Efficient Adaptation of Remote Sensing Visual Grounding

Hasan Moughnieh(1), Mohamad Chalhoub(2), Hasan Nasrallah(3),
Cristiano Nattero(4), Paolo Campanella(4), Giovanni Nico(5) and Ali J. Ghandour(6)

(1) American University of Beirut, Beirut, Lebanon
(2) Lebanese University, Beirut, Lebanon
(3) RADIS sarl, Beirut, Lebanon,
(4) WASDI sàrl, Dudelange, Luxembourg
(5) Institute for Applied Mathematics, National Research Council, Bari, Italy
(6) National Center for Remote Sensing, CNRS, Beirut, Lebanon

Introduction

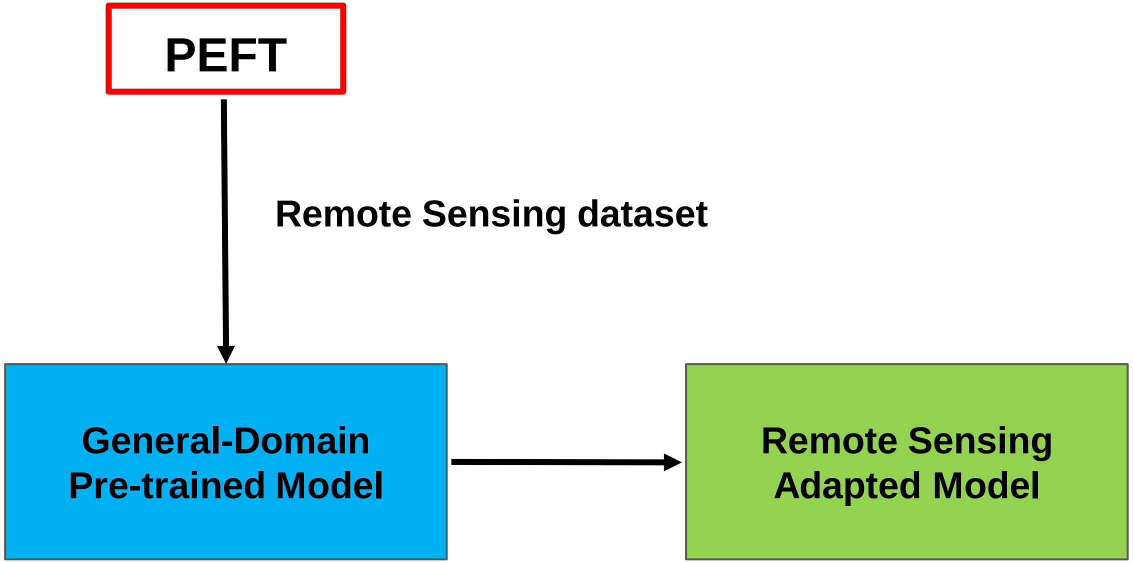
Visual grounding is a vision-language task that localizes image regions corresponding to textual descriptions.

Challenges & Limitations

- **Domain Gap**: Natural image models don't generalize to remote sensing
- **Few Labels**: Remote sensing datasets are limited
- **Dense Scenes**: Overlapping and small objects
- **Visual Variability**: Weather, lighting, and resolution changes
- **Vague Text**: Ambiguous or unclear queries
- **High Cost**: Large image sizes increase compute
- **Poor Transferability**: Weak generalization to new areas

Objective

Adapt general-domain vision-language models to remote sensing visual grounding using parameter-efficient fine-tuning (PEFT), achieving high performance with minimal computational cost.



Parameter Efficient Fine-Tuning Techniques (PEFT)

PEFT techniques fine-tune only a small subset of model parameters, reducing computational cost while maintaining performance.

- **LoRA (Low-Rank Adaptation)**:

  Injects trainable low-rank matrices into attention layers → dramatically reduces the number of updated parameters.

- **BitFit**:

  Only bias terms are fine-tuned → extremely lightweight and effective in large language and vision models.

- **Adapters**:

  Small bottleneck modules inserted between transformer layers → allow fast adaptation without modifying backbone weights.

Pre-Trained Models for PEFT Adaptation

Grounding DINO

- A strong **vision-language model** designed for open-set object detection and **visual grounding**.
- Built with **two separate encoders** (for image and text) followed by a **cross-modal decoder**, enabling fine-grained grounding of textual queries in images.
- Excels in complex grounding scenarios and is robust across various domains.

One-For-ALL (OFA)

- A **unified multimodal framework** that handles multiple tasks (e.g., captioning, grounding, VQA) within a single model.
- Trained using instruction tuning, making it **highly adaptable** across modalities.
- Used here for its **generalization capabilities** and **flexibility** in visual-language alignment.

Remote Sensing datasets

Two vision-language remote sensing datasets have been used:

- **DIOR-RSVG**
- **OPT-RSVG**

Spatial resolution: from 0.15 m to 30 m

DIOR-RSVG

Train	Val	Test
26.991	3.829	7.500

38.320 samples

OPT-RSVG

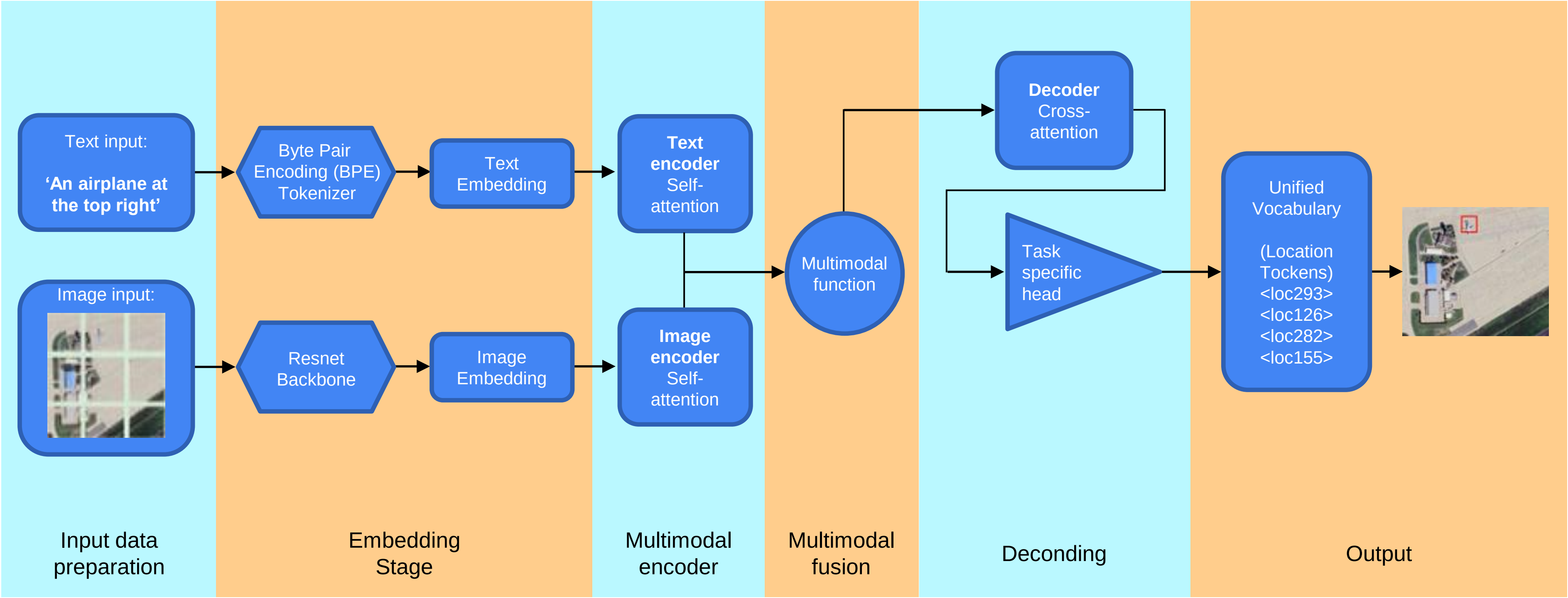
Train	Val	Test
19.500	4.895	24.477

48.952 samples

Performance Metrics

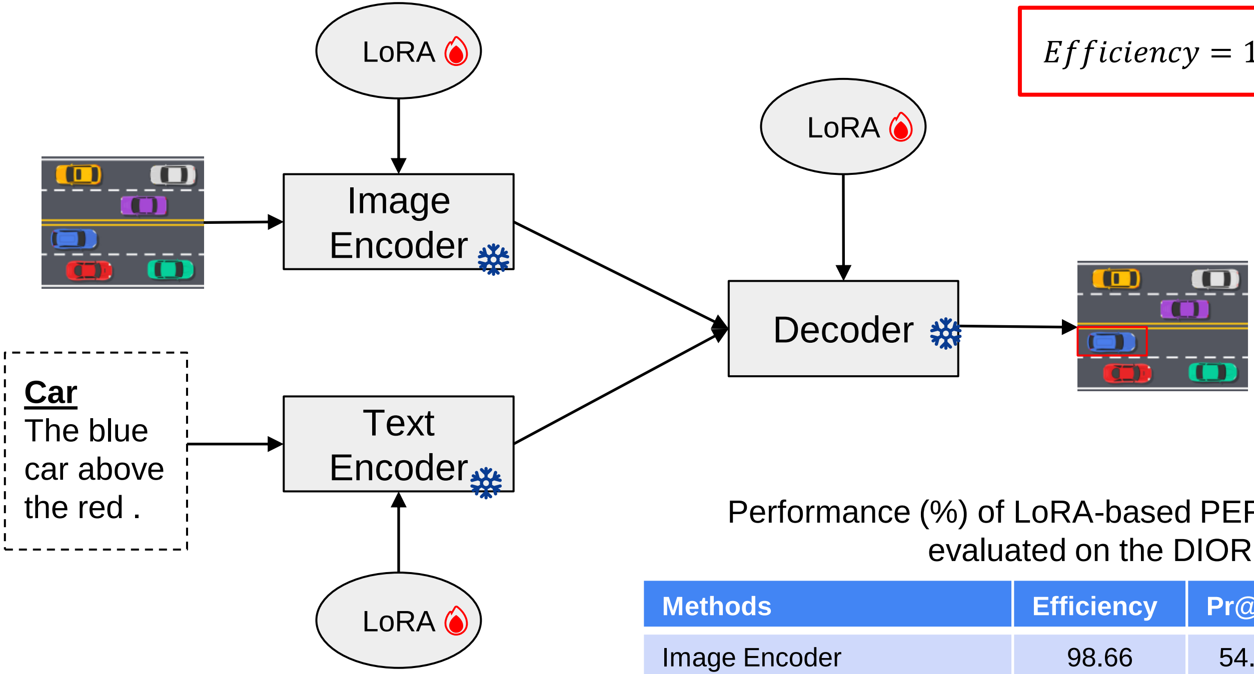
- Pr@0.5 / Pr@0.7 / Pr@0.9: Precision at IoU thresholds 0.5, 0.7, 0.9
- meanIoU: Mean Intersection over Union across all samples
- cumIoU: Cumulative IoU across image-query pairs

Adaptation of OFA Model



Simplified Architecture of OFA focusing on VG part

LoRA Placement Strategy In Grounding DINO



$$Efficiency = 100 - \left(\frac{Trainable\ Parameters\ PEFT}{Total\ Model\ Parameters} \cdot 100 \right)$$

Performance (%) of LoRA-based PEFT setups on Grounding DINO evaluated on the DIOR-RSVG test set.

Methods	Efficiency	Pr@0.5	Pr@0.7	Pr@0.9	meanIoU
Image Encoder	98.66	54.10	47.30	27.80	48.80
Decoder	99.05	78.10	71.50	43.20	70.00
Image Encoder + Decoder	97.70	81.10	74.10	44.30	82.80
Encoders + Decoders	**96.74**	**81.30**	**74.70**	**45.20**	**82.90**

Evaluation Of The Adapted Models (DIOR-RVSG)

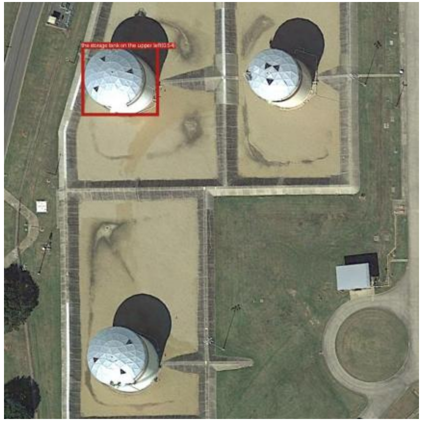
Methods	Pr@0.5	Pr@0.7	Pr@0.9	meanIoU	cumIoU
TransVG	72.41	60.05	27.84	63.56	76.27
VLTVG (ResNet-50)	69.41	58.44	24.37	59.96	71.97
VLTVG (ResNet-101)	75.79	66.33	33.11	66.32	77.85
QRNet	75.84	62.27	25.69	66.80	75.39
MGVLF	76.78	66.74	35.07	68.04	78.41
LPVA	**82.27**	72.25	39.55	72.35	**85.11**
Grounding DINO (Vanilla)	26.60	20.10	8.80	28.10	20.00
Grounding DINO (FFT) {FFT=Full Fine Tuing}	76.80	68.40	38.10	67.50	76.30
Grounding DINO+LoRA (Ours)	81.3	**74.70**	**45.20**	**82.90**	80.10
OFA + Adapter (Ours)	76.72	63.14	30.07	62.23	72.33
OFA + BitFit (Ours)	56.97	44.22	18.95	37.70	40.20

Evaluation Of The Adapted Models (OPT-RSVG)

Methods	Pr@0.5	Pr@0.7	Pr@0.9	meanIoU	cumIoU
TransVG	69.96	54.68	12.75	59.80	69.31
VLTVG (ResNet-50)	71.84	57.79	14.53	61.44	70.69
VLTVG (ResNet-101)	73.50	63.11	16.31	62.48	73.86
MGVLF	72.19	58.86	15.10	61.51	71.80
LPVA	74.69	60.56	15.84	63.78	**74.42**
OFA + Adapter (Ours)	66.38	46.70	12.86	41.67	66.39
Grounding DINO (Ours)	**75.81**	**66.47**	**26.39**	**65.24**	69.53

Inference Results (adapted Grounding DINO)

OPT-RSVG dataset



🡐The storage tank on the upper left

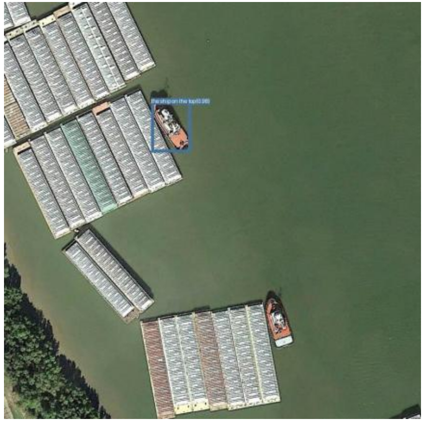
🡐The storage tank on the lower left

🡐The harbor on the right side

🡐The harbor on the left side

Inference Results (adapted Grounding DINO)

DIOR-RSVG dataset



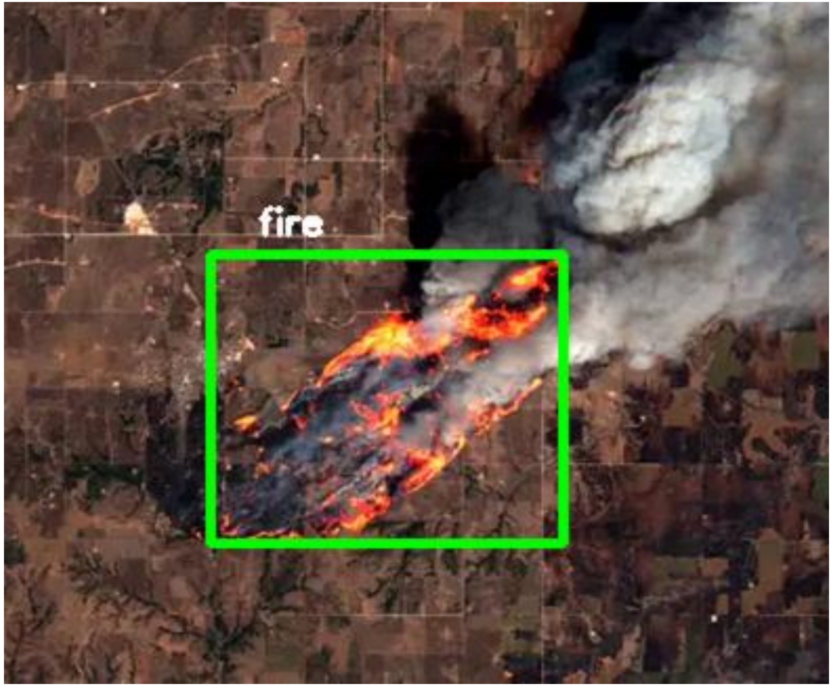
🡐The ship on the top

🡐The ship at the bottom

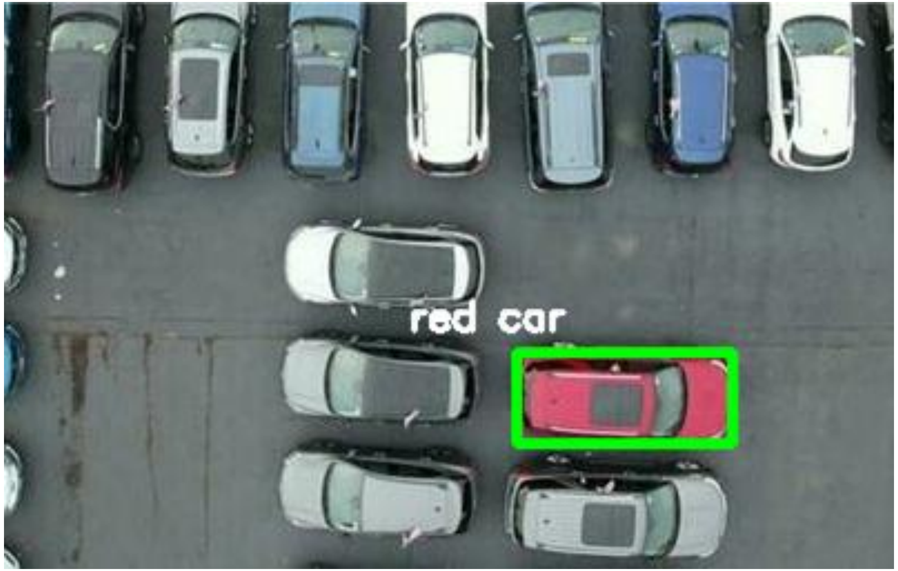
🡐The airplane on the right

🡐The airplane on the left

Inference Results (OFA)



Inference Results (OFA)



Conclusion & Key Findings

- PEFT techniques enable efficient adaptation of large vision-language models for remote sensing tasks.
- LoRA on Grounding DINO achieves state-of-the-art performance with minimal parameter updates.
- Adapters provide a strong balance of accuracy and efficiency when fine-tuning OFA.
- BitFit offers resource-efficient tuning but with reduced accuracy compared to other PEFT methods.
- PEFT presents a low-cost, practical solution for domain-specific visual grounding.
- Future work: explore hybrid PEFT approaches and extend to other remote sensing vision-language tasks.

Thank you for your attention!