allowing models to adaptively highlight relevant visual regions based on textual cues and vice versa. Representative methods include VLT [18], which frames RIS as a multi-head attention problem with query generation and balancing modules. LAVT [20], which performs early fusion of linguistic and visual features in Swin Transformer. ReMamber [24], which employs Mamba's linear computation with a channel–spatial twisting mechanism for deep cross-modal fusion.

### B. Visual Grounding in Remote Sensing

Remote Sensing Visual Grounding (RSVG) involves localizing and identifying regions in an image corresponding to a given natural language description [29]. Zhan *et al.* [30] construct the large-scale DIOR-RSVG benchmark and integrate multi-scale visual features with multi-granularity textual embeddings to handle scale variations and background interference adaptively. LPVA [31] employs a language-guided progressive visual attention mechanism to adjust visual features during the feature extraction dynamically, mitigating attention drift and improving the localization accuracy of referential expressions. While RSVG shares similarities with RIS, it is generally less granular, typically emphasizing bounding boxes rather than pixel-level segmentation.

### C. Referring Remote Sensing Image Segmentation

Referring Remote Sensing Image Segmentation (RRSIS) extends the RIS task to remote sensing, where models must segment regions described by text queries within the more complex, large-scale, and noisy context of remote sensing imagery. Yuan et al. [32] pioneered RRSIS by establishing the RefSegRS dataset and designing a language-guided cross-scale enhancement module to better segment small and scattered

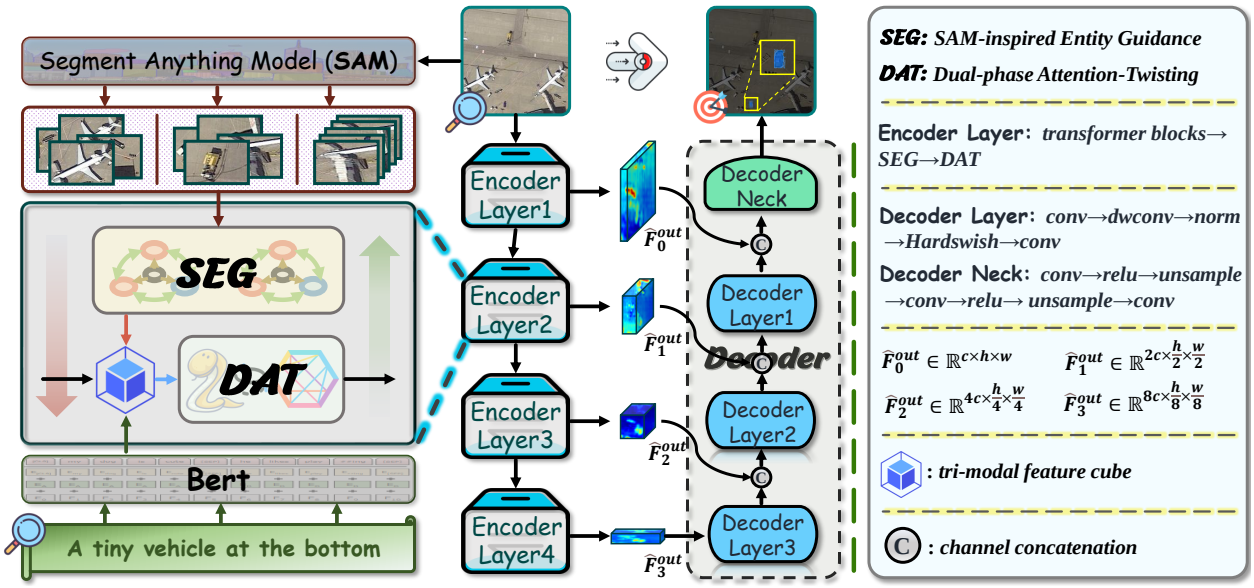



Fig. 2. The overall pipeline of the proposed Enti-TwistNet. Our core contributions are encapsulated within each multi-stage encoder, comprising two key designs: **(a) SAM-inspired Entity Guidance (SEG)** module, which introduces coherent and robust entity-guided cues via a self-reasoning mechanism, encouraging the model to attend to semantically consistent entity regions, thereby mitigating semantic dispersion. **(b) Dual-phase Attention-Twisting (DAT)** mechanism, which hierarchically alleviates the modality gap by first employing Mamba-based channel-wise sequence scanning to enable cross-modal semantic propagation, followed by cross-attention to model spatial positional dependencies.

objects. RMSIN [33] tackles complex spatial scales and orientations via intra-/cross-scale interaction and adaptive rotational convolution, while CroBIM [34] improves multi-scale fusion through context-aware prompts and language-guided aggregation. Beyond task-specific models, vision–language foundation models such as GeoGround [35] have been explored for remote sensing, leveraging large-scale multimodal pre-training for several downstream tasks. Although adaptable to RRSIS without task-specific training, GeoGround's general-purpose design and large model size limit its efficiency and fine-grained adaptability—motivating the development of lightweight, effective architectures such as our proposed Enti-TwistNet.

## III. Methodology

### A. Architecture Overview

The core challenge of Referring Remote Sensing Image Segmentation (RRSIS) lies in achieving fine-grained alignment between the natural language expression $E = \{e_i | i \in 0, \ldots, N\}$ and the queried image $I \in \mathbb{R}^{H \times W \times 3}$, where $N$ denotes the length of the description, while $H$ and $W$ represent the height and width of the image, respectively. The queried image $I$ is progressively transformed into visual embeddings $F_t^{in}$ by a hierarchical encoder $\mathcal{E}_t$ composed of four stages, where $t \in 0, 1, 2, 3$. As highlighted in Fig. 1, the RRSIS task is fundamentally challenged by *semantic dispersion*—arising from low target–background contrast, and a pronounced *modality gap* between unstructured language and complex geospatial patterns.

To address this, we propose Enti-TwistNet, a multi-stage framework that embeds two plug-and-play modules into each encoder stage: the SAM-inspired Entity Guidance (SEG) and the Dual-phase Attention-Twisting (DAT) mechanism. SEG (Sec.III-B) derives entity cues via self-reasoning and fuses them with visual–textual embeddings to form a tri-modal feature cube, enforcing semantic consistency in vision–language interactions. DAT (Sec.III-D) performs hierarchical cross-modal interaction: a linear-complexity Mamba block propagates semantics along channels, followed by spatial attention twisted under entity guidance to capture fine-grained positional dependencies. Multi-scale interactive features from all stages are finally fed into the segmentation decoder.

### B. SAM-inspired Entity Guidance

In remote sensing scenarios, the high spectral similarity of land-cover types and complex background clutter often result in low target–background contrast, leading to ***semantic dispersion*** where predictions include irrelevant or adjacent regions. Existing SAM-based prompting methods—using SAM as either a direct decoder or fixed prompts—are suboptimal for RRSIS: trained on natural scenes, SAM tends to over-segment geospatial objects, while fixed prompts overlook cross-modal gaps. In this work, we are the first to adapt SAM for RRSIS as an *entity-guidance* mechanism, treating multiple SAM proposals as *entity priors* and refining them via self-reasoning to yield coherent, text-aligned spatial cues for robust cross-modal interactions.

*1) Entity Priors Extraction:* At the encoder stage $t$, after extracting visual features $F_t^{in}$ from the encoder $\mathcal{E}_t$, we begin by feeding the queried image $I$ into SAM to generate an initial mask set $\{m_k\}_{k=1}^{K}$, where $m_k \in \{0, 1\}^{H \times W}$. We omit $t$ for brevity. Following GPRN [36], masks are sorted by area, and overlapping pixels are assigned to the smallest mask, ensuring uniqueness. Mask average pooling (MAP) converts each mask into a compact entity prior:

$$p_k = \mathcal{MAP}(F^{in}, m_k) \in \mathbb{R}^{1 \times d}, \quad k = 1, \ldots, K, \qquad (1)$$

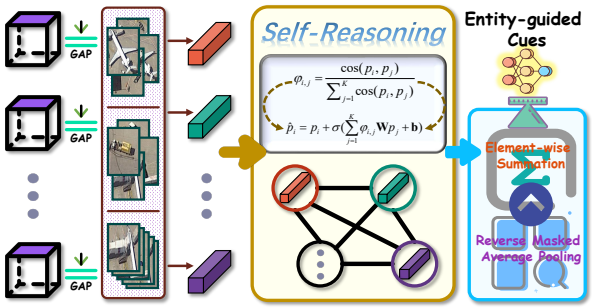



Fig. 3. Configuration diagram of the SEG module.

where $\mathcal{MAP}(\cdot)$ denotes the mask average pooling operator that extracts the mean-pooled feature vector from $F^{in}$ within the binary mask $m_k \in \{0,1\}^{H\times W}$, and $d$ is the feature channel dimension.

*2) Self-reasoning Refinement:* Due to SAM's tendency to fragment entities in remote sensing scenes, using these entity priors generated by raw masks risks introducing noise and inconsistencies into subsequent multimodal fusion. We address this via a self-reasoning mechanism that models spatial and semantic relations among priors to aggregate a more coherent and robust set of entity-aware guidance. For entity prior $p_i$, its relational affinity with $p_j$ is:

$$\varphi_{i,j} = \frac{\cos(p_i, p_j)}{\sum_{l=1}^{K} \cos(p_i, p_l)}, \quad i,j \in \{1,\dots,K\}, \tag{2}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity function, and higher values of $\varphi_{i,j}$ correspond to stronger relational affinity between $p_i$ and $p_j$. Based on the computed affinity, we perform a weighted aggregation to facilitate information exchange across entity priors:

$$\hat{p}_i = p_i + \sigma\Big(\sum_{j=1}^{K} \varphi_{i,j} \boldsymbol{W} p_j + \boldsymbol{b}\Big), \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{b} \in \mathbb{R}^{d}$ are learnable weights and bias, applied through a single-layer linear transformation. Since the encoder stages produce features with different channel dimensions $d$, $(\mathbf{W}, \mathbf{b})$ are *not* shared across stages; instead, each stage has its own independently initialized and trained lightweight linear layer. Here, $\sigma(\cdot)$ denotes a non-linear activation function, *e.g.*, ReLU.

Finally, we employ the reverse process of MAP to restore the spatial information of the refined entity priors:

$$g_k = \frac{\sum_{h,w}^{\acute{H},\acute{W}} \hat{p}_k \otimes m_k(h,w)}{\sum_{h,w}^{\acute{H},\acute{W}} m_k(h,w)}, \tag{4}$$

where $k \in 1,\dots,K$, $\otimes$ stands for the element-wise multiplication, $(h,w)$ denotes the spatial coordinates, and $(\acute{H},\acute{W})$ represents the spatial resolution of the visual features $F^{in}$ at the current stage. The entity-aware guidance can be summarized as $G = \sum_{k=1}^{K} g_k \in \mathbb{R}^{\acute{H}\times\acute{W}\times d}$.

### C. Tri-modal Feature Cube Construction

In the Enti-TwistNet framework's cross-modal interaction, language expression provides semantic cues, visual features reveal image content, and entity-aware guidance (generated by the SEG module) supplies crucial spatial structural priors. To fully leverage the strengths of all three modalities, it is essential to integrate them into a unified representation space, forming a tri-modal feature cube for multimodal interaction. As previously discussed, we have already obtained the visual features $F^{in}$ and the entity-aware $G$; we consider the textual embeddings from the given language expression $E$.

Initially, the language description $E$ is encoded into an original textual embedding $T^{ori} \in \mathbb{R}^{N\times d_t}$ through a text encoder, *e.g.,* BERT [37], where $d_t$ is the channel dimension of the original textual embedding. Given that local attributes such as color, shape, and position in the language expression are crucial for referring segmentation, we propose the construction of local textual embeddings to capture their fine-grained relationships with visual features. Formally, we compute the local correlation map via matrix multiplication:

$$T^{cor} = F^{in}\boldsymbol{W}_0^{cor} \cdot (T^{ori}\boldsymbol{W}_1^{cor})^{\top} \in \mathbb{R}^{\acute{H}\times\acute{W}\times N}, \tag{5}$$

where $\boldsymbol{W}_0^{cor} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{W}_1^{cor} \in \mathbb{R}^{d_t\times d}$ are learnable parameters specific to the current stage, given the stage-dependent channel dimension $d$. To facilitate high-dimensional feature processing while aligning with the channel dimensions of the visual features, a single convolutional layer is employed to project $T^{cor}$ into $T^{loc} \in \mathbb{R}^{\acute{H}\times\acute{W}\times d}$.

Ultimately, the tri-modal feature cube is constructed via channel-wise concatenation, denoted as:

$$F^{cube} = [F^{in} + G; T^{loc}] \in \mathbb{R}^{\acute{H}\times\acute{W}\times 2d}. \tag{6}$$

It aggregates information across different modalities and provides a unified and enriched feature representation for the subsequent DAT mechanism.

### D. Dual-phase Attention-Twisting

The tri-modal feature cube $F_{cube}$ integrates visual, textual, and entity-aware guidance. Effectively bridging the ***significant modality gap*** between unstructured linguistic queries and dense geospatial features, while leveraging entity priors for precise localization, necessitates a hierarchical interaction strategy.

Traditional cross-attention mechanisms typically perform 'flattened' global feature-level matching, which overlooks the inherent heterogeneity between modalities. Such coarse interactions are insufficient for the fine-grained alignment required in RRSIS tasks. To address this, DAT decomposes cross-modal interaction into two sequential phases: (i) *cross-modal semantic propagation* along the channel dimension, mitigating modality discrepancies; and (ii) *entity-aware spatial attention twisting*, explicitly aligning positional and relational dependencies across the tri-modal feature set, thereby providing superior feature representations for segmentation.

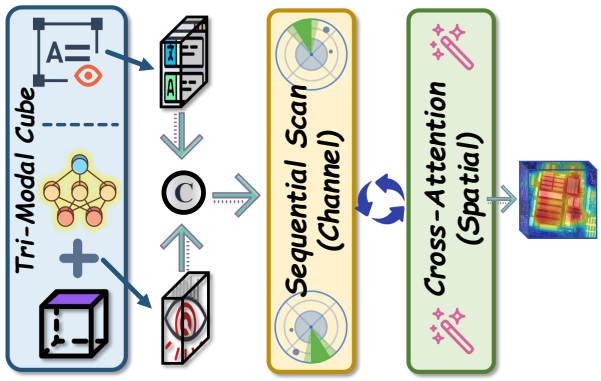



Fig. 4. Configuration diagram of the DAT module.

*1) Cross-Modal Semantic Propagation:* At each spatial location $(h, w)$ in the feature map $F^{cube}$, the feature vector $F^{cube}[h, w, :] \in \mathbb{R}^{2d}$ consists of concatenated visual, textual, and entity-aware features. To enable these modality-specific channels to exchange semantic information in a progressive, order-aware manner, we employ the Mamba state space model (SSM) with linear complexity [23]. Unlike self-attention, which may prematurely flatten modality distinctions, Mamba's sequential channel scanning preserves context and captures long-range dependencies, leading to smoother semantic propagation and reduced cross-modal conflicts. Specifically, we perform sequential channel scanning over $F^{cube}$ to facilitate cross-modal semantic propagation, thereby mitigating modality discrepancies:

$$F^{sem} = SSM_{1D}(F^{cube}) \in \mathbb{R}^{\acute{H} \times \acute{W} \times 2d}, \tag{7}$$

where $SSM_{1D}(\cdot)$ applies a 1-D selective scan along the channel dimension at each spatial position, propagating semantics across the concatenated tri-modal feature channels.

*2) Entity-Aware Spatial Attention Twist:* In contrast to ReMamber's second-phase spatial Mamba scan, we adopt *entity-aware cross-attention* to more effectively integrate structured spatial priors with visual–textual features. This mechanism explicitly establishes fine-grained positional dependencies between linguistic expressions and corresponding visual regions under entity-guided constraints. Concretely, $F^{sem}$ is evenly split along the channel dimension into an entity-guided visual feature $F^{vis}$ and a text-enhanced feature $F^{txt}$:

$$F^{vis} = F^{sem}[:, :, : d], \quad F^{txt} = F^{sem}[:, :, d :], \tag{8}$$

where $F^{vis}$ contains the first $d$ channels corresponding to entity-guided visual features, and $F^{txt}$ contains the remaining channels, enriched with text semantics. We take $F^{vis}$ as the query ($\boldsymbol{Q}$) and $F^{txt}$ as the key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$) in a cross-attention operation:

$$\boldsymbol{Q} = F^{vis}\boldsymbol{W}_Q, \quad \boldsymbol{K} = F^{txt}\boldsymbol{W}_K, \quad \boldsymbol{V} = F^{txt}\boldsymbol{W}_V, \tag{9}$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ are the linear projection matrices specific to the current stage, given the stage-dependent channel dimension $d$. Finally, the language-referred image features can be obtained as:

$$F^{out} = \text{Proj}(\mathcal{CA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + F^{vis}), \tag{10}$$

where $\mathcal{CA}(\cdot)$ denotes the cross-attention operation and $\text{Proj}(\cdot)$ is the a $1 \times 1$ convolution block.

### E. Convolution-based Decoder

The four-stage encoder jointly produces a set of multi-scale language-referred image features $\{F_t^{out} | t \in \{0, 1, 2, 3\}\}$, which are fed into the decoder to generate the final referential segmentation results. Inspired by ReMamber [24], the Enti-TwistNet framework adopts a convolution-based progressive upsampling architecture comprising four residual layers. The overall top-down prediction process can be summarized as:

$$\hat{F}_0^{out} = \mathcal{R}(\mathcal{M}_0(F_0^{out})), \tag{11}$$

$$\hat{F}_1^{out} = \mathcal{R}(\mathcal{M}_1([\hat{F}_0^{out}; F_1^{out}])), \tag{12}$$

$$\hat{F}_2^{out} = \mathcal{R}(\mathcal{M}_2([\hat{F}_1^{out}; F_2^{out}])), \tag{13}$$

$$\boldsymbol{Y} = \mathcal{N}([\hat{F}_2^{out}; F_3^{out}]) \in \mathbb{R}^{H \times W \times 2}, \tag{14}$$

where $\boldsymbol{Y}$ is the final referred segmentation mask, and $\mathcal{M}_*$ and $\mathcal{N}$ are implemented via convolution, with detailed specifications shown on the right side of Fig. 2.

## IV. Experiments

### A. Dataset

*1) RefSegRS:* It [32] is constructed from images and pixel-wise annotations of the SkyScapes dataset [38], comprising 4,420 image–language–label triplets.

*2) RRSIS-D:* It [33] expands upon previous datasets, encompassing 17,402 image-caption-mask triplets. This extensive collection offers unparalleled diversity in spatial scales and object orientations, making it particularly challenging for models aiming to localize objects in complex aerial imagery.

*3) RISBench:* It [34] is a large-scale vision-language benchmark specifically designed for RRSIS. It comprises 52,472 high-quality image-language-label triplets. All images are uniformly formatted to a resolution of 512×512 pixels, with spatial resolutions ranging from 0.1 meters to 30 meters.

### B. Implementation Details

Our proposed Enti-TwistNet is implemented using the PyTorch framework [52], with all training and testing conducted on NVIDIA A6000 GPUs. All input images are resized to 480x480. The batch size is set to 8. The model undergoes training for 40 epochs using the AdamW optimizer, configured with a weight decay of 0.01 and an initial learning rate of 5e-4. The learning rate is progressively reduced following a polynomial decay schedule. We adopt a Swin Transformer [53] pre-trained on ImageNet-22K [54] for the visual encoder, which features four distinct extraction stages. The language model is a 12-layer BERT architecture, obtained from Hugging Face's library [55], providing an output channel dimension of 768. Within the SEG module, we utilize the base version of SAM [22]. Input images are upsampled to 1024x1024 for initial mask generation, an offline process. To manage computational load, a maximum of 40 initial masks are retained; any masks exceeding this limit are discarded based on their smaller size.

TABLE I
PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS ON THE RRSIS-D DATASET.

| Method | Text Encoder | Visual Encoder | $Pr@X$ | | | | | oIoU | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | | |
| GeoGround [35] | *Remote Sensing VLMs* | | 66.06 | - | - | - | - | - | 58.93 |
| NExT-Chat [39] | | | 26.37 | - | - | - | - | - | 24.98 |
| PixelLM [40] | | | 33.46 | - | - | - | - | - | 31.65 |
| RRN [41] | LSTM | ResNet-101 | 51.07 | 42.11 | 32.77 | 21.57 | 6.37 | 66.43 | 45.64 |
| LSCM [42] | | ResNet-101 | 56.02 | 46.25 | 37.70 | 25.28 | 7.86 | 69.10 | 49.92 |
| CMPC [43] | | ResNet-101 | 55.83 | 47.40 | 35.28 | 25.45 | 9.20 | 69.41 | 49.24 |
| BRINet [44] | | ResNet-101 | 56.90 | 48.77 | 38.61 | 27.03 | 8.93 | 69.68 | 49.45 |
| CMPC+ [45] | | ResNet-101 | 57.95 | 48.31 | 37.61 | 24.33 | 7.94 | 70.13 | 50.12 |
| BKINet [46] | CLIP | ResNet-101 | 56.90 | 48.77 | 39.12 | 27.03 | 9.16 | 69.89 | 49.65 |
| ETRIS [47] | | ResNet-101 | 61.07 | 50.99 | 40.94 | 29.30 | 11.43 | 71.06 | 54.21 |
| CRIS [27] | | ResNet-101 | 54.84 | 46.77 | 38.06 | 28.15 | 11.52 | 70.46 | 49.69 |
| ReMamber [24] | | Mamba-B | 76.38 | **69.25** | 55.80 | 42.13 | 22.66 | 76.90 | 64.27 |
| LGCE [32] | BERT | Swin-B | 67.65 | 61.53 | 51.42 | 39.62 | 22.94 | 76.33 | 59.37 |
| LAVT [20] | | Swin-B | 63.98 | 57.57 | 49.30 | 38.06 | 22.29 | 76.16 | 56.82 |
| RMSIN [33] | | Swin-B | 67.16 | 60.36 | 50.16 | 38.72 | 22.81 | 75.79 | 58.79 |
| CrossVLT [48] | | Swin-B | 66.42 | 59.41 | 49.76 | 38.67 | 23.30 | 75.48 | 58.48 |
| CARIS [49] | | Swin-B | 71.50 | 63.52 | 52.92 | 40.94 | 23.90 | 77.17 | 62.12 |
| CroBIM [34] | | Swin-B | 75.00 | 66.32 | 54.31 | 41.09 | 21.78 | 76.37 | 64.24 |
| FIANet [50] | | Swin-B | 74.46 | 66.96 | 56.31 | 42.83 | **24.13** | 76.91 | 64.01 |
| **Enti-TwistNet (ours)** | CLIP | Swin-B | **78.15** | **68.96** | **57.24** | **43.93** | 24.04 | **78.21** | **66.10** |
| **Enti-TwistNet (ours)** | BERT | Swin-B | **77.94** | 68.81 | **57.07** | **44.63** | **24.80** | **78.16** | **66.83** |

TABLE II
PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS ON THE REFSEGRS DATASET.

| Method | Text Encoder | Visual Encoder | $Pr@X$ | | | | | oIoU | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | | |
| RRN [41] | LSTM | ResNet-101 | 30.26 | 23.01 | 14.87 | 7.17 | 0.98 | 65.06 | 41.88 |
| LSCM [42] | | ResNet-101 | 31.54 | 20.41 | 9.51 | 5.29 | 0.84 | 61.27 | 35.54 |
| BRINet [44] | | ResNet-101 | 20.72 | 14.26 | 9.87 | 2.98 | 1.14 | 58.22 | 31.51 |
| LSTM-CNN [51] | | ResNet-101 | 15.69 | 10.57 | 5.17 | 1.10 | 0.28 | 53.83 | 24.76 |
| BKINet [46] | CLIP | ResNet-101 | 36.12 | 20.62 | 15.22 | 6.26 | 1.33 | 63.37 | 40.41 |
| ETRIS [47] | | ResNet-101 | 35.77 | 23.00 | 13.98 | 6.44 | 1.10 | 65.96 | 43.11 |
| CRIS [27] | | ResNet-101 | 35.77 | 24.11 | 14.36 | 6.38 | 1.21 | 65.87 | 43.26 |
| ReMamber [24] | | Mamba-B | 73.96 | 61.25 | 38.78 | 19.44 | 4.41 | 73.26 | 59.00 |
| LGCE [32] | BERT | Swin-B | 50.19 | 28.62 | 17.17 | 9.36 | 2.15 | 71.59 | 46.57 |
| LAVT [20] | | Swin-B | 71.44 | 57.40 | 32.14 | 15.41 | 4.51 | 76.46 | 57.74 |
| RMSIN [33] | | Swin-B | 71.60 | 55.97 | 31.87 | 11.72 | 1.93 | 71.73 | 57.78 |
| CrossVLT [48] | | Swin-B | 71.16 | 58.28 | 34.51 | 16.35 | 5.06 | 77.44 | 58.84 |
| CroBIM [34] | | Swin-B | 64.83 | 44.41 | 17.28 | 9.69 | 2.20 | 72.30 | 52.69 |
| **Enti-TwistNet (ours)** | CLIP | Swin-B | **77.92** | **69.29** | **44.82** | **26.83** | **7.58** | **78.85** | **64.02** |
| **Enti-TwistNet (ours)** | BERT | Swin-B | **77.72** | **69.30** | **45.01** | **26.66** | **8.49** | **78.98** | **63.82** |

We adopt the evaluation metrics commonly used in prior studies [32], [33], namely overall Intersection-over-Union (oIoU), mean Intersection-over-Union (mIoU), and precision at varying threshold levels $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, denoted as $Pr@X$.

### C. Experimental Results and Comparisons

Tables I–III compare Enti-TwistNet with state-of-the-art methods. The best and second-best results are marked in red and blue, respectively. LSTM-based approaches generally underperform due to their sequential processing, which limits modeling of complex spatial–semantic dependencies. Methods specifically designed for remote sensing imagery, such as LGCE [32], RMSIN [33], CroBIM [34], and FIANet [50], achieve better results than those adapted from natural-image RIS. Regardless of encoder choice (CLIP or BERT), Enti-TwistNet delivers consistent and substantial gains, indicating strong robustness and generalization. On RRSIS-D, the CLIP-based Enti-TwistNet outperforms CroBIM by 2.59% mIoU and surpasses Robust-RefSeg by 0.81% oIoU. On RefSegRS, it further achieves a 6.01% improvement in $Pr@0.90$, underscoring its fine-grained referential segmentation capability. On the RISBench dataset, Enti-TwistNet similarly maintains leading performance, achieving an average 0.32% mIoU improvement over CroBIM, which demonstrates its adaptability to diverse remote sensing scenarios and varying text query complexities.

TABLE III
PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS ON THE RISBENCH DATASET.

| Method | Text Encoder | Visual Encoder | $Pr@X$ | | | | | oIoU | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | | |
| RRN [41] | LSTM | ResNet-101 | 55.04 | 47.31 | 39.86 | 32.58 | 13.24 | 49.67 | 43.18 |
| LSCM [42] | | ResNet-101 | 55.26 | 47.14 | 40.10 | 33.29 | 13.91 | 50.08 | 43.69 |
| BRINet [44] | | ResNet-101 | 52.87 | 45.39 | 38.64 | 30.79 | 11.86 | 48.73 | 42.91 |
| ETRIS [47] | CLIP | ResNet-101 | 60.98 | 51.88 | 39.87 | 24.49 | 11.18 | 67.61 | 53.06 |
| CRIS [27] | | ResNet-101 | 63.67 | 55.73 | 44.42 | 28.80 | 13.27 | 69.11 | 55.18 |
| LGCE [32] | BERT | Swin-B | 69.94 | 64.07 | 56.26 | 44.92 | 25.74 | 73.87 | 62.13 |
| LAVT [20] | | Swin-B | 69.40 | 63.66 | 56.10 | 44.95 | 25.21 | 74.15 | 61.93 |
| RMSIN [33] | | Swin-B | 71.01 | 65.46 | 57.69 | 45.50 | 25.92 | 74.09 | 63.07 |
| CrossVLT [48] | | Swin-B | 70.62 | 65.05 | 57.40 | 45.80 | 26.10 | 74.33 | 62.84 |
| CroBIM [34] | | Swin-B | 75.75 | 70.34 | 63.12 | 51.12 | **28.45** | 73.61 | 67.32 |
| **Enti-TwistNet (ours)** | CLIP | Swin-B | **76.84** | **71.36** | **63.50** | **52.44** | **27.59** | **74.40** | **67.64** |
| **Enti-TwistNet (ours)** | BERT | Swin-B | **76.13** | **71.47** | **63.48** | **52.38** | 27.46 | **74.07** | **67.50** |

TABLE IV
ABLATION STUDY OF CORE MODEL COMPONENTS ON THE RRSIS-D DATASET.

| SEG | DAT | $Pr@0.5$ | $Pr@0.7$ | $Pr@0.9$ | mIoU | *Time* | *FLOPs* |
|---|---|---|---|---|---|---|---|
| | | 69.68 | 51.92 | 21.79 | 59.38 | *0.65ms* | *154G* |
| ✓ | | 72.66 | 53.29 | 22.46 | 62.54 | *0.75ms* | *158G* |
| | ✓ | 75.68 | 55.98 | 23.23 | 65.02 | *0.72ms* | *157G* |
| ✓ | ✓ | **77.94** | **57.07** | **24.80** | **66.83** | *0.95ms* | *160G* |
| ReMamber [24] | | 76.36 | 55.80 | 22.66 | 64.27 | *1.15ms* | *201G* |

TABLE V
ABLATION STUDY OF THE SELF-REASONING REFINEMENT STEP WITHIN THE SEG MODULE ON THE RRSIS-D DATASET.

| Option | | $Pr@0.5$ | $Pr@0.7$ | $Pr@0.9$ | oIoU | mIoU |
|---|---|---|---|---|---|---|
| w/o DAT | w/o SR | 70.25 | 52.40 | 22.16 | 74.89 | 60.73 |
| | w/ SR | 72.66 | 53.29 | 22.46 | 75.92 | 62.54 |
| w/ DAT | w/o SR | 76.39 | 56.64 | 23.39 | 77.36 | 65.60 |
| | w/ SR | 77.94 | 57.07 | 24.80 | 78.16 | 66.83 |

We additionally evaluate three remote sensing VLMs listed in Table I. They exhibit notably lower performance on RRSIS tasks compared to Enti-TwistNet. It suggests that existing VLMs, while powerful in generic multimodal settings, are not yet well aligned to the fine-grained, instance-level RRSIS.

Furthermore, a visual comparison of segmentation results across different methods is provided in the **supplementary material**.

### D. Ablation Study

We perform comprehensive ablation studies with BERT as the text encoder.

*1) Module Ablation of Enti-TwistNet:* As shown in Table IV, ablations on the RRSIS-D dataset verify the pivotal roles of both the SEG module and the DAT mechanism. SEG introduces entity priors that alleviate semantic dispersion, while DAT leverages hierarchical interaction to bridge the cross-modal gap. Combined, they deliver a 7.45% mIoU improvement over the baseline. Under identical settings, we also compare Enti-TwistNet and ReMamber in terms of FLOPs (via `ptflops`) and average inference time. Results show that

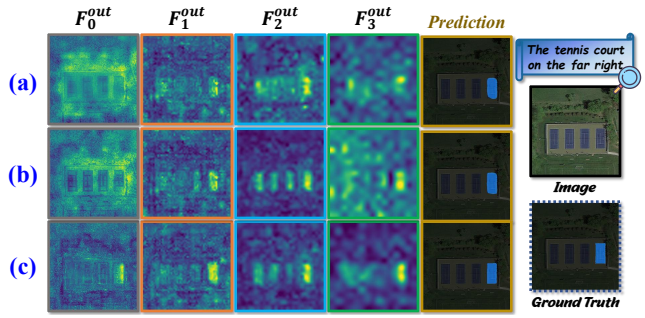



Fig. 5. Visualization of language-referred image features across the four stages. Each row corresponds to a different model configuration: (a) denotes the baseline without SEG and DAT, employing only multi-stage cross-attention for modality interaction; (b) indicates the inclusion of the SEG module; (c) represents the complete Enti-TwistNet with the SEG and the DAT.

Enti-TwistNet achieves higher accuracy and faster inference, confirming that SEG and DAT are efficiency-conscious designs offering clear gains without extra computational cost.

Fig. 5 presents feature maps from four extraction stages under SEG and DAT ablation settings. SEG enhances boundary delineation via entity-aware guidance but shows limited comprehension of referential language, leading to coarse localization. Introducing DAT further refines segmentation by enabling higher-quality cross-modal interactions.

*2) Design of the SEG Module:* To investigate the impact of entity prior granularity on segmentation performance, we conduct an ablation on the number of initial masks extracted by SAM during the **entity priors extraction** stage. As shown in Fig. 6, mIoU on three datasets increases notably as $K$ grows from 10 to 40, since more masks provide finer and more complete entity-aware guidance. However, when $K$ is too small, large or disjoint targets are under-covered, leading to incomplete spatial constraints and under-segmentation. Conversely, excessively large $K$ introduces over-segmented and irrelevant background fragments, which can mislead the self-reasoning process and cause false-positive activations. Beyond 40 masks, the gains plateau as redundancy outweighs useful new coverage, while computation increases. Thus, $K = 40$ offers the best trade-off between accuracy, robustness, and

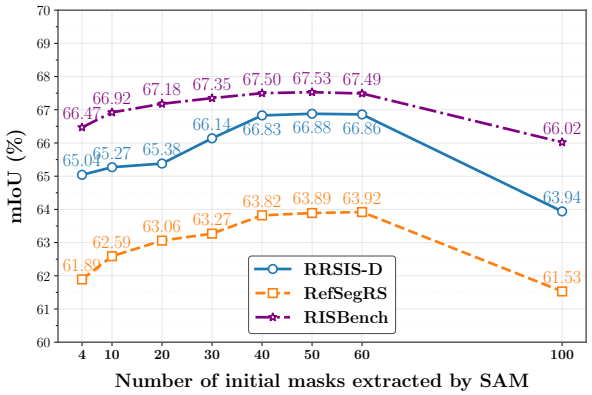



Fig. 6. Impact of initial mask numbers.

TABLE VI
ABLATION STUDY OF CROSS-MODAL SEMANTIC PROPAGATION (SP) AND ENTITY-AWARE SPATIAL ATTENTION TWIST (SA) WITHIN THE DAT MECHANISM ON THE RRSIS-D DATASET.

| SP | SA | $Pr$@0.5 | $Pr$@0.7 | $Pr$@0.9 | oIoU | mIoU |
|---|---|---|---|---|---|---|
| ✓ | | 70.2 | 46.33 | 11.42 | 68.53 | 53.67 |
| | ✓ | 75.71 | 56.21 | 23.31 | 77.22 | 65.45 |
| Parallel | | 77.25 | 57.02 | 23.79 | 77.90 | 66.47 |
| SA→SP | | 77.62 | **57.16** | 24.31 | 77.85 | 66.23 |
| SP→SA | | **77.94** | 57.07 | **24.80** | **78.16** | **66.83** |
| ReMamber-V | | 76.49 | 56.23 | 23.26 | 77.61 | 65.12 |

efficiency, and is adopted as the default in all experiments.

Table V quantitatively demonstrates the critical role of the self-reasoning refinement step within our SEG module. Regardless of whether the DAT mechanism is integrated, models incorporating self-reasoning (w/ SR) consistently outperform those without it (w/o SR) across all metrics. This improvement confirms that self-reasoning effectively exploits the inherent spatial and semantic relationships among the initial masks, thereby aggregating a more coherent and robust set of entity-aware guidance. This refined guidance, in turn, consistently leads to improved segmentation performance.

*3) Design of the DAT mechanism:* Table VI presents a comparative analysis of different DAT design strategies. Segmentation results for the individual effects of cross-modal semantic propagation (SP) and entity-aware spatial attention twist (SA) are provided first. We observe a pronounced performance decline when only SP is applied, which can be attributed to semantic propagation altering the data distribution and mitigating the modality gap, while lacking modeling of cross-modal positional dependencies. Additionally, we discuss alternative combination strategies for SP and SA. The 'parallel' strategy executes both steps simultaneously and fuses their outputs by summation. Overall, performing semantic propagation before spatial attention ("SP→SA") yields superior results, indicating that spatial positional dependencies are better established after modal alignment. To further verify the effectiveness of the proposed DAT design over Mamba-based twisting schemes such as ReMamber, we introduce an additional baseline **ReMamber-V** in Table VI. This variant extends the original ReMamber by applying the same "channel–spatial" Mamba twisting mechanism on our tri-modal feature cube (visual–text–entity). As reported, DAT consistently outperforms ReMamber-V by 1.71% in mIoU on the RRSIS-D dataset. It validates the key advantage of DAT: the replacement of the spatial Mamba scan with *entity-aware cross-attention* yields stronger fine-grained alignment between language, vision, and entity priors.

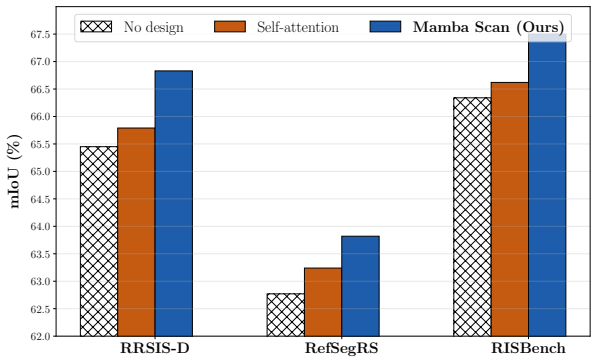



Fig. 7. Strategy comparison of cross-modal semantic propagation within the DAT mechanism. 'No design' refers to the absence of cross-modal semantic propagation, while 'Self-attention' denotes applying a channel-wise self-attention mechanism to the tri-modal feature cube $F^{cube}$.

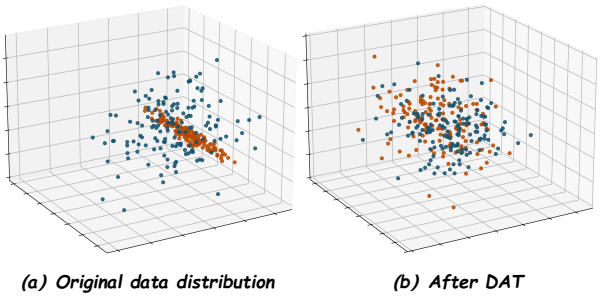



Fig. 8. Data distribution after the DAT mechanism.

To examine different cross-modal semantic propagation strategies, we conduct a comparative analysis of three approaches across three datasets. As shown in Fig. 7, the Mamba scanning strategy outperforms channel-wise self-attention. This superiority can be attributed to its linear computational complexity, which enables more efficient processing of feature channels, particularly as channel dimensionality increases. More importantly, Mamba's data-driven selective mechanism dynamically filters and highlights critical cross-modal semantic information, enabling more targeted and fine-grained feature fusion.

Fig. 8 employs PCA to project features into 3D space, intuitively demonstrating the effectiveness of the DAT mechanism. Visual and textual features are shown as blue and red dots, respectively. In Fig. 8(a), textual features form a linear structure in 3D space—caused by replication across $H \times W$—yet remain clearly separated from visual features, revealing a substantial modality gap. In Fig. 8(b), after DAT processing, visual and textual features are more dispersed and intermixed, indicating that DAT promotes fine-grained cross-modal alignment.

## V. Conclusion

In this paper, we present Enti-TwistNet, a novel framework tailored to address the challenges of semantic dispersion and modality gaps in RRSIS. Our key contributions are twofold. First, we introduce the SAM-inspired Entity Guidance module, which harnesses fine-grained entity priors and incorporates a self-reasoning process to deliver robust, coherent, and entity-aware guidance. Second, we propose the Dual-phase Attention-Twisting mechanism, which alleviates the modality gap by decoupling cross-modal semantic propagation from fine-grained spatial dependency modeling. Extensive experiments on two challenging RRSIS datasets demonstrate that both SEG and DAT substantially enhance segmentation performance, with their synergy setting a new state-of-the-art. Ablation studies further substantiate the effectiveness and precise role of each proposed component. While our framework shows strong performance, its current evaluation is limited to available public RRSIS datasets and relies on SAM-derived priors, which may not fully capture the diversity of real-world scenarios. Future work will explore more lightweight or task-specific entity prior generation and broader cross-dataset validation to further enhance generalization.

## References


[1] C. Mulverhill, N. C. Coops, and A. Achim, "Continuous monitoring and sub-annual change detection in high-latitude forests using harmonized landsat sentinel-2 data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 309–319, 2023.

[2] M. Liu, Z. Hu, B. Zhou, H. Hu, C. Qiu, and X. Zhang, "Cross-modal event extraction based on adaptive feature selection and semantic-aware graph," *Knowledge-Based Systems*, vol. 326, p. 114038, 2025.

[3] T. Sung, Y. Kang, and J. Im, "Enhancing satellite-based wildfire monitoring: Advanced contextual model using environmental and structural information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[4] Z. Wang, J. Yi, A. Chen, L. Chen, H. Lin, and K. Xu, "Accurate semantic segmentation of very high-resolution remote sensing images considering feature state sequences: From benchmark datasets to urban applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 824–840, 2025.

[5] Y. Zhang, X. Li, J. Huang, H. Guan, and H. Huang, "A novel framework for urban land cover change detection with nasa's black marble nighttime lights product," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–9, 2023.

[6] Y. He, J. Wang, Y. Zhang, and C. Liao, "An efficient urban flood mapping framework towards disaster response driven by weakly supervised semantic segmentation with decoupled training samples," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 207, pp. 338–358, 2024.

[7] S. Voigt, T. Kemper, T. Riedlinger, R. Kiefl, K. Scholte, and H. Mehl, "Satellite image analysis for disaster and crisis-management support," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1520–1528, 2007.

[8] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[9] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu, "Decouple and weight semi-supervised semantic segmentation of remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 212, pp. 13–26, 2024.

[10] C. Yang, X. Han, T. Han, H. Han, B. Zhao, and Q. Wang, "Edge approximation text detector," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 9, pp. 9234–9245, 2025.

[11] W. Xuan, H. Qi, and A. Xiao, "Tsg-seg: Temporal-selective guidance for semi-supervised semantic segmentation of 3d lidar point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 216, pp. 217–228, 2024.

[12] Y. Jia, J. Li, and Q. Wang, "Generalized few-shot semantic segmentation for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–10, 2025.

[13] Y. Jia, J. Gao, Y. Yuan, and Q. Wang, "Holistic mutual representation enhancement for few-shot remote sensing segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[14] Q. Wang, Y. Jia, W. Huang, J. Gao, and Q. Li, "Embedding generalized semantic knowledge into few-shot remote sensing segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.

[15] S. Huang, S. He, and B. Wen, "Zori: Towards discriminative zero-shot remote sensing instance segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3724–3732.

[16] C. Yang, B. Zhao, Q. Zhou, and Q. Wang, "Mmo-ig: Multiclass and multiscale object image generation for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.

[17] L. P. Osco, Q. Wu, E. L. De Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. M. Junior, "The segment anything model (sam) for remote sensing applications: From zero to one shot," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103540, 2023.

[18] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 16 321–16 330.

[19] M. Liu, B. Zhou, H. Hu, C. Qiu, and X. Zhang, "Cross-modal event extraction via visual event grounding and semantic relation filling," *Information Processing and Management*, vol. 62, no. 3, p. 104027, 2025.

[20] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 155–18 165.

[21] M. Liu, X. Jiang, and X. Zhang, "Cadformer: Fine-grained cross-modal alignment and decoding transformer for referring remote sensing image segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 14 557–14 569, 2025.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[23] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[24] Y. Yang, C. Ma, J. Yao, Z. Zhong, Y. Zhang, and Y. Wang, "Remamber: Referring image segmentation with mamba twister," in *Computer Vision – ECCV 2024*. Springer Nature Switzerland, 2025, pp. 108–126.

[25] L. Ji, Y. Du, Y. Dang, W. Gao, and H. Zhang, "A survey of methods for addressing the challenges of referring image segmentation," *Neurocomputing*, vol. 583, p. 127599, 2024.

[26] N. Kim, D. Kim, C. Lan, W. Zeng, and S. Kwak, "Restr: Convolution-free referring image segmentation using transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 145–18 154.

[27] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 686–11 695.

[28] M. Qu, Y. Wu, Y. Wei, W. Liu, X. Liang, and Y. Zhao, "Learning to segment every referring object point by point," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 3021–3030.

[29] M. Lan, F. Rong, H. Jiao, Z. Gao, and L. Zhang, "Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[30] Y. Zhan, Z. Xiong, and Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[31] K. Li, D. Wang, H. Xu, H. Zhong, and C. Wang, "Language-guided progressive attention for visual grounding in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[32] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "Rrsis: Referring remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[33] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 658–26 668.

[34] Z. Dong, Y. Sun, Y. Gu, and T. Liu, "Cross-modal bidirectional interaction model for referring remote sensing image segmentation," *arXiv preprint arXiv:2410.08613*, 2024.

[35] Y. Zhou, M. Lan, X. Li, L. Feng, Y. Ke, X. Jiang, Q. Li, X. Yang, and W. Zhang, "Geoground: A unified large vision-language model for remote sensing visual grounding," 2025. [Online]. Available: https://arxiv.org/abs/2411.11904

[36] S.-F. Peng, G. Sun, Y. Li, H. Wang, and G.-S. Xie, "Sam-aware graph prompt reasoning network for cross-domain few-shot segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6488–6496.

[37] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[38] S. M. Azimi, C. Henry, L. Sommer, A. Schumann, and E. Vig, "Skyscapes fine-grained semantic understanding of aerial scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7393–7403.

[39] A. Zhang, Y. Yao, W. Ji, Z. Liu, and T.-S. Chua, "Next-chat: An lmm for chat, detection and segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2311.04498

[40] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin, "Pixellm: Pixel reasoning with large multimodal model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 374–26 383.

[41] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5745–5753.

[42] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han, "Linguistic structure guided context modeling for referring image segmentation," in *European conference on computer vision*. Springer, 2020, pp. 59–75.

[43] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li, "Referring image segmentation via cross-modal progressive comprehension," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 488–10 497.

[44] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4424–4433.

[45] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4761–4775, 2022.

[46] H. Ding, S. Zhang, Q. Wu, S. Yu, J. Hu, L. Cao, and R. Ji, "Bilateral knowledge interaction network for referring image segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 2966–2977, 2024.

[47] Z. Xu, Z. Chen, Y. Zhang, Y. Song, X. Wan, and G. Li, "Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 17 503–17 512.

[48] Y. Cho, H. Yu, and S.-J. Kang, "Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 5823–5833, 2024.

[49] S.-A. Liu, Y. Zhang, Z. Qiu, H. Xie, Y. Zhang, and T. Yao, "Caris: Context-aware referring image segmentation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, p. 779–788.

[50] 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 Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–11, 2025.

[51] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 108–124.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

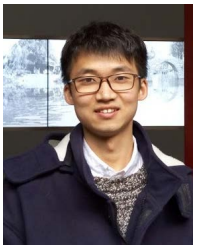
[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[55] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.


**Yuyu Jia** received the B.E. degree and the M.S. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree at the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include few-shot learning, deep learning, and remote sensing.

**Qing Zhou** is currently pursuing the Ph.D degree in computer science and technology with the school of Artificial Intelligence, Optics and Electronics (iOPEN). His research interests include computer vision and pattern recognition.



**Junyu Gao** received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015 and 2021, respectively. He is currently an associate professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

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

# Supplemental Material

Figs. 1 and 2 visually demonstrate Enti-TwistNet's superior segmentation capabilities compared to LAVT, LGCE, and RMSIN across various challenging remote sensing scenarios. While all algorithms perform comparably for large, regularly shaped objects, Enti-TwistNet truly excels in complex situations. Thanks to its **entity-aware guidance (SEG module)**, our method produces significantly smoother and more precise segmentation regions in cluttered or low-contrast backgrounds. Furthermore, the **powerful cross-modal interaction design (DAT mechanism)** enables Enti-TwistNet to more accurately comprehend complex referring logic, leading to precise identification of target objects.

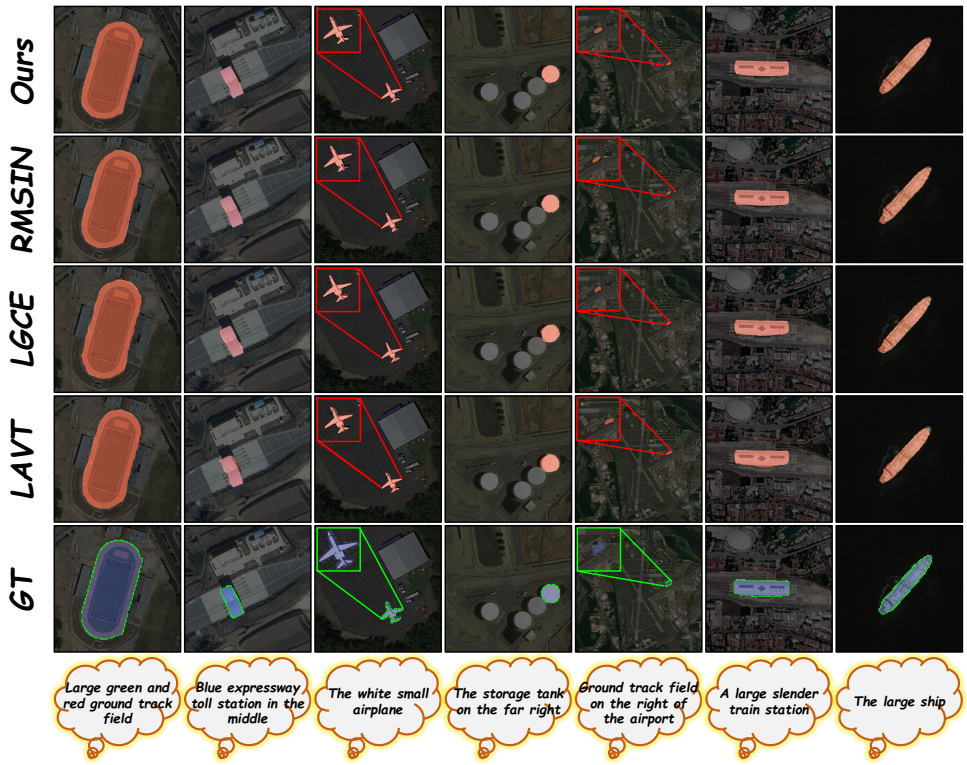



Fig. 1. Qualitative comparison of various methods on the RRSIS-D dataset.

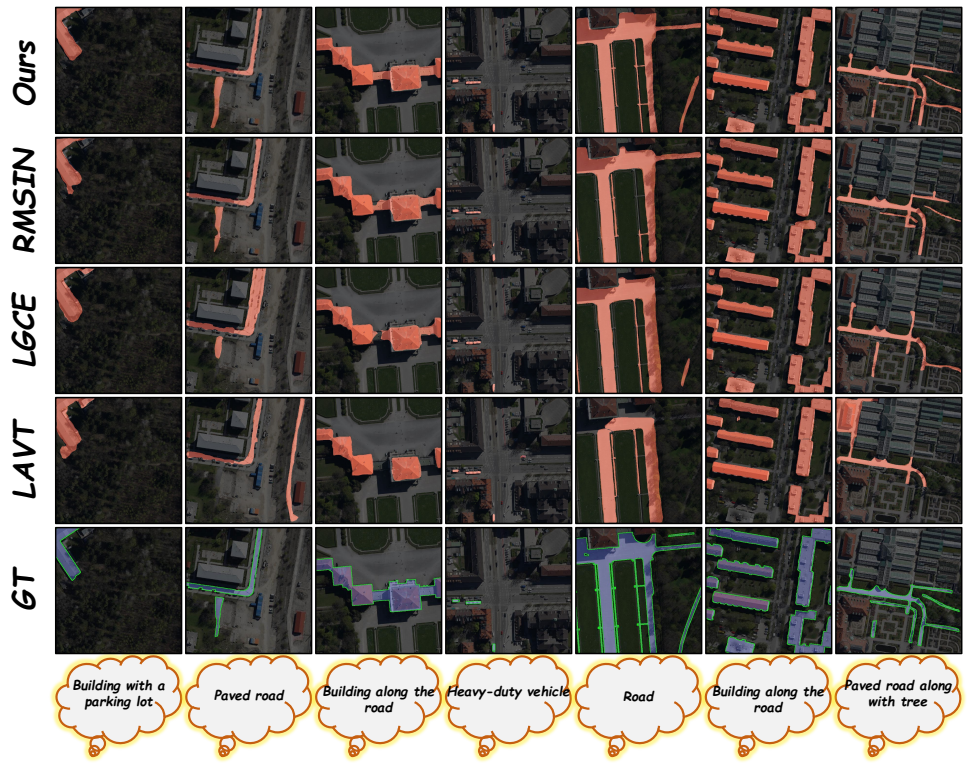



Fig. 2. Qualitative comparison of various methods on the RefSegRS dataset.