Efficient Grounding DINO: Efficient Cross-Modality Fusion and Efficient Label Assignment for Visual Grounding in Remote Sensing

Zibo Hu, Kun Gao, *Member, IEEE*, Xiaodian Zhang, *Member, IEEE*, Zhijia Yang, Mingfeng Cai, Zhenyu Zhu, and Wei Li, *Senior Member, IEEE*

***Abstract*— Visual grounding for remote sensing (RSVG) aims to detect objects in remote sensing scenes based on textual descriptions. While existing methods perform well on RSVG datasets, they are limited to single-object predictions, making them unsuitable for multi-object candidate category datasets. Open-set methods can be applied to both RSVG and candidate datasets, but their use in remote sensing remains rare. To bridge this gap, we introduce the open-set approach to RSVG and propose Efficient Grounding DINO, using Grounding DINO as a baseline. Open-set methods rely on two key modules: cross-modality fusion and label assignment. Existing cross-modality fusion methods simultaneously update text and multi-scale visual features, which hampers the model's ability to generalize under different texts and increases learning complexity. Existing methods predict a single object, allowing direct use as a positive example for loss calculation, while open-set methods for multi-objects require one-to-one matching to assign positive and negative samples. However, background interference in the RSVG datasets causes frequent misassignments, slowing model convergence. We address these issues with two innovations: the multi-scale image-to-text fusion module (MSITFM), which updates text features using self-attention to maintain independence from visual features and employs scale-specific cross-attention for multi-scale visual feature fusion to reduce learning complexity, achieving a 3% parameter and 21.6% GFLOPs reduction. Text confidence matching (TCM) incorporates IoU-based confidence into label assignment to reduce mismatches and enhance model performance. Experiments on DIOR-RSVG, RSVG-HR, and DOTA datasets validate the effectiveness of our approach.**

***Index Terms*— Cross-modality fusion module, misassignment, multi-scale image-to-text fusion module (MSITFM), text confidence matching (TCM), visual grounding for remote sensing (RSVG).**

I. Introduction

RECENTLY, visual grounding for remote sensing (RSVG) has developed rapidly, which aims to locate objects described by natural language in remote sensing images. Candidate category object detection focuses on all objects of some category in an image. Compared with traditional candidate category object detection tasks [1], [2], [3], RSVG can go beyond discrete labels and locate specified objects more flexibly based on natural language containing the topology and distribution of multiple objects and their attributes [4], as shown in Fig. 1(a) and (b). In addition, natural language is consistent with human communication, so RSVG enhances the interactivity of humans and machines, making it have broad application prospects in scenarios such as object localization and recognition [5], traffic monitoring [6], and agricultural monitoring [7]. Some previous works [8], [9], [10] have made considerable progress in RSVG. Sun et al. [8] first introduced the concept of visual grounding in remote sensing images and proposed the RSVG dataset and the GeoVG method. Subsequently, Zhan et al. [9] proposed the DIOR-RSVG dataset and the MGVLF method. Recently, Lan et al. [10] proposed the RSVG-HR dataset and the LQVG method. These works provide a dataset foundation for this field and provide improvement plans, achieving remarkable results. However, these methods predict only a single object, which is sufficient for single-object prediction on the RSVG dataset but inadequate for multi-object prediction on candidate category datasets. Open-set methods can be applied to both the RSVG datasets and candidate category datasets, such as GLIP [11] and Grounding DINO [12]. However, open-set methods are rarely used in RSVG.

In this article, we introduce an open-set method to RSVG and use Grounding DINO [12] as a benchmark. The cross-modality fusion and label assignment strategy are two key modules for achieving visual grounding. Existing cross-modality fusion methods [9], [10], [12], as illustrated in Fig. 1(c), update both text features and multi-scale visual features simultaneously through a fusion strategy. However, since the text is independent of the image, the same object can be described by different texts, and the same text can correspond to different objects in different images. The excessive influence of visual features on text feature updates may impair the model's ability to generalize effectively under different texts. Additionally, visual features at different scales have varying levels of semantic density. Using a single set of parameters to perform interactive learning across multiple scales of visual features increases learning complexity and slows convergence,

Received 2 December 2024; accepted 24 January 2025. Date of publication 29 January 2025; date of current version 13 February 2025. This work was supported by the National Natural Science Foundation of China under Grant U2241275. *(Corresponding authors: Kun Gao; Zhenyu Zhu.)*

Zibo Hu, Kun Gao, Xiaodian Zhang, Zhijia Yang, Mingfeng Cai, and Wei Li are with the School of Optics and Photonics and the Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing Institute of Technology, Beijing 100081, China (e-mail: gaokun@bit.edu.cn).

Zhenyu Zhu is with the Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: zhuzhenyu2021@wust.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3536015

1558-0644 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

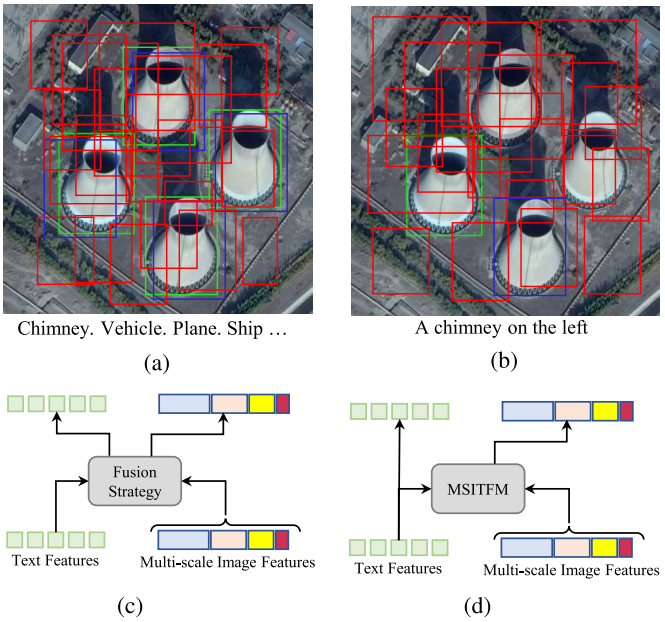


Fig. 1. (a) Label assignment of the open-set method on a candidate category dataset. (b) Misassignment issues of the open-set method on the RSVG dataset. The green boxes represent the ground truth, and the blue/red boxes are positive and negative examples, respectively. The text below the image is the prompt. (c) and (d) Cross-modality fusion strategy of existing methods and the cross-modality fusion strategy proposed in this article. (a) Label assignment for the candidate category dataset. (b) Misassignment for the RSVG dataset. (c) Fusion strategy of existing methods. (d) Our proposed fusion strategy.

which can adversely affect the model's detection performance. Regarding the label assignment strategy, the existing methods [9], [13], [14] only predict a single object, so it can be directly used as a positive example to calculate the loss value. However, the open-set method predicts multiple objects and adopts a one-to-one matching mechanism, so label assignment is required during training to select positive and negative samples to calculate the loss value. Due to background interference and similar object disturbances, the open-set method leads to numerous misassignments in the RSVG dataset, while only a few misassignments occur in the candidate category dataset. These mismatches increase the difficulty of model learning, thus slowing down the model convergence speed.

To solve these issues, we propose Efficient Grounding DINO, which includes two specific components to improve performance. Specifically, we propose the multi-scale image-to-text fusion module (MSITFM) to efficiently fuse multi-scale visual features and text features. In this approach, text features are updated through self-attention to prevent them from being overly influenced by visual features. Additionally, for multi-scale visual features, each scale interacts with text features using a separate set of cross-attention parameters, thereby reducing the learning difficulty associated with features from different scales. Compared with the baseline, the proposed MSITFM not only reduces the parameter count by 3% but also decreases GFLOPs by 21.6%. Moreover, as for the misassignment phenomenon, we adapted the Stable Matching [15] and Rank-DETR [16] from natural image object detection to visual grounding. We proposed text confidence matching (TCM), treating IoU as the confidence between the predicted object and the text description, and integrated it into the bipartite matching process. This effectively reduces misassignments and improves model performance. A substantial number of experiments demonstrate that our method achieves improvements compared with the baseline on the DIOR-RSVG [9] and RSVG-HR [9] while delivering comparable performance on the DOTA [17] dataset.

In summary, the key contributions are as follows.

1) We propose the MSITFM, which enhances the efficiency of multi-scale visual and text feature fusion by employing self-attention for text feature updates and scale-specific cross-attention for image features updates, reducing learning complexity, parameter count by 3%, and GFLOPs by 21.6%.

2) We proposed TCM to reduce the misassignment phenomenon of the open-set method when training the RSVG dataset, which effectively improved the model training efficiency.

3) Our method achieves a significant improvement over the baseline on the RSVG dataset while delivering comparable performance on candidate category datasets.

The remaining chapters of this article are introduced as follows. We briefly introduce related work on visual grounding in Section II. Section III describes our proposed method in detail. Section IV gives detailed comparison experiments and ablation experiments and analyzes the experimental results. Finally, we summarize the work of this article in section V.

II. Related Work

A. Open-Set Object Detection

Open-set object detection generalizes beyond candidate categories by leveraging text descriptions to enable adaptive predictions based on the provided textual content. OV-DETR [18], built on the DETR [19] framework, utilizes image and text embeddings encoded by the CLIP [20] as queries to predict objects. OVR-CNN [21] introduces a novel approach by pretraining the backbone on image captioning data to acquire knowledge from natural language vocabularies. GLIP [11] achieves superior performance by unifying object detection and phrase grounding under a single paradigm and pretraining on large-scale datasets. Grounding-DINO [12] enhances the integration of visual and textual information by performing vision–language modality fusion across multiple stages, enabling more robust grounding capabilities. However, these methods are primarily applied to natural scenes, with limited exploration of open-set methods in the remote sensing domain. This article builds on Grounding DINO [12] to investigate the issues of cross-modality fusion and label assignment in existing open-set object detection methods in the remote sensing domain. We propose the MSITFM and TCM to address these challenges.

B. Cross-Modality Fusion in Visual Grounding

Visual grounding requires the integration of visual and text information, so making their fusion is a critical technology. Current fusion techniques fall into three main categories: contrastive learning, feature interaction, and hybrid approaches

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

combining these two. Contrastive learning [22] operates without direct feature interaction for parameter tuning. It assesses the similarity between potential image region features and text features to rank candidates, thus identifying the object. For example, CLIP [20] jointly encodes images and texts to obtain positive and negative samples and realizes cross-retrieval between images and texts by maximizing the similarity of positive samples and minimizing the similarity of negative samples. Contrastive learning does not directly use feature interaction, but it is easy to accumulate errors in multiple stages. Feature interaction [9], [13], [14] updates parameters of different modalities through feature-level interaction. For example, TransVG [13] uses Transformers to reason about the relationship between text features and visual features within and between modalities and directly regresses the box coordinates to determine the object. However, TransVG [13] uses single-scale visual features to adapt to the scale changes and complex backgrounds of remote sensing images. Zhan et al. [9] proposed the MGVLF module to introduce multi-scale visual features and multi-granularity text embedding to learn more discriminative representations. Feature interaction can effectively improve learning efficiency, but direct regression and coordinates make the model converge more slowly. The method [23], [24], [25] combining contrastive learning and feature interaction can exchange information twice, at the feature level and at prediction time. Grounding DINO [12] uses feature enhancers, language-guided query selection, and cross-modality decoders for feature interaction and uses contrastive loss for classification between predicted objects and language tags. This article focuses on efficient feature interaction between multi-scale features and text features. Although existing methods [9], [10], [12] have achieved good results in the interaction between multi-scale features and text features, there is little research on efficient feature interaction in open-set methods. This article analyzes common feature interaction methods and proposes MSITFM to efficiently interact with multi-scale features and text features.

C. Label Assignment Strategy in Visual Grounding

The label assignment strategy defines positive and negative samples during training. For candidate category object detection, the label assignment strategies are commonly one-to-one strategy and one-to-many strategy. The one-to-one strategy [26] assigns a positive sample to an object during training. For instance, DETR [19] and DETR-like methods [27], [28] implement one-to-one strategy using bipartite graph matching. The one-to-many strategy [29], [30] is to assign one true value to multiple positive examples, such as Oriented RCNN [31] and R3Det [32] maximum assign positive and negative examples through the borrowing threshold. For visual grounding, the common label assignment is one-to-one matching. TransVG [13], TransVG++ [14], and MGVLF [9] only predict one bounding box and use it as a positive example. Glip [11], Grounding DINO [12], and LQVG [10] use bipartite graph matching to achieve a one-to-one assignment. Since the positive example of one-to-one matching is unique, the quality of matching will affect the convergence of the model [33], [34]. There have been some works on improving the quality of one-to-one matching in natural image candidate category object detection, such as Stable Matching [15] and Rank-DETR [16], which improve the matching quality by introducing IoU into the classification cost. However, there are few studies on improving matching quality in the field of RSVG. This article focuses on improving the quality of bipartite graph matching and introduces text confidence into matching based on the ideas of Stable Matching [15] and Rank-DETR [16] to improve matching quality.

III. Method

This section introduces the overall framework and motivation of our method. We first outline the pipeline framework of the overall model in Section III-A. Then, we introduce the data input and feature extraction modules in detail in Section III-B. Section III-C introduces the motivation and design of our proposed MGVLF module. Section III-D introduces the structure of the decoder. Finally, Section III-E introduces our proposed TCM and training loss function.

A. Overview

We introduce the open-set method, Grounding DINO [12], as a benchmark to explore the application of open-set methods in remote sensing. We found that current cross-modality fusion modules and label assignment strategies are inefficient. Therefore, we propose Efficient Grounding DINO to solve the above issues, as shown in Fig. 2. The text encoder and the image encoder map text and image to feature space, respectively (see Section III-B). The cross-modality encoder is used to fuse visual features and text features (see Section III-C). Query generation uses the updated visual and text features to generate cross-modality queries. The cross-modality decoder updates cross-modality queries, and the head network predicts bounding boxes corresponding to phrases from the cross-modality decoder's output (see Section III-D). During training, TCM is used to assign labels, and the corresponding loss is calculated to update parameters (see Section III-E). Our innovations lie in the MSITFM within the cross-modality encoder and TCM. The rest of the design follows Grounding DINO and is briefly introduced.

B. Feature Extraction

For a dataset $D = \{(x_i, t_i, b_i)\}_{i=1}^{N}$ sampled from an unknown joint distribution, where x_i is the image, t_i is the language expression, and b_i is the bounding box corresponding to the language expression. For the candidate category dataset, it is defined as an object label given as an object category such as [car, ship, plane, . . . , harbor]. We concatenate each category together using "." as shown in the input part of Fig. 2 and the following equation:

$$\text{prompt} = \text{“car. ship. plane. } \ldots \text{ harbor”}. \tag{1}$$

We take (x_i, t_i) pair as input and use an image encoder and the text encoder to extract features from x_i and t_i, respectively.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

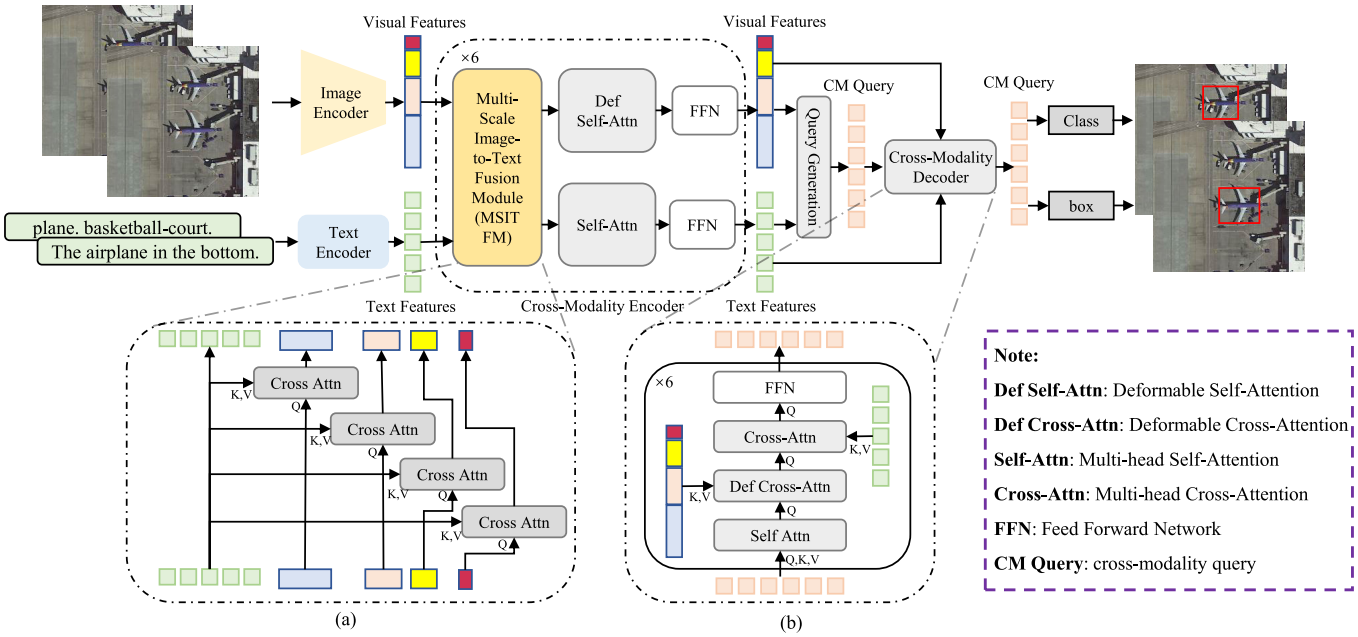


Fig. 2. Illustration of efficient grounding DINO. The detailed explanations of feature extraction, the cross-modality encoder, and cross-modality decoder and prediction in Sections III-B–III-D, respectively. (a) MSITFM. (b) Cross-modality decoder.

1) Image Encoder: For the input image, we use ResNet [35] as the image encoder to extract image features and obtain multi-scale visual features $F_v = \{\boldsymbol{F}_v^i\}_{i=1}^4$ as the output of the image encoder. The downsampling stride corresponding to the F_v multi-scale feature is {8, 16, 32, 64}. Since the dimensions of features at different scales of ResNet are inconsistent, they are unified to dimension $c = 256$ through the 1×1 convolution layer.

2) Text Encoder: For the language description, we use BERT [36] as a text encoder to extract text features. Before inputting into BERT [36], we embed each word into a one-hot embedding vector following Grounding DINO [12] and then convert each one-hot vector into a language token and add [CLS] token and [SEP] token. We input the tokens into BERT to get the text features. The BERT output feature dimension is 768, and the dimension is adjusted to $c = 256$ through the fully connected layer to obtain the text feature F_t.

C. Cross-Modality Encoder

Cross-modality encoder is to efficiently fuse multi-scale visual features (F_v) and text features (F_t). We compared and analyzed common fusion modules and their variants in Section IV-E. Our findings suggest that a more efficient fusion method involves interacting multi-scale visual features with text features separately and updating visual features independently. Text features should be updated without being affected by visual features. Then, we believe that the receptive fields and semantic information of feature layers differ across scales. If multi-scale visual features interact with text features simultaneously, it becomes difficult to understand the interactions. By having each scale's visual features interact with text features separately, we can update each scale's visual features more effectively. Moreover, unlike candidate category detection, the same text description can apply to different images, and the same image can use different text descriptions for the same or different objects for RSVG. If updates to text features are overly influenced by visual features, it can compromise the model's ability to generalize effectively, especially when dealing with diverse or unseen textual descriptions. Therefore, for updating text features, we propose that this should be done independently of visual features.

Therefore, we propose a cross-modality encoder to efficiently fuse visual features and text features and update them, as shown in Fig. 2. It consists of six encoding layers. Each layer consists of two submodules. We will introduce its three submodules in detail later.

1) Multi-Scale Image-to-Text Fusion Module: The goal of MSITFN is to efficiently fuse visual features and text features, as shown in Fig. 2(a). Its submodules perform cross-attention on visual features F_v^i of different scales and text features F_t, respectively, so that visual features of different scales can achieve cross-modality interaction. For visual features $F_v = \{\boldsymbol{F}_v^i\}_{i=1}^4$ and text features F_t, where F_v^i is regarded as a query (Q) and F_t is regarded as a key–value pair (K, V), the cross-attention mechanism is calculated to obtain the fused visual feature $F_v^{i'}$, as shown as follows:

$$\begin{aligned} F_v^{i'} &= \text{Cross Attn}\left(F_v^i, F_t\right) \\ &= F_v^i + \text{Dropout}\left(\text{Attn}\left(\widetilde{F_v^i}, \widetilde{F_t}, F_t\right)\right) \end{aligned} \tag{2}$$

where $\widetilde{F_v^i}$ and $\widetilde{F_t}$ are the sum of F_v^i and F_t and their corresponding positional encoding (PE)

$$\widetilde{F_v^i} = F_v^i + PE\left(F_v^i\right) \tag{3}$$

$$\widetilde{F_t} = F_t + PE(F_t). \tag{4}$$

After MSITFM, we get the visual feature $F_v^{'} = \{\boldsymbol{F}_v^{i'}\}_{i=1}^4$ after the fusion layer.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

2) Self-Attention (Self-Attn) and Feed-Forward Network (FFN): For the fused visual features $F_v^{'}$, we use deformable self-attention (Def Self-Attn) [27] to extract their features. Then, we use the FFN to deepen the features to obtain the visual feature output F_v^u of the cross-modality encoding layer, as shown as follows:

$$\begin{aligned} F_v^u &= \mathrm{LN}\big(\mathrm{FFN}\big(\mathrm{LN}\big(\text{Def Self Attn}\big(F_v'\big)\big)\big)\big) \\ &= \mathrm{LN}\Big(\mathrm{FFN}\Big(\mathrm{LN}\Big(\text{Def Attn}\Big(\widetilde{F_v'}, F_v', p_v\Big)\Big)\Big)\Big) \end{aligned} \tag{5}$$

$$\mathrm{FFN}(F) = F + \mathrm{Linear}(\mathrm{RL}(\mathrm{Linear}(F))) \tag{6}$$

where LN is the layer normalization layer, Linear is the linear layer, RL is the ReLU activation function, and p_v is the reference point corresponding to $F_v^{'}$.

For text features F_t, we use Self-Attention to extract its features, and then use the FFN to obtain the text feature output F_t^u of the cross-modality encoding layer, as shown as follows:

$$\begin{aligned} F_t^u &= \mathrm{LN}(\mathrm{FFN}(\mathrm{LN}(\text{Self Attn}(F_t)))) \\ &= \mathrm{LN}\Big(\mathrm{FFN}\Big(\mathrm{LN}\Big(\mathrm{Attn}\Big(\widetilde{F}_t, \widetilde{F}_t, F_t\Big)\Big)\Big)\Big). \end{aligned} \tag{7}$$

We obtain the visual features F_v^u and text features F_t^u after a single cross-modality encoding layer. The entire encoder consists of 6 identical encoding layers, and the updated visual features F_v^e and text features F_t^e are obtained after six encoding layers.

D. Cross-Modality Decoder and Head Network

The cross-modality decoder is same with Grounding DINO [12], illustrated in Fig. 2(b). It processes three inputs: cross-modality query (CM query), visual features F_v^e, and text features F_t^e. CM query $q \in R^{N\times c}$ is generated by Language-Guided Query Selection [12], where N is the number of CM queries. The entire decoder consists of six identical decoding layers. The CM query first encounters the self-attention unit, then integrates with the visual features through the deformable cross-attention unit, followed by an interaction with the textual features via the cross-attention unit, and culminates in feature enhancement through the FFN.

Same with Grounding DINO [12], the head network consists of a classification head and a regression head. The classification head compares each query with text features using a dot product to determine the score for each text element. The regression head predicts the center and size of an object.

E. TCM and the Loss Function

As illustrated in Fig. 3, we compared different methods of label assignment. Existing approaches, such as TransVG [13] and MGVLF [9], demonstrate effective matching for single-object prediction when trained on the RSVG dataset [see Fig. 3(a)]. However, single-object prediction methods fail to perform effectively on candidate category datasets with multiple objects. Open-set methods, such as Grounding DINO [12], can predict multiple objects and effectively handle both the RSVG datasets and candidate-category datasets. The label assignment during training for these methods is shown in Fig. 3(b) and (c).

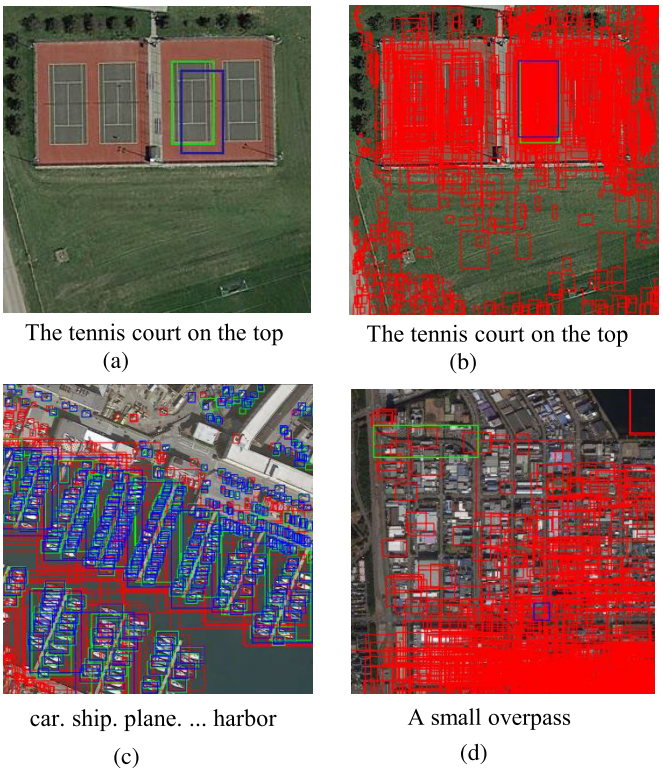


Fig. 3. Label assignment results under different datasets and mechanisms in training. Green boxes are ground-truth values, and blue/red boxes are positive and negative examples, respectively. The text below the image is the prompt. (a) Label assignment results of existing methods such as TransVG [13] on the DIOR-RSVG dataset when predicting a single object. (b) and (c) Label assignment results on the DIOR-RSVG and DOTA datasets when grounding DINO [12] adopts multiple prediction objects, respectively. (d) Misalignment phenomenon. (a) Single-object prediction label assignment for DIOR-RSVG. (b) Multi-object predictions label assignment for DIOR-RSVG. (c) Multiple-object prediction label assignment for DOTA. (d) Misalignment for DIOR-RSVG.

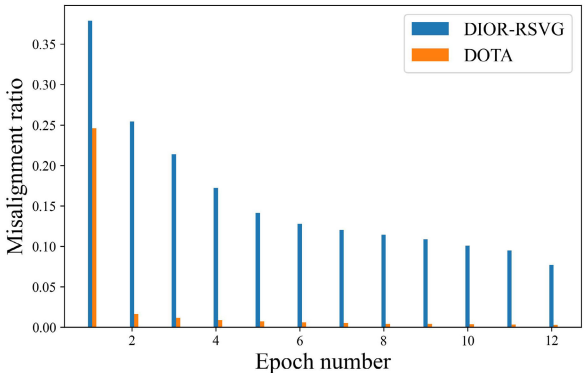


Fig. 4. Comparison of misalignment ratio between the DOTA and DIOR-RSVG datasets. The misalignment ratio indicates the ratio of misaligned samples to all objects in an epoch.

Due to the variability in the number of objects, the maximum number of predicted objects N, is set slightly larger than the maximum number of objects to ensure all objects can be predicted. Under the current one-to-one matching mechanism, this setup results in matched positive examples and unmatched negative examples for accurate loss computation. In contrast,

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

single-object prediction only involves positive examples without any negatives.

Existing works [2], [15] show that the quality of label assignment affects model convergence. To analyze the quality of label assignment in bipartite matching on RSVG and candidate category datasets, we measured the proportion of misassignments during training, defined as the matched prediction box having an IoU of 0 with the ground truth, as shown in Fig. 3(d). We used Grounding DINO [12] to track the misassignment proportions on the DIOR-RSVG [9] and DOTA [17] datasets during training, as shown in Fig. 4. It shows that there is a significant number of misassignments in the RSVG dataset, whereas the candidate category dataset exhibits only a few misassignments. We believe that due to background interference and similar object disturbances, existing matching methods fail to achieve satisfactory alignment in the RSVG dataset.

Grounding DINO [12] employs the Hungarian algorithm [37] to achieve the one-to-one matching, with the set matching cost calculated as follows:

$$L_{\text{match}} = \lambda_1 \cdot \mathcal{C}_{\text{cls}} + \lambda_2 \cdot \mathcal{C}_{L1} + \lambda_3 \cdot \mathcal{C}_{\text{GIOU}}. \tag{8}$$

Here, $\mathcal{C}_{\text{cls}}$ represents the classification cost, $\mathcal{C}_{L1}$ denotes the L1 cost, and $\mathcal{C}_{\text{GIOU}}$ indicates the GIoU cost. The coefficients λ_1, λ_2, and λ_3 correspond to the weights of the respective costs.

Since text information is difficult to understand, the classification cost accounts for a large proportion, and it is easy to be disturbed by the background during matching, resulting in misalignment. Recognizing that IoU effectively reflects textual information between predicted and ground-truth boxes, we introduced IoU into the classification cost to incorporate textual information into the matching process. Specifically, the calculation of the classification cost is expressed as (9), where $f(p)$ represents the treatment of the predicted probability p

$$\begin{aligned}\mathcal{C}_{\text{cls}} = &|1 - f(p)|^{\gamma}\,\text{BCE}(f(p), 1) \\ &- (f(p))^{\gamma}\,\text{BCE}(1 - f(p), 1).\end{aligned} \tag{9}$$

The original form of the classification cost [$f(p)$ in (10)] fails to consider IoU

$$f(p)^{\text{ori}} = p. \tag{10}$$

In this article, the following equation is employed to introduce IoU into the classification cost. Calculating $\mathcal{C}_{\text{cls}}$ using $f(p)^{\text{iou}}$ is referred to as the text confidence classification cost:

$$f(p)^{\text{iou}} = p \cdot \text{IoU}^{\sigma}. \tag{11}$$

This modification ensures that predicted boxes with higher IoU values have higher predicted probabilities, bringing them closer to the interested object, and conversely, those with lower IoU values are assigned probabilities closer to the background.

This article is not the first to introduce IoU into the classification cost. Previous works, such as Stable Matching [15] and Rank-DETR [16], have introduced IoU into the classification cost in candidate category object detection. The main contribution of this part is to discover the high proportion of mismatches and migrate existing methods [15], [16] to visual grounding, and propose TCM.

Therefore, TCM uses the Hungarian algorithm [37] to assign labels. The cost calculation includes text confidence classification cost, L1 cost, and GIoU cost, with weights of 2, 5, and 2, respectively. The loss incorporates both classification and regression components. For classification, Focal Loss [32] is applied, and for regression, Smooth L1 Loss along with Generalized IoU Loss are utilized, with respective weights of 1, 5, and 2.

IV. Experiment

In this section, we conduct extensive experiments to verify the effectiveness of our proposed Efficient Grounding DINO. First, we introduce the datasets and related details from Sections IV-A to IV-C. Then, we show the comparison results of Efficient Grounding DINO with other SOTA methods on different datasets in Section IV-D. Moreover, we design ablation experiments to analyze the effectiveness of each part of Efficient Grounding DINO in Section IV-E. Finally, we give some qualitative results in Section IV-F to supplement the qualitative analysis of the effectiveness of Efficient Grounding DINO.

A. Datasets

To better verify the effectiveness of our method, we used the DIOR-RSVG dataset [9], the RSVG-HR dataset [10], and the DOTA dataset [17].

1) DIOR-RSVG [9]: The DIOR-RSVG dataset released by Northwestern Polytechnical University is obtained by using language expressions for objects based on the DIOR dataset [44]. DIOR-RSVG consists of 17 402 remote sensing images with 800 × 800 image size, including 20 object categories, and describes the objects in them through 38 320 language expressions. The total vocabulary size of this dataset is 100, and the average degree of each expression is 7.47. The entire dataset is divided into the training set, validation set, and testing set. The training set contains 26 991(70%) image–language pairs, the validation set contains 3829(10%) image–language pairs, and the test set contains 7500(20%) image–language pairs. We use the training set and validation set for training, and the test set for testing.

2) RSVG-HR [10]: The RSVG-HR dataset was released by Wuhan University in [10]. This dataset is a redesigned language annotation using high-resolution remote sensing images in the RSVG dataset [8]. Its purpose is to solve the problem that the text description in the RSVG dataset [8] is not clear enough. The image resolution of the RSVG-HR dataset is 1024 × 1024. It contains seven object categories, and the average length of the language description is 19.6 words. The entire dataset consists of 2650 image–language pairs, of which the training set consists of 2151 image–language pairs and the test set consists of the remaining 499 image–language pairs. Due to the small number of datasets, we repeated the training set five times for training.

3) DOTA [17]: The DOTA dataset [17] is a large-scale aerial object detection dataset released by Wuhan University. It contains three versions: DOTAv1, DOTAv1.5, and DOTAv2. We use DOTAv1 as our dataset. The DOTAv1 dataset consists of 2806 aerial images with image sizes ranging from 800 ×

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

Fig. 5. Visual comparison of grounding DINO and our proposed method on the DIOR-RSVG test set.

TABLE I

COMPARISON RESULTS WITH STATE-OF-THE-ART METHODS ON THE DIOR-RSVG TEST SET. * INDICATES THAT GROUNDING DINO AND OUR METHOD USE DATA AUGMENTATION. RED AND BLUE REPRESENT THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN, RESPECTIVELY

Methods	Venue	Visual Enc.	Text Enc.	Params(M)	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
ReSC [23]	ECCV20	DarkNet-53	BERT	179.9	72.71	68.92	63.01	53.70	33.37	64.24	68.10
LBYL-Net [38]	CVPR21	DarkNet-53	LSTM	-	73.29	69.92	63.97	48.07	16.60	65.86	75.45
LBYL-Net [38]	CVPR21	DarkNet-53	BERT	163.8	73.78	69.22	65.56	47.89	15.69	65.92	76.37
TransVG [13]	ICCV21	ResNet-50	BERT	149.7	72.41	67.38	60.05	49.10	27.84	63.56	76.27
VLTVG [39]	CVPR22	ResNet-50	BERT	152.2	69 41	65.16	58.44	46.56	24.37	59.96	71.97
VLTVG [39]	CVPR22	ResNet-101	BERT	-	75.79	72.22	66.33	55.17	33.11	66.32	77.85
QRNet [40]	CVPR22	Swin-S	BERT	178.4	69.77	63.31	54.57	42.82	18.66	59.05	71.69
MGVLF [9]	TGRS23	ResNet-50	BERT	152.5	76.78	72.68	66.74	56.42	35.07	68.04	78.41
LQVG [10]	TGRS24	ResNet50	BERT	166.3	83.41	81.03	75.91	65.52	43.53	74.02	82.22
LPVA [41]	TGRS24	ResNet50	BERT	156.2	82.27	77.44	72.25	60.98	39.55	72.35	85.11
MSAM [42]	GRSL24	ResNet50	BERT	-	73.46	68.42	60.22	47.31	22.33	63.38	76.24
QAMFN [43]	TGRS24	ResNet50	BERT	128.4	81.67	75.15	68.56	57.43	34.51	71.78	84.55
Ours											
Grounding DINO [12]	arXiv23	ResNet50	BERT	172.4	75.99	73.12	67.83	57.57	37.36	67.22	77.71
Grounding DINO* [12]	arXiv23	ResNet50	BERT	172.4	80.27	77.87	72.77	62.81	41.68	71.21	80.49
Efficient Grounding DINO(Ours)	-	ResNet50	BERT	169.3	80.81	77.88	72.45	61.25	40.37	71.41	80.70
Efficient Grounding DINO(Ours)*	-	ResNet50	BERT	169.3	83.05	80.60	75.00	63.56	42.27	73.41	81.06

800 to 4000 $\times$ 4000 pixels. The entire dataset contains 188 282 object annotations. These objects consist of 15 object categories, whose short names can be defined as full names (abbreviations): plane (PL), helicopter (HC), baseball diamond (BD), ship (SH), storage tank (ST), swimming pool (SP), tennis court (TC), harbor (HB), ground track field (GTF), large vehicle (LV), bridge (BR), small vehicle (SV), soccer ball field (SBF), basketball court (BC), and roundabout (RA). DOTAv1 consists of a training set, validation set, and testing set, of which the test set is not publicly labeled and needs to be uploaded to the server for verification. We use the training set for the training and the validation set for testing. The DOTA dataset contains horizontal box annotations and rotated box annotations. We use horizontal box annotations for training and testing.

B. Implementation Details

We developed our method using the MMdetection [45]. Experiments were conducted on 4 NVIDIA 3090 GPUs. We employed the AdamW optimizer [46], starting with a learning rate of 0.0001 and a weight decay of 0.0001, using a batch size of 16 (16 images across 4 GPUs). The training lasted for 12 epochs, with the learning rate halved at the 11th epoch.

The parameter σ in TCM is 2. For the CM query, N is set to 900 during training. For the DIOR-RSVG and RSVG-HR datasets, we use the object with the highest confidence as the prediction result during inference. For the DOTA dataset, we select the 300 objects with the highest confidence as the results during inference.

Unless stated otherwise, we applied no data augmentation to DIOR-RSVG and RSVG-HR datasets, except for random flipping in DOTA. For DIOR-RSVG, we adhered to MGVLF's setup [9], using a 640 $\times$ 640 image size for training and inference. RSVG-HR and DOTA used a 1024 $\times$ 1024 image size for both training and testing. Augmentation was only used for DIOR-RSVG and RSVG-HR, where image sizes were randomly set to one of several values: [480, 560, 640, 720, 800] for DIOR-RSVG and [768, 840, 920, 1024] for RSVG-HR.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

Fig. 6. Visual comparison of grounding DINO and our proposed method on the RSVG-HR test set.

TABLE II

COMPARISON RESULTS WITH STATE-OF-THE-ART METHODS ON THE RSVG-HR TEST SET. * INDICATES THAT GROUNDING DINO AND OUR METHOD USE DATA AUGMENTATION. FINETUNE INDICATES FINE-TUNING BASED ON THE BEST MODEL OF OUR METHOD ON THE DIOR-RSVG DATASET. RED AND BLUE REPRESENT THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN, RESPECTIVELY

Methods	Venue	Visual Enc.	Text Enc.	Params(M)	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
TransVG [13]	ICCV21	ResNet50	BERT	149.7	46.26	34.76	19.05	5.63	0.30	40.72	30.47
LBYL-Net [38]	CVPR21	DarkNet-53	BERT	163.8	48.17	35.84	20.33	5.46	0.30	41.57	30.88
VLTVG [39]	CVPR22	ResNet50	BERT	152.2	48.76	36.33	21.43	7.43	0.50	42.06	31.92
QRNet [40]	CVPR22	Swin-S	BERT	178.4	50.35	37.76	21.77	7.92	0.50	42.87	32.74
MGVLF [9]	TGRS23	ResNet50	BERT	152.5	50.70	38.88	23.45	8.22	0.60	43.11	33.23
LQVG [10]	TGRS24	ResNet50	BERT	166.3	87.37	82.57	72.34	51.10	14.43	71.55	70.06
Ours											
Grounding DINO [12]	arXiv23	ResNet50	BERT	172.4	60.52	55.31	45.09	28.46	8.02	49.51	38.89
Grounding DINO* [12]	arXiv23	ResNet50	BERT	172.4	66.13	59.12	49.70	31.06	7.82	54.73	43.50
Grounding DINO*(Finetune) [12]	arXiv23	ResNet50	BERT	172.4	79.96	74.15	65.33	45.29	14.83	65.45	57.39
Efficient Grounding DINO(Ours)	-	ResNet50	BERT	169.3	67.33	60.72	49.90	31.66	6.21	53.33	44.92
Efficient Grounding DINO(Ours)*	-	ResNet50	BERT	169.3	74.55	67.94	56.11	36.67	8.62	60.53	52.33
Efficient Grounding DINO(Ours)*(Finetune)	-	ResNet50	BERT	169.3	85.77	80.76	68.94	49.30	15.03	70.89	69.82

C. Evaluation Metrics

For the DIOR-RSVG and RSVG-HR datasets, we used the same evaluation metrics as in MGVLF [9], which include Pr@0.5 to Pr@0.9, meanIoU, and cumIoU. These metrics measure the precision of the IoU at different thresholds (0.5 to 0.9). MeanIoU and cumIoU are defined as follows:

$$\text{meanIoU} = \frac{1}{M}\sum_{t} I_t / U_t \tag{12}$$

$$\text{cumIoU} = \left(\sum_{t} I_t\right) / \left(\sum_{t} U_t\right) \tag{13}$$

where t is the index for each image–text pair and M is the total number of test samples. U_t and I_t denote the union and intersection of the predicted and actual bounding box areas, respectively.

For the DOTA dataset, we use mAP with an IoU threshold of 0.5 as the evaluation metric.

D. Comparison With State-of-the-Art Methods

To verify the effectiveness of our method, we compared it with some state-of-the-art methods on the DIOR-RSVG test set, the RSVG-HR test set, and the DOTA validation set.

1) DIOR-RSVG: The comparison results of the DIOR-RSVG test set are presented in Table I. Compared with the benchmark Grounding DINO, our proposed method demonstrates improvements across various metrics with or without data augmentation. Visual comparisons between our method and Grounding DINO are illustrated in Fig. 5, highlighting our method's capability to predict more accurate bounding boxes and detect objects that Grounding DINO fails to identify.

Compared with existing methods, the results of our method are second only to LQVG [10]. Compared with LQVG [10], which requires 70 epochs of training, our method only needs 12 epochs of training to achieve competitive results. The comparative results on the DIOR-RSVG dataset show that our method is effective and comparable.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

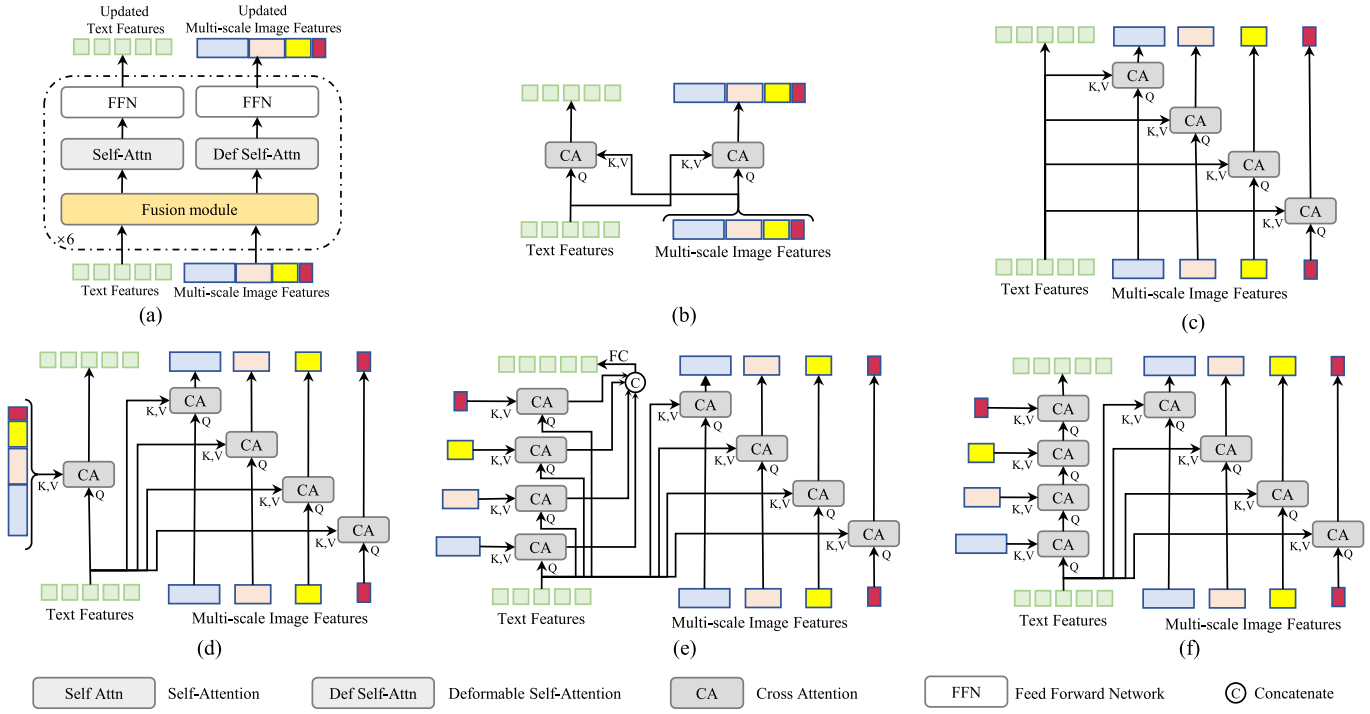


Fig. 7. Comparison of different fusion modules. (a) Schematic of the cross-modality encoding layer. (b)–(f) Various fusion modules in (a). (b) Feature enhancer in grounding DINO [12]. (c) MSITFM proposed in this article. (d) MSVCTFM. (e) MSVMLTFM. (f) MSCMAM in LQVG [10].

TABLE III

COMPARISON RESULTS ON THE DOTA-v1.0 VALIDATION SET UNDER DIFFERENT METHODS. RED AND BLUE REPRESENT THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN, RESPECTIVELY

Method	Backbone	Params(M)	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
CNN-Based																		
RetinaNet [32]	ResNet50	36.62	87.00	58.10	36.90	25.40	43.90	68.80	58.60	88.90	52.30	55.60	38.70	56.50	68.80	50.60	9.70	53.32
Faster R-CNN [47]	ResNet50	41.42	89.50	72.30	47.80	67.70	47.30	77.60	61.10	90.70	62.00	62.00	60.90	61.40	75.20	53.30	55.20	65.60
Dynamic RCNN [48]	ResNet50	41.42	89.00	73.10	43.80	64.70	45.70	76.80	60.50	90.80	61.00	60.30	55.60	61.20	67.70	51.00	56.30	63.83
Transformer-Based																		
Deformable DETR [27]	ResNet50	40.36	87.50	62.30	44.20	55.10	55.20	74.60	77.70	89.00	54.00	59.50	50.60	52.40	70.80	49.20	58.00	62.67
Dab-DETR [49]	ResNet50	43.71	85.30	61.30	32.80	46.10	32.10	62.90	51.10	87.60	55.90	47.70	39.80	52.30	65.10	50.20	55.30	55.03
DINO [28]	ResNet50	47.57	85.00	61.00	41.20	44.40	64.80	77.70	83.40	86.70	45.70	71.60	47.10	55.00	70.40	54.90	52.70	62.77
Ours																		
Grounding DINO [12]	ResNet50	172.4	84.30	63.00	40.40	50.40	67.20	78.00	83.90	87.80	40.80	72.30	44.60	52.00	69.50	55.10	43.30	62.17
Efficient Grounding DINO(Ours)	ResNet50	169.3	84.40	62.70	36.20	48.40	63.80	78.30	84.30	87.20	49.80	71.60	51.80	49.20	68.90	56.50	46.30	62.63

2) RSVG-HR: We evaluate performance on the RSVG-HR test set, with results summarized in Table II. Compared with the baseline Grounding DINO, we analyze performance across direct training, training with data augmentation, and fine-tuning. Across all scenarios, our method demonstrates significant improvements. For instance, Pr@0.5 improves by 6.81%, 8.42%, and 5.81%, respectively, under direct training, training with data augmentation, and fine-tuning. Visual comparisons in Fig. 6 illustrate that our method predicts more accurate bounding boxes and identifies objects that Grounding DINO may miss, even when faced with complex text descriptions.

The results of existing methods in Table II are adapted from [10], which initially pretrains on the DIOR-RSVG training set followed by fine-tuning on the RSVG-HR training set. Our method achieves results second only to LQVG [10] overall and attains state-of-the-art performance in Pr@0.9. Notably, our approach outperforms most methods that require fine-tuning, even without fine-tuning. Comparative results on the RSVG-HR dataset demonstrate the effectiveness and competitiveness of our method.

3) DOTA: In addition to the RSVG dataset, we also evaluated the effectiveness of our method on the candidate category dataset DOTA. Comparison results are presented in Table III. The results show that on the DOTA dataset, the CNN-based method performs better than the Transformer-based method, and the proposed method in this article achieves comparable results with the Transformer method. Our method improves mAP by 0.46% compared with the baseline. However, unlike the significant improvements observed on the RSVG datasets, our method exhibits relatively modest enhancements on the candidate category dataset. We attribute this to two main factors: 1) the fixed input text focuses on objects across all categories, potentially weakening the role of our fusion module and 2) as illustrated in Fig. 4, the DOTA dataset features a very low misalignment rate, diminishing the impact of our proposed text-confidence matching. Therefore, our method shows slight improvements in DOTA.

E. Ablation Study

1) Comparison With Multi-Scale Fusion Module: For the fusion module, we compared different model structures to

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON RESULTS OF DIFFERENT FUSION MODULES ON THE DIOR-RSVG DATASET. FEATURE ENHANCER IS THE STRUCTURE IN FIG. 7(b), MSCMAM IS THE STRUCTURE IN FIG. 7(f), MSVMLTFM IS THE STRUCTURE IN FIG. 7(e), MSVCTFM IS THE STRUCTURE IN FIG. 7(d), AND MSITFM IS THE STRUCTURE IN FIG. 7(c). RED AND BLUE REPRESENT THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN, RESPECTIVELY

Method	Pr@0.5	Pr@0.6	Pr@0.7	meanIoU	cumIoU
Feature Enhancer [12]	75.99	73.12	67.83	67.22	77.71
MSCMAM [10]	77.69	74.16	67.96	67.94	77.98
MSVMLTFM	78.83	75.65	70.29	69.75	79.45
MSVCTFM	79.12	75.60	69.87	69.68	78.74
MSITFM	80.27	77.11	71.43	70.96	80.18

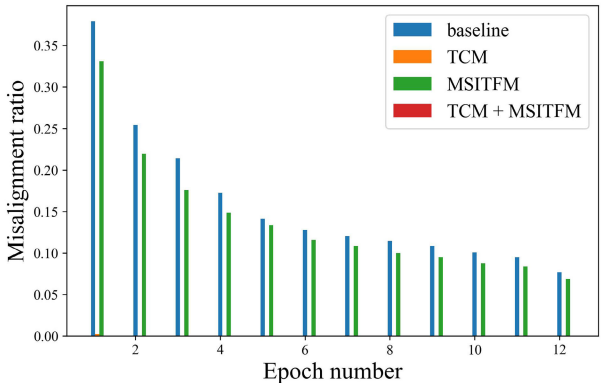


Fig. 8. Comparison of misalignment rates among different methods on the DIOR-RSVG dataset. Baseline represents the Grounding DINO, TCM represents the use of text confidence matching, MSITFM represents the use of a multi-scale image-to-text fusion module, and TCM+MSITFM is the combination of TCM and MSITFM.

explore the state of text information and image information when they are fused. We compared five fusion structures, as shown in Fig. 7, namely the feature enhancer structure in Grounding DINO [12] in Fig. 7(b), the MSITFM proposed in this article in Fig. 7(c), the Multi-scale visual–cross-text fusion module (MSVCTFM) in Fig. 7(d), the Multi-scale vision and multilevel text fusion module (MSVMLTFM) in Fig. 7(e), and the Multi-scale Cross-Modality Alignment Module (MSCMAM) in LQVG [10] in Fig. 7(f). We conducted comparative experiments on the five structures on the DIOR-RSVG dataset, and the results are shown in Table IV.

Comparing Fig. 7(b) and (d), breaking down features at different scales and applying cross-attention to text features can notably improve the model's performance. Looking at Fig. 7(c) and (d), updating text features without relying on visual features also enhances the model's effectiveness. However, comparing Fig. 7(d)–(f), updating text features too closely with visual ones can reduce the model's detection ability, with the cascade structure in Fig. 7(f) leading to the worst results.

In summary, the MSITFM described in this article works best when combining text features with multi-scale visual features. The optimal performance is achieved when text features are updated independently and when each scale of visual features interacts with text features on its own.

In evaluating the aforementioned structures across different performances, we attribute their performance mainly to two factors:

a) Independence of text from images: Text often functions as retrieval information and tends to be relatively independent of specific images. Specifically, the same object can be described in different contexts, and similarly, the same text can correspond to different objects in different images. Therefore, if updates to text features are overly influenced by visual features, it can compromise the model's ability to generalize effectively, especially when dealing with diverse or unseen textual descriptions. To validate that updating text features independently of visual features enhances the model's generalization ability, we trained our model on the DIOR-RSVG training set and tested it on both the DIOR-RSVG and RSVG-HR test sets, as shown in Table V. The RSVG-HR dataset features long textual descriptions, contrasting with the short descriptions in DIOR-RSVG, offering a meaningful test of the model's generalization capability. The results in Table V demonstrate that the proposed architecture significantly improves the model's generalization ability, achieving a 10.42% increase in Pr@0.5 on an unseen dataset compared to the baseline.

b) Visual features at different scales exhibit varying concentrations of semantic information: Visual features at varying scales possess distinct semantic information concentrations due to differences in their scales, receptive fields, and semantic content, as shown in Fig. 9. Concurrent interaction of text features with visual features across multiple scales increases learning complexity and slows convergence speed, thereby impacting the model's detection capabilities.

Based on these observations, we recommend that when designing fusion networks, text features should be updated independent of visual features to increase the generalization ability of the model under diverse texts. Moreover, for multi-scale visual features, each scale should undergo separate updates while interacting with text features to optimize performance and maintain efficiency.

2) Ablation Study for TCM: Regarding the hyperparameter σ for TCM, we conducted a comparison of its effects using different σ, as shown in Table VI. Across four different parameter settings, we observed similar outcomes, leading us to select $\sigma = 2$ as a representative value.

To analyze the effect of TCM, we compared the misassignment ratios across different components, depicted in Fig. 8. The results in Fig. 8 indicate that using TCM alone significantly reduces the misassignment rate, although some residual misassignments initially remain. Similarly, using MSITFM alone slightly decreases the misassignment rate. However, combining the two methods almost completely eliminates misassignments. This observation strongly validates the effectiveness of our proposed approach.

To explore why TCM is effective, we examined the cost composition of the existing bipartite matching method and found that the classification branch contributes disproportionately to the overall cost. Consequently, compared with TCM, we utilize a low weight of the classification cost. The results, presented in Table VII, show that both methods yield similar improvements. Thus, we believe that TCM enhances model

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

TABLE V

COMPARISON RESULTS OF TRAINING ON THE DIOR-RSVG TRAINING SET AND TESTING ON THE DIOR-RSVG AND RSVG-HR TEST SETS

Method	Train Dataset	Test Dataset	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
Grounding DINO [12]	DIOR-RSVG	DIOR-RSVG	75.99	73.12	67.83	57.57	37.36	67.22	77.71
	DIOR-RSVG	RSVG-HR	12.02	9.82	8.02	5.61	2.40	12.07	7.93
Grounding DINO + MSITFM	DIOR-RSVG	DIOR-RSVG	80.27(+4.28)	77.11(+3.99)	71.43(+3.60)	60.47(+2.90)	40.36(+3.00)	70.96(+3.74)	80.18(+2.47)
	DIOR-RSVG	RSVG-HR	22.44(+10.42)	18.84(+9.02)	14.23(+6.21)	10.02(+4.41)	3.01(+0.60)	21.73(+9.66)	14.58(+6.65)

TABLE VI

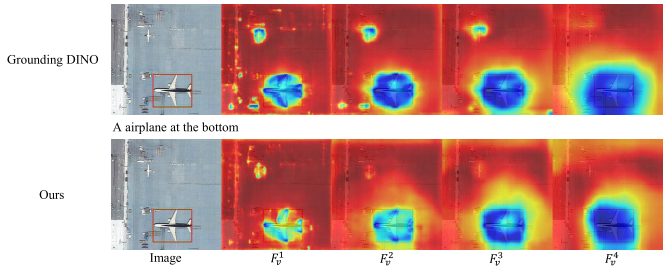
COMPARISON RESULTS AT DIFFERENT WEIGHT OF CLASSIFICATION COST IN BIPARTITE MATCHING. RED AND BLUE REPRESENT THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN, RESPECTIVELY

σ	Pr@0.5	Pr@0.6	Pr@0.7	meanIoU	cumIoU
1	80.51	77.31	71.96	71.06	80.54
2	80.81	77.88	72.45	71.41	80.70
3	80.84	78.13	72.56	71.47	80.43
4	80.84	77.71	72.60	71.25	80.57

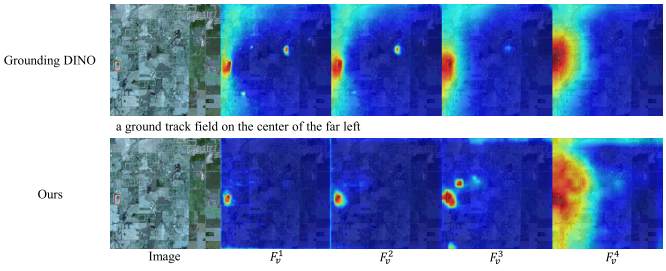
TABLE VII

COMPARISON RESULTS AT DIFFERENT WEIGHT OF CLASSIFICATION COST IN BIPARTITE MATCHING. THE BEST RESULTS IN EACH COLUMN ARE INDICATED IN BOLD

Method	Pr@0.5	Pr@0.6	Pr@0.7	meanIoU	cumIoU
TCM	78.16	75.21	**70.19**	69.03	**79.77**
Low weight 0.5	**78.31**	**75.69**	70.08	**69.28**	79.23



(a)

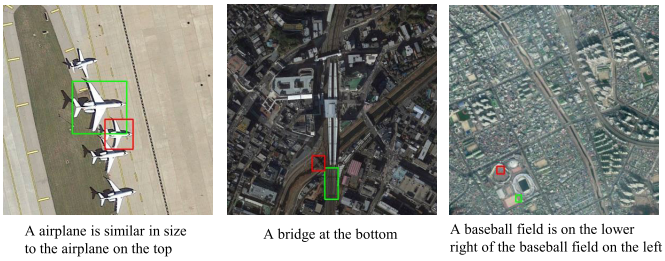


(b)

Fig. 9. Visualization results of Eigen-CAM [51] on the DIOR-RSVG test set and RSVG-HR test set. The green/red boxes are the prediction/ground-truth boxes. F_v^i represents different feature layers. As i increases, the size of the feature layers decreases. Text description in the middle of the image predicted by different methods. (a) Visualization results on the DIOR-RSVG test set. (b) Visualization results on the RSVG-HR test set.

performance primarily by dynamically adjusting weights, thereby achieving more efficient matching.

3) Ablation Study for MSITFM and TCM: We evaluated the effectiveness of the two proposed components on the DIOR-RSVG dataset, as detailed in Table VIII. Both MSITFM and TCM significantly enhance model performance individually, with their combination yielding the best results. Notably, MSITFM predominantly contributes to performance gains.



(a)

(b)

Fig. 10. Failure of our method on the (a) DIOR-RSVG test set and (b) RSVG-HR test set. The green box is the predicted box and the red box is the grounding truth.

Additionally, we found that MSITFM cuts down the model's parameters by 1.8%, while TCM keeps the parameter count unchanged. This means our method successfully reduces the model's parameters by 1.8%. And the GFLOPs are reduced by 21.6% in the case of input images (640 640). It proves that our method can not only improve performance but also reduce inference time.

F. Qualitative Results

During inference, we utilized Eigen-CAM [51] to display the features from the final fusion module in the cross-modality encoder, as illustrated in Fig. 9(a) and (b). Our method shows a better focus on the object areas described by the text postfusion, compared to Grounding DINO, with less attention to irrelevant areas.

In addition, we also analyzed the objects that our method failed to detect during inference, as shown in Fig. 10(a) and (b). We found that most of the missed objects belong to the following categories: 1) color recognition failure; 2) relative size recognition failure; 3) background interference; and 4) relative position recognition failure. This means that our method has limitations and needs to be further studied and alleviated.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII

COMPARISON RESULTS OF THE EFFECTIVENESS OF DIFFERENT COMPONENTS ON THE DIOR-RSVG DATASET. MSITFM REPRESENTS THE USE OF A MULTI-SCALE IMAGE-TO-TEXT FUSION MODULE AND TCM REPRESENTS THE USE OF TCM. RED AND BLUE REPRESENT THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN, RESPECTIVELY. NOTE: MMDETECTION UTILIZES FAIRSCALE [50] TO ENHANCE EFFICIENCY. FOR A FAIR COMPARISON, FAIRSCALE [50] IS NOT USED IN THE GFLOPS STATISTICS

MSITFM	TCM	Input shape	Params(M)	GFLOPs(G)	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	meanIoU	cumIoU
✗	✗	(640,640)	172.4	162.4	75.99	73.12	67.83	57.57	37.36	67.22	77.71
✓	✗	(640,640)	169.3	127.4	80.27	77.11	71.43	60.47	40.36	70.96	80.18
✗	✓	(640,640)	172.4	162.4	78.16	75.21	70.19	59.48	39.33	69.03	79.77
✓	✓	(640,640)	169.3	127.4	80.81	77.88	72.45	61.25	40.37	71.41	80.70

V. CONCLUSION

In this article, we propose an open-set-based method, Efficient Grounding DINO, for visual grounding in remote sensing images. Our approach overcomes the limitations of existing methods on multi-object candidate category datasets. It addresses two key challenges in open-set methods: cross-modality fusion and label assignment. To improve cross-modality fusion, we developed the MSITFM, which employs self-attention to independently update text features and utilizes scale-specific cross-attention for multi-scale visual feature fusion. This design reduces learning complexity, achieving a 3% reduction in parameters and a 21.6% reduction in GFLOPs. For label assignment, we introduced TCM, incorporating IoU-based confidence into the bipartite matching process to minimize mismatches and enhance performance. Extensive experiments on the DIOR-RSVG, RSVG-HR, and DOTA datasets validate the effectiveness of our method, demonstrating its potential to bridge the gap between RSVG tasks and open-set multi-object detection.

Limitations: Our method has a good improvement on the RSVG dataset, but the improvement on the candidate category dataset is limited. Although the number of parameters of our method is reduced compared to the baseline, it is still large compared to the simple object detection model. In future work, we will try to achieve comprehensive improvements on the RSVG dataset and candidate category dataset, on the one hand, and, on the other hand, we will try to reduce model parameters to accelerate model training and reasoning.

REFERENCES

[1] G. Sahbeni, M. Ngabire, P. K. Musyimi, and B. Székely, "Challenges and opportunities in remote sensing for soil salinization mapping and monitoring: A review," *Remote Sens.*, vol. 15, no. 10, p. 2540, May 2023.

[2] Z. Hu et al., "EMO2-DETR: Efficient-matching oriented object detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616814.

[3] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, "AO2-DETR: Arbitrary-oriented object detection transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2342–2356, May 2023.

[4] L. Bashmal, Y. Bazi, F. Melgani, M. M. Al Rahhal, and M. A. Al Zuair, "Language integration in remote sensing: Tasks, datasets, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 4, pp. 63–93, Dec. 2023.

[5] F. Liu et al., "RemoteCLIP: A vision language foundation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5622216.

[6] X. Lu and Q. Weng, "Multi-LoRA fine-tuned segment anything model for urban man-made object extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5637519.

[7] Y. Xu, T. Bai, W. Yu, S. Chang, P. M. Atkinson, and P. Ghamisi, "AI security for geoscience and remote sensing: Challenges and future trends," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 2, pp. 60–85, Jun. 2023.

[8] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang, and X. Huang, "Visual grounding in remote sensing images," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 404–412.

[9] Y. Zhan, Z. Xiong, and Y. Yuan, "RSVG: Exploring data and models for visual grounding on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604513.

[10] 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 Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5626513.

[11] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10965–10975.

[12] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Nov. 2024, pp. 38–55.

[13] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "TransVG: End-to-end visual grounding with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1769–1779.

[14] J. Deng et al., "TransVG++: End-to-end visual grounding with language conditioned vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13636–13652, Nov. 2023.

[15] S. Liu et al., "Detection transformer with stable matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6468–6477.

[16] Y. Pu et al., "Rank-DETR for high quality object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Jan. 2024, pp. 1–14.

[17] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[18] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary DETR with conditional matching," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 106–122.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Switzerland: Springer, 2020, pp. 213–229.

[20] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[21] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14393–14402.

[22] H. Jiang, Y. Lin, D. Han, S. Song, and G. Huang, "Pseudo-Q: Generating pseudo language queries for visual grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 15513–15523, Jun. 2022.

[23] Z. Yang, T. Chen, L. Wang, and J. Luo, "Improving one-stage visual grounding by recursive sub-query construction," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 387–404.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

[24] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR–modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1760–1770.

[25] F. Shi, R. Gao, W. Huang, and L. Wang, "Dynamic MDETR: A dynamic multimodal transformer decoder for visual grounding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1181–1198, Feb. 2024.

[26] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng, "End-to-end object detection with fully convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15849–15858.

[27] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, Oct. 2020, pp. 1–16.

[28] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[29] Z. Zhao and S. Li, "OASL: Orientation-aware adaptive sampling learning for arbitrary oriented object detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 128, Apr. 2024, Art. no. 103740.

[30] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[31] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3520–3529.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[33] Q. Chen et al., "Group DETR: Fast DETR training with group-wise one-to-many assignment," 2022, *arXiv:2207.13085*.

[34] D. Jia et al., "DETRs with hybrid matching," 2022, *arXiv:2207.13080*.

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

[36] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[37] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, pp. 83–97, Mar. 1955.

[38] B. Huang, D. Lian, W. Luo, and S. Gao, "Look before you leap: Learning landmark features for one-stage visual grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16883–16892.

[39] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9489–9498.

[40] J. Ye et al., "Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15481–15491.

[41] K. Li, D. Wang, H. Xu, H. Zhong, and C. Wang, "Language-guided progressive attention for visual grounding in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5631413.

[42] F. Wang, C. Wu, J. Wu, L. Wang, and C. Li, "Multistage synergistic aggregation network for remote sensing visual grounding," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[43] C. Li et al., "Injecting linguistic into visual backbone: Query-aware multimodal fusion network for remote sensing visual grounding," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5637814.

[44] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[45] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2017, pp. 1–18.

[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.

[48] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Nov. 2020, pp. 260–275.

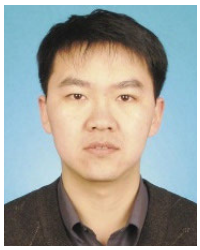
[49] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–19.

[50] FairScale. (2021). *Fairscale: A General Purpose Modular PyTorch Library for High Performance and Large Scale Training*. [Online]. Available: https://github.com/facebookresearch/fairscale

[51] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

Zibo Hu is currently pursuing the Ph.D. degree with the College of Optics and Photonics, Beijing Institute of Technology, Beijing, China.

His research interests include remote sensing image scene object detection, noise robustness, instance segmentation, and machine learning.



Kun Gao (Member, IEEE) received the B.E. degree in electrical engineering and the Ph.D. degree in instrument science and engineering from Zhejiang University, Hangzhou, China, in 1995 and 2002, respectively.

From 2002 to 2004, he was a Post-Doctoral Fellow at Tsinghua University, Beijing, China. Since 2005, he has been with Beijing Institute of Technology, Beijing, working on infrared technology and real-time image processing.

Dr. Gao is a member of the Optical Society of China.

Xiaodian Zhang (Member, IEEE) received the B.E. degree in optoelectronics information science and engineering from Beijing Institute of Technology, Beijing, China, in 2018, where he is currently pursuing the M.S. degree in optical engineering with the School of Optoelectronics. His research interests include deep learning and object detection in natural and remote sensing scenes.

Zhijia Yang received the B.E. degree from the School of Optics, Beijing Institute of Technology, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Photoelectronic Imaging Technology and Systems.

His research interests include image fusion, image stitching, and deep learning.

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.

Mingfeng Cai is currently pursuing the master's degree with Beijing Institute of Technology, Beijing, China.

His research interests include polarization imaging and Fourier light field microscopy.

Zhenyu Zhu received the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2020.

From 2021 to 2024, he was Post-Doctoral Fellow at Tsinghua University, Beijing. Since 2025, he has been with Wuhan University of Science and Technology, Wuhan, China, working on multisource information fusion and machine vision.

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

Subsequently, he spent one year as a Post-Doctoral Researcher at the University of California at Davis, Davis, CA, USA. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include hyperspectral image analysis, pattern recognition, and target detection.

Dr. Li received the JSTARS Best Reviewer in 2016, the TGRS Best Reviewer Award in 2020 from the IEEE Geoscience and Remote Sensing Society (GRSS) and the Outstanding Paper Award at the IEEE International Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers) in 2019. He has served as an Associate Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) and IEEE SIGNAL PROCESSING LETTERS (SPL). He is also serving as an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS).

Authorized licensed use limited to: UNIVERSITY OF ALBERTA. Downloaded on September 04,2025 at 03:30:46 UTC from IEEE Xplore. Restrictions apply.