



HAL
open science

MB-ORES: A Multi-Branch Object Reasoner for Visual Grounding in Remote Sensing

Karim Radouane, Hanane Azzag, Mustapha Lebbah

► **To cite this version:**

Karim Radouane, Hanane Azzag, Mustapha Lebbah. MB-ORES: A Multi-Branch Object Reasoner for Visual Grounding in Remote Sensing. 2025. <hal-05015328>

HAL Id: hal-05015328

<https://hal.science/hal-05015328v1>

Preprint submitted on 1 Apr 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

MB-ORES: A Multi-Branch Object Reasoner for Visual Grounding in Remote Sensing

Karim Radouane¹, Hanane Azzag¹, and Mustapha lebbah²

¹ University Sorbonne Paris Nord - LIPN, Villetaneuse, France

² University Paris-Saclay - DAVID Lab, UVSQ Versailles, France

Abstract. We propose a unified framework that integrates object detection (OD) and visual grounding (VG) for remote sensing (RS) imagery. To support conventional OD and establish an intuitive prior for VG task, we fine-tune an open-set object detector using referring expression data, framing it as a partially supervised OD task. In the first stage, we construct a graph representation of each image, comprising object queries, class embeddings, and proposal locations. Then, our task-aware architecture processes this graph to perform the VG task. The model consists of: (i) a multi-branch network that integrates spatial, visual, and categorical features to generate task-aware proposals, and (ii) an object reasoning network that assigns probabilities across proposals, followed by a soft selection mechanism for final referring object localization. Our model demonstrates superior performance on the OPT-RSVG and DIOR-RSVG datasets, achieving significant improvements over state-of-the-art methods while retaining classical OD capabilities. The code will be available in our repository: <https://github.com/rd20karim/MB-ORES>.

Keywords: Object detection, Visual Grounding, Referring Expression Comprehension, Remote Sensing.

1 Introduction

Object detection (OD), a well-established task in computer vision with a wide range of applications [10] involves predicting bounding boxes and category labels for objects of interest. It began with simple problems like frontal face detection [30] and expanded to diverse categories. Traditional OD methods were designed to recognize objects given a fixed set of predefined categories (closed-set). Early contributions, such as [30], led to advances using convolutional neural networks (CNNs), including SSD [17], YOLO [21], and the RCNN family [23,25]. Building on these foundational systems, recent research has shifted towards open-set OD, where models identify both predefined and novel categories [16,2,38,32]. This transition has been driven by large pretrained vision-language models [41]. Their incorporation into OD tasks has not only improved detection accuracy but also expanded the applicability of OD to more diverse scenarios through language integration. As OD systems evolve to handle an increasingly open set

of categories, a closely related challenge emerges in visual grounding which aims to link textual descriptions to image region. While significant progress has been made in natural image datasets [33], visual grounding in remote sensing (RS) remains an emerging research area, first introduced as a novel task by [26].

To bridge the gap between the extensive advancements made in optical images compared to remote sensing (RS), our study will focus on the REC grounding task within the RS domain. Specifically, given a language expression that describes an object within an RS image, we aim to localize the single referred object while simultaneously allowing for the detection of all available objects in the image.

2 Main Contributions

We propose a flexible and novel approach that uses an open-set object detector, fine-tuned on referring expression data formulated as partially supervised object detection. Instead of depending solely on generated proposals, we incorporate object queries, initial bounding boxes, and class name embeddings, structuring them as a graph where each node captures visual (object query), spatial (bounding box coordinates), and categorical attributes. Unlike prior task-specific models such as MGVLFF [39] and LPVA [12], which are designed for referring object localization given non-ambiguous language expressions, thus disregarding classical object detection capabilities. These methods require users to visually inspect the RS image beforehand, formulate a targeted query, and have prior knowledge of RS images and their categories. In contrast, our method retains object detection capabilities while performing the REC task on demand. This approach conceptually illustrated in Figure 1 implicitly enables users to discover all objects in the image and/or target a specific object using a language query.

For this goal, we build a task-aware design that integrates specific features required by our REC task through a multi-branch network connected to a reasoner, across object proposals, along with a selector mechanism and a regressor for the final referred object localization. The technical details of each component are detailed in Section 4.

3 Related Work

3.1 Object detection

Early architectural designs for object detection used an initial set of default boxes/anchors [17] or region proposals [25], to predict relative object locations. The first transformer based OD model, DETR [3], replaced these traditional techniques with object queries and formulated OD as a set prediction problem using Hungarian matching. Many studies [15,11] have aimed to improve DETR training and accuracy. Drawing inspiration from deformable convolution [5], Deformable DETR [46] incorporates deformable attention mechanisms to enhance feature representation. DINO [40] introduces denoising training and improved

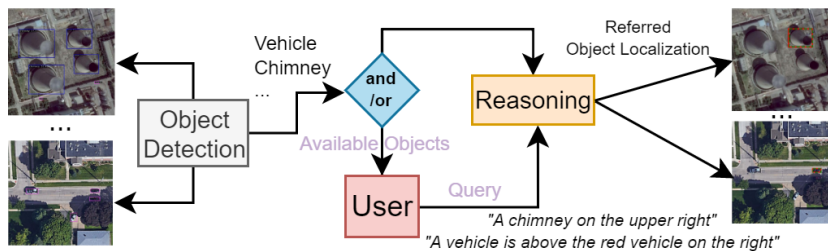


Fig. 1: Unlike previous approaches, our framework is designed to retain object detection capabilities while providing users with essential information to simplify query formulation for their object of interest.

query initialization techniques. GLIP [13] reformulates object detection as a phrase grounding task, aligning textual descriptions with corresponding image regions. Building on these techniques, GroundingDINO [16] was proposed as an effective framework for open-set object detection. However, while [16] demonstrates robust OD capabilities it struggles to accurately isolate a single referred object in REC tasks. Its performance as an open-set OD reveals a significant gap when the text prompt targets a unique object, as highlighted by the authors in [16] (Section D.3/C.6). The model generates bounding boxes for all objects mentioned in the text description, rather than isolating the specific one that satisfies the spatial/visual constraints. While it’s possible to filter the output by the highest text score, this approach consistently fails when non-referred objects receive higher confidence scores. This challenge is particularly evident in remote sensing datasets like OPT-RSVG [12] and DIOR-RSVG [39], which contain many ambiguous cases with spatial/visual constraints (see Appendix C).

3.2 Visual Grounding

At the intersection of computer vision and natural language processing, visual grounding involves localizing specific regions or objects within an image based on a given textual description [45]. This broad task can encompass several specific tasks, including Referring Expression Comprehension (REC), which aims to locate a specific target object in an image guided by a natural language query [31]. Phrase Grounding (PG) focuses on identifying multiple regions in an image mentioned in a sentence [27]. General Referring Expression Comprehension (GREC) [8], extends the scope of REC by addressing more complex scenarios where a sentence can have multiple targets or, no target at all. VG approaches are classically divided into two categories:

Two-stage VG. This approach involves two steps: first, generating a set of region proposals from the image using a pre-trained object detector; second, ranking these proposals based on their alignment with the referring expression and selecting the proposal with the highest alignment score for referring object localization [4,14].

One-stage VG. In contrast, one-stage methods are designed to directly predict the grounding bounding box by jointly processing the image and the referring expression in a single step, without relying on an intermediate proposal generation phase [36,35,9,6,37].

4 Methods

Although GroundingDINO as open-set object detector has practical limitations for REC tasks, it still demonstrates satisfactory ability in generating object proposals for optical images. To leverage this capability and maintain its core functionalities for object detection, we retain its original design and apply slight fine-tuning for transfer to RS domain using REC data as partially supervised OD. Then GroundingDINO outputs are structured as graph-based representation of image objects, where each object proposal node contains information about its bounding box, object query, and class name embedding. In the second stage, we incorporate a task-aware design that processes this graph input to target specific referred objects and regress their bounding boxes. Differing from previous RS approaches [12,39], our final framework unifies OD and REC for remote sensing through effective representation learning and reasoning processes, significantly enhancing REC performance. Our approach integrates explicit spatial/visual reasoning, semantic alignment, and robust bounding box refinement. The following provides an overview (Section 4.1), followed by a detailed explanation of each framework component: the first stage in Section 4.2, the second stage consisting of representation learning (Section 4.3), reasoning and selection (Section 4.4), and finally, referred object localization (Section 4.5).

4.1 Overview

We propose a two-stage framework, Multi-Branch Object REaSoner (MB-ORES), for the REC grounding task. MB-ORES leverages explicit prior knowledge and cross-modal alignment, as illustrated in Figure 2. First, we fine-tune GroundingDINO to generate object query proposals, bounding box coordinates, and class name embeddings. Each input is processed separately through our three input branches, which handle different types of information, generating updated representations that integrate language, spatial, and visual awareness. Next, this information is fused, and the object reasoner aligns the fused representations with the referring expression, producing a probability distribution over object queries. A soft query selection mechanism then aggregates the object queries, weighted by these probabilities, into a refined representations that is input to an FFN regression head for precise bounding box refinement. The core of our two-stage approach relies on the following techniques:

1. **Graph Representation of RS Image.** We fine-tune an open-set object detector on REC data as a partially supervised OD task, and structure its outputs as a graph representation, where each node encodes an object’s visual, spatial, and categorical attributes as a separate modalities.

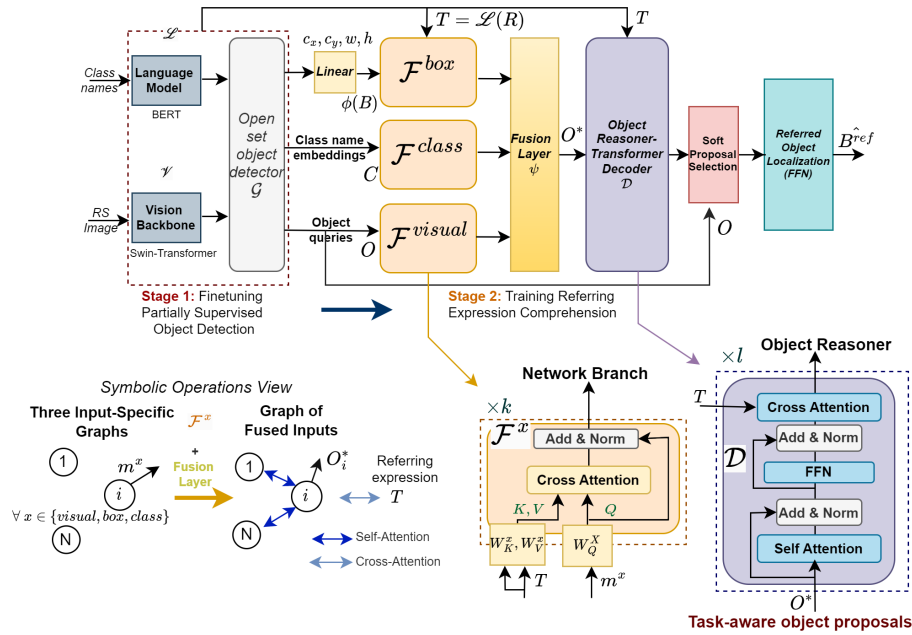


Fig. 2: Our Overall Framework (MB-ORES): In the first stage, the object detector is trained on partially annotated images from the REC data, producing output structured as a graph. In the second stage, these outputs are processed through a multi-branch network, fused into task-aware object proposals, and refined using reasoning and selection modules to generate the final representation for referred object localization.

- 2. Multi-branch Network and Cross-Modal Fusion.** Each node’s input representation is processed by a separate network branch, which is updated with referring text information. The output features from the multi-branch network are then fused to generate task-aware object proposal representations.
- 3. Object Reasoning and Soft Selection.** This component models the reasoning process to identify the referred object among object proposals and applies a soft object query selection mechanism.
- 4. Referred Object Localization.** A final specialized feedforward network (FFN) head predicts the bounding box of the referred object.

4.2 Fine-Tuning

We use the pretrained GroundingDINO, which employs the tiny version of Swin-Transformer [18] as the visual backbone and BERT [7] as the language model. We fine-tune GroundingDINO on the REC datasets (DIOR-RSVG and OPT-RSVG)

and denote this model as \mathcal{G} . Fine-tuning is performed separately for each dataset to prevent data leakage, as OPT-RSVG partially overlaps with DIOR-RSVG.

Formally, given the concatenation of all possible class names t_c , for each image I , the object detector \mathcal{G} produces a set of object queries, bounding boxes and class embeddings.

$$\{O \in \mathbb{R}^{N \times D_{obj}}, B \in \mathbb{R}^{N \times 4}, C \in \mathbb{R}^{N \times D_{obj}}\} = \mathcal{G}(I, t_c) \quad (1)$$

where N is the number of object queries and D_{obj} is the dimensionality of each query, B the bounding box coordinates and C class name embedding.

4.3 Cross-Multimodal Branches and Fusion

In this section, we provide formal details on our three-branch network \mathcal{F} for integrating prior knowledge to form task-aware object representations.

Network Branches. We use the notation $\mathcal{F} = \{\mathcal{F}^{\text{box}}, \mathcal{F}^{\text{class}}, \mathcal{F}^{\text{visual}}\}$ and for each object node we consider, the following:

- Visual attributes: Object query $O_i \in \mathbb{R}^{D_{obj}}$.
- Spatial attributes: Predicted bounding box $B_i = [c_x^i, c_y^i, w^i, h^i]$.
- Categorical attributes: Class name embedding $C \in \mathbb{R}^{D_{obj}}$.

Particularly, the bounding box coordinates are projected using a linear function $\phi: \mathbb{R}^4 \rightarrow \mathbb{R}^{D_{obj}}$ yielding $\phi(B_i) \in \mathbb{R}^{D_{obj}}$.

Given a referring expression query text R , tokenized into n_k tokens with an embedding dimension of d_t , and a language model \mathcal{L} , the token representations are denoted as $\mathbf{T} \in \mathbb{R}^{n_k \times d_t}$, which encodes the semantic meaning of each token in the referring expression R :

$$\mathbf{T} = \mathcal{L}(R) \quad (2)$$

Each network branch \mathcal{F}^x , where $x \in \{\text{box}, \text{class}, \text{visual}\}$, models the interaction and alignment with the referring expression representation \mathbf{T} separately using a multi-head cross-attention operation, denoted as \mathcal{A} . We define \mathcal{F}^x as follows:

$$\mathcal{F}^x(m^x, T) = m^x + \mathcal{A}(Q^x, K^x, V^x) \quad \forall x \in \{\text{box}, \text{class}, \text{visual}\} \quad (3)$$

Cross-Modal Interaction. We employ h attention heads, each with its own set of learned projection matrices. The attention for the i -th head is computed as follows:

$$\mathcal{H}_i = \text{softmax} \left(\frac{(m^x W_{Q,i}^x)(T W_{K,i}^x)^\top}{\sqrt{D_{obj}}} \right) (T W_{V,i}^x) \quad \forall i \in [1, H] \quad (4)$$

- m^x refers to the input of branch x and acts as the query source, while the token representations T are used to extract the keys and values.

- $W_{Q,i}^x$, $W_{K,i}^x$, and $W_{V,i}^x$ are learned projection matrices for branch x that map the inputs to the query, key, and value spaces respectively for each head i .

The outputs from the h heads are then aggregated via concatenation and passed through a final projection matrix W_O to produce the overall multi-head attention output:

$$\mathcal{A}(Q^x, K^x, V^x) = \text{Concat}(\mathcal{H}_1, \dots, \mathcal{H}_h)W_O \quad (5)$$

We described the cross-modal operations for a single layer $k = 1$ for notation simplicity; however, \mathcal{F}^x generally consists of k multiple layers, forming a multi-layer network branch defined as

$$\mathcal{F}^x = \mathcal{F}_k^x \circ \dots \circ \mathcal{F}_1^x.$$

This multi-layer structure, along with the incorporation of a multi-head attention mechanism in each layer, enables the model to capture diverse features from different subspaces of the input, thereby enriching its representation of complex cross-modal interactions (cf. Figure 2). Our final outputs of interest are defined by the following equation:

$$\{\tilde{B}, \tilde{C}, \tilde{O}\} = \{\mathcal{F}^{\text{box}}(\phi(B), T), \mathcal{F}^{\text{class}}(C, T), \mathcal{F}^{\text{visual}}(O, T)\} \quad (6)$$

Fusion Layer. These features are concatenated and fused with a projection layer function $\psi : m \mapsto m.W^\psi$ of learnable weights $W^\psi \in \mathbb{R}^{(3 \cdot D_{obj}) \times D_{obj}}$:

$$O^* = \psi \left(\text{Concat}(\tilde{B}, \tilde{C}, \tilde{O}) \right) \in \mathbb{R}^{N \times D_{obj}}, \quad (7)$$

O^* is the task-aware updated object proposal representations.

4.4 Object Reasoner Network and Selection

Given the outputs from the fusion layer ψ , this step consists of two key elements, object reasoner and selection mechanism:

Object Reasoner Network. Given the updated proposals O^* from the multi-modal cross-fusion network, the goal of the object reasoner network is to output a probability distribution, guided by the referring expression T , across all proposals. Formally, our objective is to learn a function f parameterized by θ that predicts a probability distribution P over the object candidates given the text T :

$$P(y | \mathbf{T}, \mathbf{O}^*; \theta) = f(T, \mathbf{O}^*; \theta) \quad (8)$$

where $y \in \{1, 2, \dots, N\}$ indexes the N object proposals.

We define the function f as a transformer decoder \mathcal{D} , formally described in [29], which has been successful for a wide range of applications. In our case, self-attention is used to model communication between object nodes O^* and understand their respective locations and visual characteristics in the image,

while cross-attention updates these representations based on their relevance to the text tokens of the referring expression T .

$$\{s_1, \dots, s_N\} = \mathcal{D}(O^*, T, T) \quad (9)$$

Where s_i is the decoder output logit score for the i -th object proposal, softmax function gives the probability distribution over object proposals:

$$p_i = \frac{\exp(s_i)}{\sum_{j=1}^N \exp(s_j)}, \quad \forall i = 1, \dots, N \quad (10)$$

Soft Proposal Selection. Instead of the non-differentiable hard selection based on the maximum score, we employ a soft selection mechanism that allows adjusting the selection process based on the localization precision of the referred object during optimization.

$$O_{ref} = \sum_{i=1}^N p_i \cdot O_i \quad (11)$$

Note that we use the original object queries $O = \{O_i, \forall i \leq N\}$ from the fine-tuned GroundingDINO, which leads to faster convergence for localization (represented as a skip connection in Figure 2). However, the object detector weights are frozen during this second stage, and only our lightweight model is updated.

4.5 Referred Object Localization

Finally, given the soft query-aware visual representation O_{ref} , which encodes prior knowledge about the objects' distribution in the image conditioned on the referring expression query, we use it as input to a regression head modeled as a simple feed-forward network (FFN) that predicts the refined bounding box coordinates:

$$\hat{B}^{ref} = \mathcal{FFN}(O_{ref}), \quad (12)$$

where $\hat{B}^{ref} = [\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h}]$ denotes the predicted bounding box of the referred object.

4.6 Loss Function

The overall loss \mathcal{L} is composed of three terms:

$$\mathcal{L} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{giou} \mathcal{L}_{giou} + \mathcal{L}_{L1}. \quad (13)$$

Classification Loss \mathcal{L}_{cls} : This loss aims to maximize the logits corresponding to the GroundingDINO object query O_k associated with the bounding box B_k that has the highest Intersection over Union (IoU) with the ground truth bounding box B^{gt} .

$$\mathcal{L}_{\text{cls}} = -\log(p_r), \quad r = \arg \max_{k=1, \dots, N} \{\text{IoU}(B_k, B^{\text{gt}})\} \quad (14)$$

Localization losses $\mathcal{L}_{\text{giou}}$ and \mathcal{L}_{L1} : The regression loss consists of the GIoU loss [22] and the L1 loss computed between the predicted bounding box \hat{B}^{ref} and the ground truth B^{gt} of the referred object:

$$\mathcal{L}_{\text{giou}} = 1 - \text{GIoU}(\hat{B}^{\text{ref}}, B^{\text{gt}}), \quad \mathcal{L}_{\text{L1}} = \|\hat{B}^{\text{ref}} - B^{\text{gt}}\|_1. \quad (15)$$

5 Datasets benchmarks

Table 1 presents the different splits used in the literature for the RSVG datasets. For DIOR-RSVG, we used the original split proposed in [39], which is the standard adopted split for this dataset in model performance comparisons. For OPT-RSVG, a larger dataset, we used its predefined split. For evaluation we use the same metrics as defined in [39]. We briefly recall that $\text{meanIoU} = (\sum_s I_s / U_s) / N_r$ and $\text{cmIoU} = \sum_s I_s / \sum_s U_s$, computed over all split samples, where I_s and U_s are, respectively, the Intersection and Union of each sample s with its referred object ground truth. N_r is the number of referred objects in the entire test set.

Table 1: Split statistics for each dataset.

| Dataset | Train | Validation | Test | Total |
|-----------|-------|------------|-------|-------|
| OPT-RSVG | 19580 | 4895 | 24477 | 48952 |
| DIOR-RSVG | 15328 | 3832 | 19160 | 38320 |

6 Implementation Details

In this section, we describe the experimental settings for each training stage.

First Stage. We finetune GroundingDINO with $l_r = 10^{-5}$ learning rate with a batch size of 8. This task is considered partially supervised object detection because we use only the referred object annotations from the training split, which typically do not cover all objects in each image.

Second Stage. Given the outputs from the first stage, for each image, we select the top N object queries with the highest classification scores from the fine-tuned model. We set $N = 300$, which provides the best trade-off between average recall and computational efficiency.

Multi-branch Network: We experiment with the use of multiple cross-attention layers $k \in \{1, 3\}$ and also analyze the effect of omitting these network branches. We don't use a higher number of layers k to maintain a lightweight model.

Object Reasoner: We experiment with different numbers of layers $l \in \{3, 6\}$ and attention heads $h \in \{4, 8\}$. The object feature dimension is set to $D_{\text{obj}} = 256$, defined by the finetuned model.

Referred Object Regression: Our specialized FFN head for referred object localization is initialized with the parameters from the frozen FFN regression head of GroundingDINO, leveraging its fine-tuned initial localization capabilities.

Optimization: We use a batch size of 8 with an initial learning rate of 1×10^{-4} in the AdamW optimizer [19]. The loss weights are set to $\lambda_{\text{cls}} = 100$, $\lambda_{\text{giou}} = 5$. The classification loss has the highest weight because the model, having already been fine-tuned for localization, should focus in the early training stages on correctly selecting the best proposal for the referred object in the referring expression, then refine the localization precision through remaining losses.

7 Experimental Results

Table 2 presents the results on the DIOR-RSVG dataset, where our method outperforms the current best model, LPVA [12], across various threshold levels by clear margins (+3.38% up to +14.89%). However, a discrepancy is observed in the *meanIoU* (+5.38%) and *cmuIoU* (-2.05%) values. This could suggest that our model performs better on smaller objects, but less on larger objects compared to LPVA. However, when trained on the OPT-RSVG dataset, our model achieves superior performance across all metrics with clear margins, as shown in Table 3. In particular, we observe a +6.98% and +3% increase in the global metrics meanIoU and cmuIoU, respectively, while precision/accuracy improvements range from +5.78% to +10.4%.

When compared to the recent vision-language model GeoGround [44], which achieves 77.73% Pr@0.5³, our approach achieves 82.78% accuracy, yielding an improvement of +5.05%, while also outperforming EarthGPT [42] on all reported metrics.

Table 2: Comparison with state-of-the-art (SOTA) methods for our model on the original split of DIOR-RSVG.

| Methods | Venue | Visual Encoder | Language Encoder | Pr@0.5 | Pr@0.6 | Pr@0.7 | Pr@0.8 | Pr@0.9 | meanIoU | cmuIoU |
|--------------------------------|---------|----------------|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Vision-language models: | | | | | | | | | | |
| EarthGPT [42] | TGRS'24 | ViT [1] | Llama-2 [28] | 76.65 | 71.93 | 66.52 | 56.53 | 37.63 | 69.34 | 81.54 |
| GeoGround [44] | - | CLIP-ViT [20] | Vicuna 1.5 [43] | 77.73 | - | - | - | - | - | - |
| Specialist models: | | | | | | | | | | |
| ZSNet [24] | ICCV'19 | ResNet-50 | BiLSTM | 51.67 | 48.13 | 42.30 | 32.41 | 10.15 | 44.12 | 51.65 |
| FAOA [36] | ICCV'19 | DarkNet-53 | BERT | 67.21 | 64.18 | 59.23 | 50.87 | 34.44 | 59.76 | 63.14 |
| ReSC [35] | ECCV'20 | DarkNet-53 | BERT | 72.71 | 68.92 | 63.01 | 53.70 | 33.37 | 64.24 | 68.10 |
| LBYL-Net [9] | CVPR'21 | DarkNet-53 | BERT | 73.78 | 69.22 | 65.56 | 47.89 | 15.69 | 65.92 | 76.37 |
| TransVG [6] | CVPR'21 | ResNet-50 | BERT | 72.41 | 67.38 | 60.05 | 49.10 | 27.84 | 63.56 | 76.27 |
| QRNet [37] | CVPR'22 | Swin | BERT | 75.84 | 70.82 | 62.27 | 49.63 | 25.69 | 66.80 | 83.02 |
| VLTVG [34] | CVPR'22 | ResNet-50 | BERT | 69.41 | 65.16 | 58.44 | 46.56 | 24.37 | 59.96 | 71.97 |
| VLTVG [34] | CVPR'22 | ResNet-101 | BERT | 75.79 | 72.22 | 66.33 | 55.17 | 33.11 | 66.32 | 77.85 |
| MGVLF [39] | TGRS'23 | ResNet-50 | BERT | 76.78 | 72.68 | 66.74 | 56.42 | 35.07 | 68.04 | 78.41 |
| LPVA [12] | TGRS'24 | ResNet-50 | BERT | 82.27 | 77.44 | 72.25 | 60.98 | 39.55 | 72.35 | 85.11 |
| MB-ORES (Ours) | - | Swin-T | BERT | 85.65 | 83.89 | 80.87 | 73.00 | 54.39 | 77.73 | 83.06 |

Evaluating Object Detection. Although trained only with partially annotated images, Table 4 shows good OD performance relative to our challenging

³ The only reported metric.

Table 3: Comparison with SOTA methods on the test set of OPT-RSVG shows a significant improvement with our model, especially at higher thresholds.

| Methods | Venue | Visual Encoder | Language Encoder | Pr@0.5 | Pr@0.6 | Pr@0.7 | Pr@0.8 | Pr@0.9 | meanIoU | cmIoU |
|-----------------------|---------|----------------|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| NMTree [14] | ICCV'19 | ResNet-101 | BiLSTM | 69.28 | 64.17 | 55.22 | 40.31 | 12.90 | 60.12 | 69.85 |
| Ref-NMS [4] | AAAI'21 | ResNet-101 | Bi-GRU | 70.59 | 65.61 | 58.01 | 41.36 | 14.58 | 60.42 | 70.72 |
| ZSGNet [24] | ICCV'19 | ResNet-50 | BiLSTM | 48.64 | 47.32 | 43.85 | 27.69 | 6.33 | 43.01 | 47.71 |
| FAOA [36] | ICCV'19 | DarkNet-53 | BERT | 68.13 | 64.30 | 57.15 | 41.83 | 15.33 | 58.79 | 65.20 |
| LBLY-Net [9] | CVPR'21 | DarkNet-53 | BERT | 70.22 | 65.39 | 58.65 | 37.54 | 9.46 | 60.57 | 70.28 |
| TransVG [6] | CVPR'21 | ResNet-50 | BERT | 69.96 | 64.17 | 54.68 | 38.01 | 12.75 | 59.80 | 69.31 |
| VLTVG [34] | CVPR'22 | ResNet-50 | BERT | 71.84 | 66.54 | 57.98 | 42.15 | 14.63 | 61.47 | 71.10 |
| VLTVG [34] | CVPR'22 | ResNet-101 | BERT | 73.50 | 68.31 | 59.93 | 43.45 | 15.31 | 62.84 | 73.80 |
| MGVLF [39] | TGRS'23 | ResNet-50 | BERT | <u>72.19</u> | 66.86 | 58.02 | 42.51 | 15.30 | 61.51 | 71.80 |
| LPVA [12] | TGRS'24 | ResNet-50 | BERT | 78.03 | 73.32 | 62.22 | 49.60 | 25.61 | 66.20 | 76.30 |
| MB-ORES (Ours) | - | Swin-T | BERT | 83.81 | 81.54 | 76.40 | 63.82 | 36.01 | 73.18 | 79.29 |

case of partially annotated images. However, the computed metrics are likely an underestimate, as some true detections may be incorrectly marked as false positives or not counted, leading to a lower reported performance than the model’s actual capability. Qualitative visualizations for OD are presented in Appendix B.

Table 4: Evaluating object detection (OD) using only the available annotated objects from the REC test set (approximation). Trained with only a few annotated objects per image (REC train set).

| Dataset | AP@0.5 | mAP | AR@100 | AR@300 |
|-----------|--------|------|--------|--------|
| DIOR-RSVG | 67.9 | 55.8 | 84.9 | 85.0 |
| OPT-RSVG | 65.5 | 47.7 | 78.3 | 78.6 |

8 Ablation studies

In this section, we investigate the effect of various hyperparameters on each block of MB-ORES (cf. Figure 2) using both DIOR-RSVG and OPT-RSVG.

First, Table 5 highlights and demonstrates that our multi-branch network integration consistently enhances grounding accuracy across both datasets, yielding a significant improvement in all performance metrics ($>+4\%$). In the following, we provide a detailed analysis for each dataset:

DIOR-RSVG: The optimal configuration (4 heads, 3 layers, multi-branch) achieves 77.73% MeanIoU and 83.06% CmuIoU with only 7.97M, making it a lightweight model. Compared to the (4,1) multi-branch variant, this corresponds to a +0.55% and +1.39% improvement in MeanIoU and CmuIoU, respectively. Increasing to 8 heads and 6 layers (11.13M parameters) offers negligible gains. Crucially, removing the multi-branch network leads to a substantial performance drop: MeanIoU falls by -4.23%, and CmuIoU by -4.63%, highlighting the importance of our multi-branch based reasoning for our REC task.

OPT-RSVG: The same (4,3) multi-branch model achieves 73.18% MeanIoU and 79.29% CmuIoU, outperforming the (4,1) counterpart by +0.81% and +0.98%, respectively. Notably, removing the multi-branch network results in significant performance drop than in DIOR-RSVG, with MeanIoU decreasing by -6.8% and CmuIoU by -6.03%. These results indicate that multi-branch reasoning is particularly beneficial for complex expressions in OPT-RSVG, where contextual dependencies are crucial for accurate localization. While Table 6 demonstrates that even with few proposals, the model is already precise in selecting the referred object highlighting the benefits of our finetuning stage in enhancing the quality of generated proposals.

Table 5: Impact of using multiple layers in each branch and in the object reasoner network. The effect of the multi-modal branches (4,3) and fusion on performance shows a significant improvement.

| Dataset | | | | | DIOR-RSVG | | OPT-RSVG | |
|----------|--------------|-----------------|----------|--------------|--------------|--------------|--------------|--|
| # Heads | Multi-Branch | Object Reasoner | #Params. | MeanIoU | CmuIoU | MeanIoU | CmuIoU | |
| (h, l/k) | (4,1) | (4,3) | 6.38M | 77.18 | 81.67 | 72.15 | 78.27 | |
| | | (8,6) | 11.13M | 77.26 | 81.71 | 72.37 | 78.31 | |
| | (4,3) | (4,3) | 7.97M | 77.73 | 83.06 | <i>72.73</i> | <i>78.60</i> | |
| | | (8,6) | 12.70M | <i>77.72</i> | <i>82.42</i> | 73.18 | 79.29 | |
| | × | (4,3) | 5.13M | 73.50 | 77.94 | 66.04 | 72.54 | |
| | | (8,6) | 9.87M | 73.93 | 78.43 | 66.38 | 73.26 | |

Table 6: At inference time, we could maintain comparable performance with very few proposals and mitigate the limitations of previous two-stage methods.

| Dataset | OPT-RSVG | | DIOR-RSVG | |
|---------|---------------------|--------------|---------------------|--------------|
| Model | MB-ORES-(4,3)-(8,6) | | MB-ORES-(4,3)-(4,3) | |
| top_N | MeanIoU (%) | CmuIoU (%) | MeanIoU (%) | CmuIoU (%) |
| 50 | 73.10 | 79.29 | 76.62 | 82.41 |
| 100 | 73.14 | 79.34 | 77.20 | 82.85 |
| 200 | 73.18 | 79.36 | 77.64 | 82.94 |
| 300 | 73.18 | 79.29 | 77.73 | 83.06 |

9 Qualitative Analysis

In this section, we present qualitative visualization of our REC results, we show challenging examples with multiple occurrences of objects from the same instance to demonstrate the effectiveness of our framework in distinguishing the target based on linguistic, spatial, and visual attributes defined by the query referring expression.

DIOR-RSVG. Figure 3 shows different REC types, object category (<class name>), location relative to the object in the image (<relative location> <obj>, <absolute>), and visual attribute-dependent features (<color>, <size>). Figure 4 shows the results for multiple queries per image.

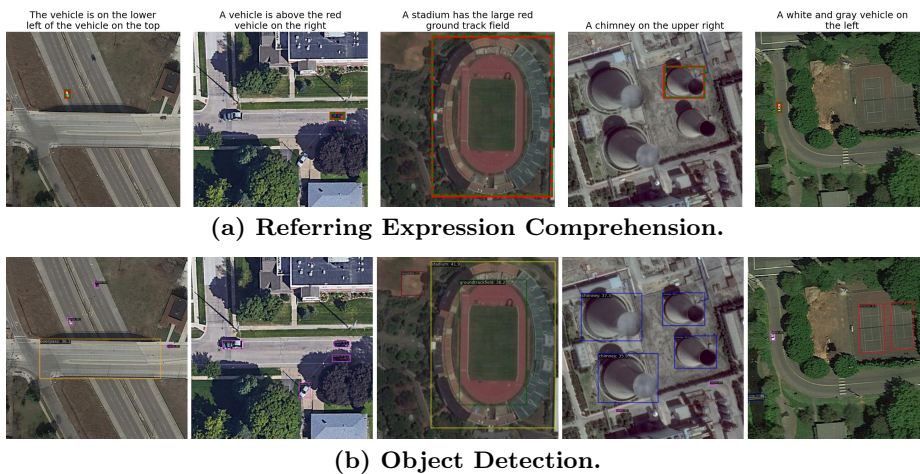


Fig. 3: DIOR-RSVG: At the top of the image, the results for the REC task are shown (prediction in red), while at the bottom, the OD task is performed simultaneously using our unified approach.

OPT-RSVG. In Figure 5, we display the grounding of multiple referring expressions for each image. Our proposed framework, with its multi-modal branch fusion, effectively disentangles the referring expressions through a correct alignment with language expression, demonstrating its ability to learn significant discriminative features that distinguish between inter- and intra-category attributes based on spatial, visual, and categorical characteristics.



Fig. 4: DIOR-RSVG: Visual Grounding of multiple referring expressions per image.

10 Conclusion

In this work, we proposed MB-ORES, a simple yet effective architecture that integrates spatial, semantic, and visual cues through a Multi-Branch network. Extensive ablation studies demonstrated its ability to significantly enhance REC performance. Moreover, our lightweight model variants maintain competitive accuracy while using fewer proposals, effectively addressing the bottlenecks of two-stage methods. Beyond achieving state-of-the-art performance on the OPT-RSVG and DIOR-RSVG datasets, our framework offers a unified solution for object detection and visual grounding. The proposed soft referring expression-aware query selection mechanism efficiently aggregates information across all object queries, refining object localization dynamically in the second training stage instead of relying on a fixed prediction from the first stage. By incorporating an open-set-based object detector, MB-ORES not only advances REC in



Fig. 5: OPT-RSVG: Visual Grounding of multiple referring expressions per image (ground-truth in dashed gray color).

remote sensing but also paves the way for future research in zero-shot reasoning and beyond.

References

- Alexey, D., Lucas, B., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., Mostafa, D., Matthias, M., Georg, H., et al., G.S.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) **10**
- Barzilai, A., Gigi, Y., Helmy, A., Silverman, V., Refael, Y., Jaber, B., Shekel, T., Leifman, G., Beryozkin, G.: A recipe for improving remote sensing vlm zero shot generalization. In: arXiv preprint arXiv:2503.08722 (2025) **1**
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) **2**
- Chen, L., Ma, W., Xiao, J., Zhang, H., Chang, S.F.: Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In: AAAI (2021) **3, 11**

5. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017) [2](#)
6. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: ICCV (2021) [4](#), [10](#), [11](#)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019) [5](#)
8. He, S., Ding, H., Liu, C., Jiang, X.: Grec: Generalized referring expression comprehension (2023) [3](#)
9. Huang, B., Lian, D., Luo, W., Gao, S.: Look before you leap: Learning landmark features for one-stage visual grounding. In: CVPR (2021) [4](#), [10](#), [11](#)
10. Kaur, J., Singh, W.: A systematic review of object detection from images using deep learning. *Multimedia Tools and Applications* (2023) [1](#)
11. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: CVPR (2022) [2](#)
12. Li, K., Wang, D., Xu, H., Zhong, H., Wang, C.: Language-guided progressive attention for visual grounding in remote sensing images. *TGRS* (2024) [2](#), [3](#), [4](#), [10](#), [11](#)
13. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022) [3](#)
14. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: CVPR (2019) [3](#), [11](#)
15. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: ICLR (2022) [2](#)
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV (2024) [1](#), [3](#)
17. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector (2016) [1](#), [2](#)
18. 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: ICCV (2021) [5](#)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) [10](#)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) [10](#)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016) [1](#)
22. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (2019) [9](#)
23. Ross, G.: Fast r-cnn. In: ICCV (2015) [1](#)
24. Sadhu, A., Chen, K., Nevatia, R.: Zero-shot grounding of objects from natural language queries. In: ICCV (2019) [10](#), [11](#)
25. Shaoqing, R., Kaiming, H., Ross, G., Jian, S.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) [1](#), [2](#)
26. Sun, Y., Feng, S., Li, X., Ye, Y., Kang, J., Huang, X.: Visual grounding in remote sensing images. In: ACM International Conference on Multimedia (2022) [2](#)
27. Tan, Y., Jiang, L., Jiang, Y.G., Feng, J.: Hierarchical semantic correspondence networks for video paragraph grounding. In: CVPR (2023) [3](#)

28. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., et al., S.B.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 10
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, , Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 7
30. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001) 1
31. Wang, Y., Tian, Z., Guo, Q., Qin, Z., Zhou, S., Yang, M., Wang, L.: Improving visual grounding with referential query (2024) 3
32. Wei, G., Yuan, X., Liu, Y., Shang, Z., Xue, X., Wang, P., Yao, K., Zhao, C., Zhang, H., Xiao, R.: Ova-det: Open vocabulary aerial object detection with image-text collaboration. In: arXiv preprint arXiv:2408.12246 (2025) 1
33. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Towards visual grounding: A survey (2024) 2
34. Yang, L., Xu, Y., Yuan, C., Liu, W., Li, B., Hu, W.: Improving visual grounding with visual-linguistic verification and iterative reasoning. In: CVPR (2022) 10, 11
35. Yang, Z., Chen, T., Wang, L., Luo, J.: Improving one-stage visual grounding by recursive sub-query construction. In: ECCV (2020) 4, 10
36. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: ICCV (2019) 4, 10, 11
37. Ye, J., Tian, J., Yan, M., Yang, X., Wang, X., Zhang, J.: Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In: CVPR (2022) 4, 10
38. Zang, Z., Lin, C., Tang, C., Wang, T., Lv, J.: Zero-shot aerial object detection with visual description regularization. AAAI (2024) 1
39. Zhan, Y., Xiong, Z., Yuan, Y.: Rsvg: Exploring data and models for visual grounding on remote sensing data. TGRS (2022) 2, 3, 4, 9, 10, 11
40. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: ICLR (2023) 2
41. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. TPAMI (2024) 1
42. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. TGRS (2024) 10
43. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. NeurIPS (2023) 10
44. Zhou, Y., Lan, M., Li, X., Ke, Y., Jiang, X., Feng, L., Zhang, W.: Geoground: A unified large vision-language model for remote sensing visual grounding (2024) 10
45. Zhu, H., Su, L., Mao, S., Ye, J.: Read before grounding: Scene knowledge visual grounding via multi-step parsing. In: COLING (2025) 3
46. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021) 2

A Appendix

This supplement provides additional visualizations for object detection. We also compare the object detection capabilities of GroundingDINO with the REC task on optical images and explain why it is better suited as a region proposal method for REC tasks, along with the intuition behind our current design.

B Object detection visualization

In Figures 6,7 we visualize the object detection qualitative results of our proposed approach alongside the previously analyzed REC task.



Fig. 6: DIOR-RSVG: Object detection results on different samples.

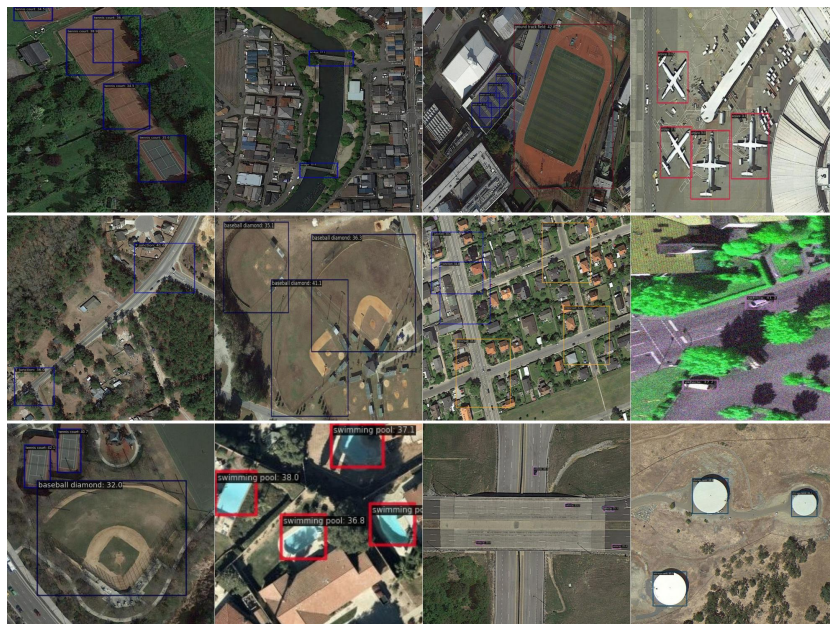
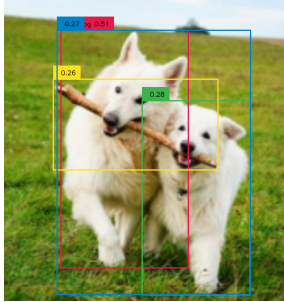


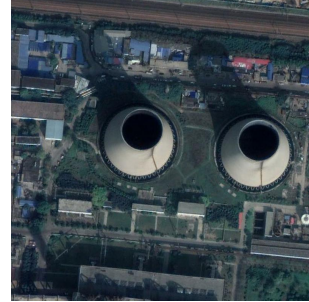
Fig. 7: OPT-RSVG: Object detection results on different samples.

C GroundingDINO limitations

Observing the practical limitations of GroundingDINO as an open-set object detector for the REC task, as illustrated in Figures 8 and 9, even with extensive pretraining on large optical image datasets. However, despite these limitations, its core design offers valuable capabilities for object detection, making it an effective region proposal network and enabling the establishment of prior knowledge about object distribution in the image. Therefore, we retain its core design and fine-tune it for the remote sensing domain. For the REC task, we introduce a task-aware, lightweight design to enable accurate referring expression comprehension within a two-stage paradigm.

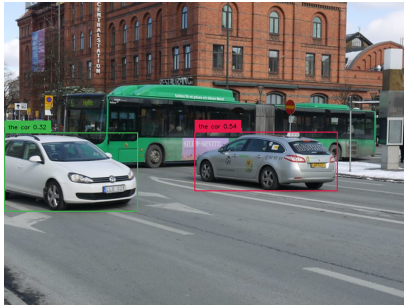


(a) Given the prompt *"the dog on the right"*, the model outputs boxes for many *"dog"* objects. Filtering by maximum score incorrectly selects a different object.

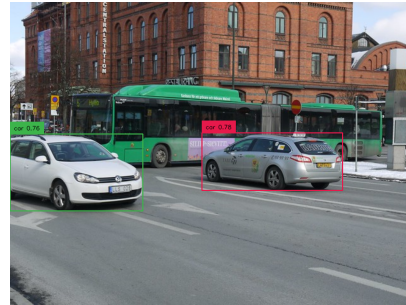


(b) Remote sensing transfer case with similar objects, targeting one specific object using the prompt *"Chimney on the right"*. OPT and DIOR-RSVG has several such cases.

Fig. 8: Examples illustrating the limitations of GroundingDINO approach and challenges for transfer case for remote sensing.



(a) REC: Given the prompt *"the car on the left."* and filtering by maximum score incorrectly selects the car on the right.



(b) Object detection with the prompt *"car."* provides better region proposals, which can be further processed to enable accurate REC.

Fig. 9: GroundingDINO faces the same challenge of isolating a single referred object across many images, yet it remains effective as a region proposal mechanism.