Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

RS-OVC – Open-Vocabulary Counting for Remote-Sensing Data

Tamir Shor[1,2], *George Leifman*[2], *Genady Beryozkin*[2]

1-*Technion*, 2-*Google Research*

Motivation

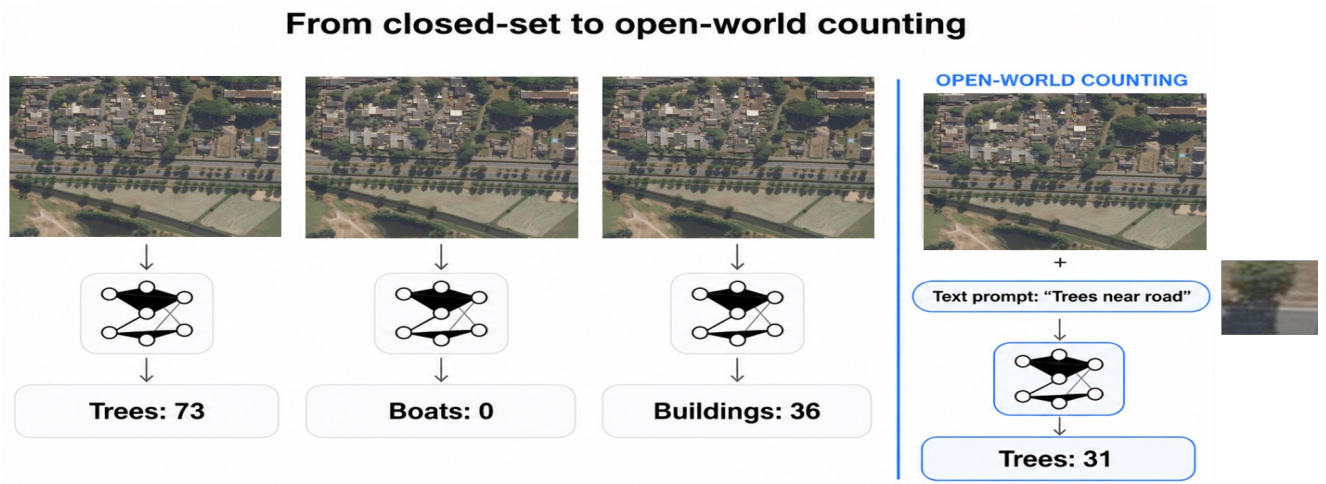
- Counting is central in a variety of tasks involving Remote-Sensing (RS) data – e.g. disaster control, urban planning, maritime surveillance, wildlife tracking, and ecological monitoring.

- Satellite and aerial imagery are exploding in volume and resolution, demanding automated quantitative analysis.

- Existing approaches for RS object-counting are mostly **closed-set**. Adaptations require expensive re-annotation and model retraining.

Motivation

- Many real-world applications in RS are **dynamic**, and require:
  - Extensions to novel classes, beyond a fixed object set.
  - Custom conditioning (e.g. the *white* boat, rather than boat).
  - Visual conditioning, alongside text.
- To address these gaps, we develop **RS-OVC** – the first open-world counting model for RS data.



Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

Prior Work

- Open Vocabulary **Detection** (OVD) models do exist – so why develop RS-OVC?
- The reason is that a perfect detection model is a perfect counting model, but we don't have those. Object detection models are sub-optimal for counting:
  - Localization makes optimization noisier under occlusions, scale diversity and low-res data.
  - They struggle with sub-pixel counting.
  - Data is harder to acquire (expensive annotation), especially dense scenes.
- OVC models have been developed for **general CV**. We show that trivially-applying such a model to RS data leads to poor performance.

Main Challenges

- Training an open-world RS counting model requires counitng data that is diverse in semantic traits (e.g. object types), scale, density and object count, among other conditions.
- Current publicly-available object counting dataset do not address these needs.
- In our work we adopt two orthogonal courses of action to address this difficulty:
  1. We **leverage broader open-world counting knowledge** learned by a pre-trained OVC model developed for general CV (for which much richer and more diverse data is available). We inject this model with RS knowledge to adapt it to our domain.
  2. We **curate several existing RS detection and counting datasets** into a single benchmark dataset that addresses diversity challenges raised before.

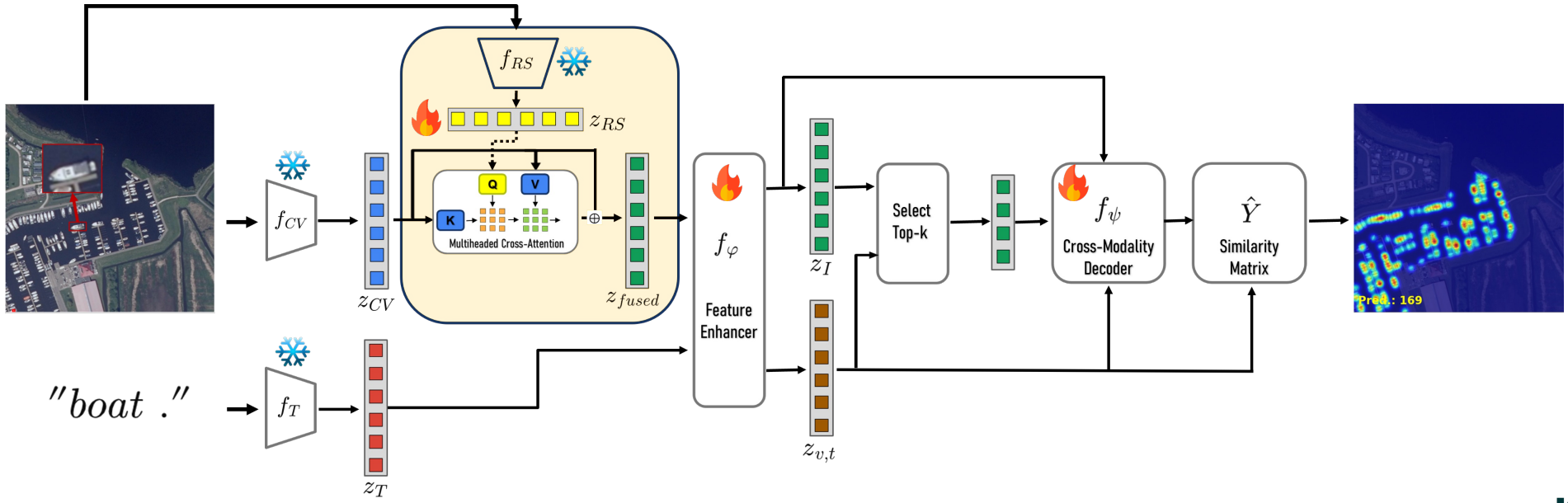
Method – RS Feature Injection

- Teaching a model to attribute visual features and free text requires large diverse corpora, currently unavailable for RS data.
- Rather than training a model from scratch, we start by using a pre-trained **CountGD** – an OVC model developed for general CV [1].
- CountGD allows zero-shot (pure text) and n-shot conditioning (with n visual exemplars).
- Naïve finetuning over RS data causes catastrophic forgetting (and is also costly) – our approach in this work is introducing RS knowledge with **minimal intervention**.
- We therefore propose **Feature Injection** – We freeze the original text and image encoders, add a pre-trained RS image encoder, and fuse RS features with original CV features using cross attention.

[1] -https://arxiv.org/pdf/2407.04619

Method – RS Feature Injection

- The original CountGD uses a feature enhancer (transformer encoder-decoder) for textual and visual feature fusion, where the $k = 900$ most correlated features are def to a transformer decoder for similarity scores.
- RS-OVC (yellow background) optimizes a cross-attention module with learned projections.



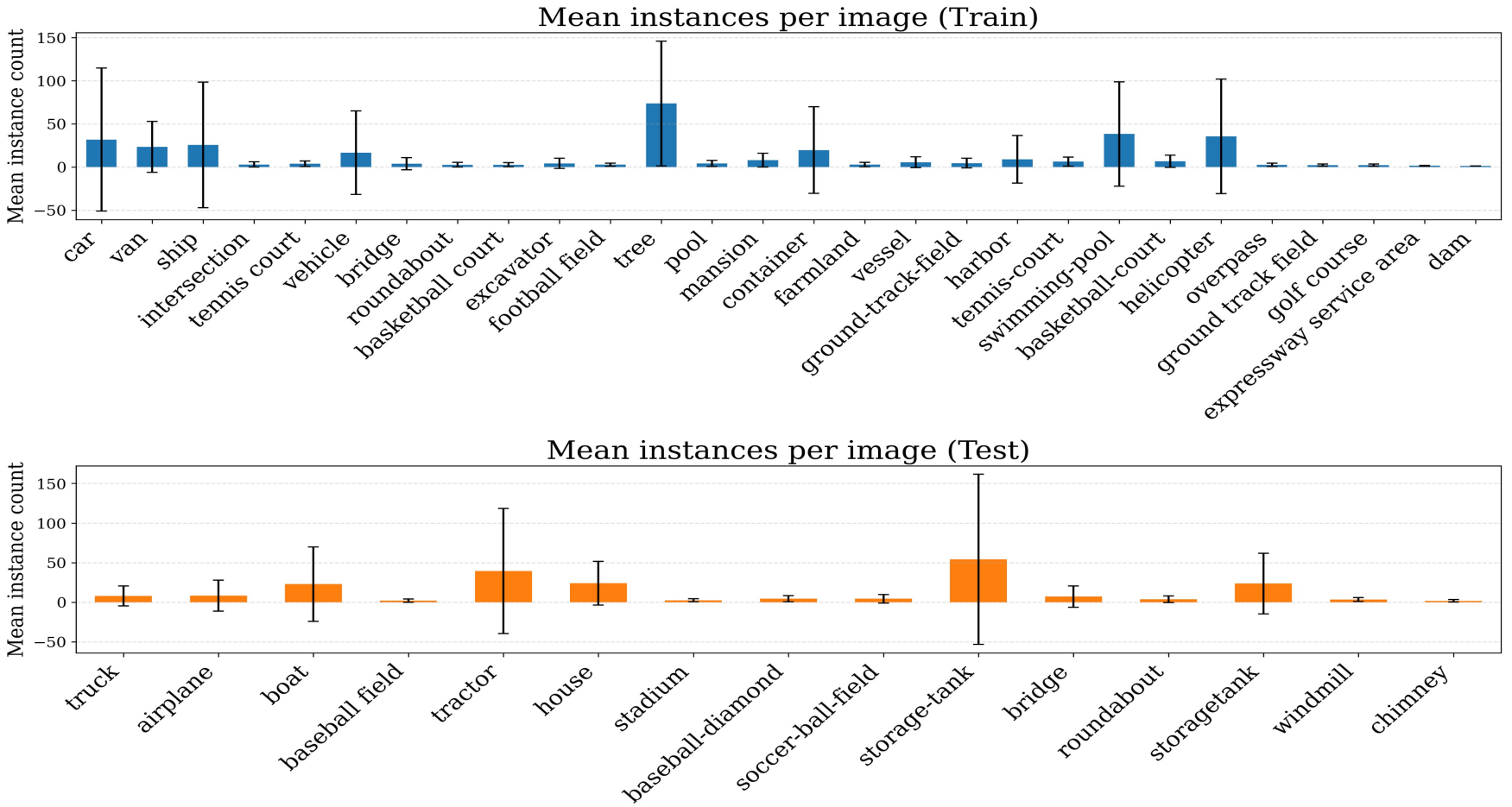
Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

Method – Data Curation

- Leveraging general CV knowledge and features helps mitigate previously mentioned challenges, but does not address them fully.
- We curate several existing RS datasets into a unified benchmark. These rely on semantically diverse OD datasets and one object counting dataset with denser scenes.
- For each dataset we curate labels to be object centroids for counting, and keep only samples with at least four object instances. Train-val-test split is done based on **object classes** (rather than the original sample-based split).

Method – Data Curation



Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

Experimental Setup

- Our curated dataset includes 23 train classes, 2 validation classes and13 test classes curated from NWPU-MOC, FAIR-1M, DIOR and DOTA, as well as ~1200 test-only samples from the RSOC-building dataset.
- We train and test in a 3-shot and zero-shot setup, and compare test MAE and RMSE of RS-OVC against four baselines:
  - LAE – a SOTA OVD model (AAAI 2025)
  - Off-the-shelf CountGD – CountGD trained for general CV data and applied to RS data without any adaptation.
  - CountGD-RSFT - Original model finetuned on our curated RS dataset *without* RS-informed feature aggregation.
  - CountGD-RS-only - RS-OVC simplified to use *only* a single frozen RS image encoder, relinquishing feature injection and the dual-encoder setup.

Quantitative Results

- RS-OVC substantially outperforms all baselines in most scenes, slightly underperforming baselines in low-density scenes (where OVD models already do well).

NWPU-MOC

Method	Boat		House		Truck		Stadium		Airplane		Mean	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
CountGD	32.58	63.56	23.54	35.66	11.38	29.59	2.5	3.16	1.97	2.52	20.16	37.1
CountGD (RS-FT)	36.25	76.01	21.75	34.88	8.80	18.58	**1.45**	**2.13**	**1.19**	**1.85**	18.71	36.89
LAE	40.03	77.90	27.11	38.73	12.72	20.70	5.50	5.83	4.96	5.21	23.44	39.73
RS-OVC (RS-only)	32.47	72.56	18.56	31.13	10.21	**17.32**	4	4.69	6.5	7.0	17.07	34.01
RS-OVC (Ours)	**7.5**	**15.44**	**15.52**	**26.87**	**8.74**	24.2	5.52	6.37	3.08	3.55	**12.42**	**24.72**

FAIR-1M

Method	Tractor		Boat		Truck		Airplane		Baseball-Field		Mean	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
CountGD	52.74	73.42	18.22	35.57	39.78	55.90	3.89	9.93	7.61	14.81	21.65	39.93
CountGD (RS-FT)	37.03	86.25	18.12	45.42	5.79	12.61	4.27	9.64	1.39	1.68	6.57	19.06
LAE	41.58	89.57	19.16	45.53	10.65	15.88	**1.29**	**6.74**	**0.48**	**0.82**	7.56	19.68
RS-OVC (RS-only)	37.55	86.66	18.14	45.23	6.04	12.72	4.39	9.75	1.58	1.88	6.74	19.08
RS-OVC (Ours)	**26.84**	**47.98**	**12.79**	**27.45**	**5.54**	**10.41**	4.21	10.22	3.00	3.06	**5.81**	**13.58**

Quantitative Results

- Conclusions from previous slide replicate – our model underperforms only on DIOR, that is characterized by much lower object counts and scene densities than other datasets.

DOTA

Method	Storage Tank		Airplane		Bridge		Soccer Ball Field		Roundabout		Baseball Diamond		Mean	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
CountGD	57.71	112.68	33.21	55.19	14.74	25.14	56.39	92.20	21.29	43.15	52.09	87.43	36.46	71.97
CountGD (RS-FT)	51.47	116.75	35.60	77.28	11.54	20.41	10.57	28.00	**4.18**	**5.36**	8.00	17.40	28.14	72.72
LAE	57.18	121.67	**18.86**	59.40	14.80	23.91	7.21	9.01	6.82	8.06	**4.67**	6.69	23.59	67.94
RS-OVC (RS-only)	56.18	111.30	32.14	73.74	16.75	21.58	18.75	26.80	9.22	9.88	10.63	11.80	23.82	60.85
RS-OVC (Ours)	**49.76**	**111.28**	24.13	**41.66**	**11.51**	**18.61**	**5.50**	**6.27**	5.67	6.93	5.33	**6.65**	**23.16**	**58.15**

DIOR

Method	Storage Tank		Bridge		Airplane		Stadium		Chimney		Baseball Field		Windmill		Mean	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
CountGD	14.28	**30.17**	13.93	35.29	3.43	8.67	8.00	8.00	2.02	3.61	28.00	81.35	3.10	4.08	11.70	39.02
CountGD (RS-FT)	18.94	39.14	3.29	3.70	6.24	10.69	6.00	6.00	5.40	5.74	4.38	4.83	2.63	3.23	9.18	22.96
LAE	**12.80**	36.24	4.15	5.00	**1.16**	**3.24**	**3.00**	**3.00**	**1.09**	**2.17**	**1.14**	**2.79**	**0.89**	**2.12**	**5.05**	**20.68**
RS-OVC (RS-only)	20.46	42.42	7.76	9.50	6.65	11.42	4.00	4.00	2.98	3.24	2.60	2.98	2.01	2.53	8.37	22.71
RS-OVC (Ours)	19.95	42.20	**2.50**	**2.94**	6.23	11.10	7.00	7.00	2.86	3.08	3.05	3.37	2.69	3.21	9.15	24.62

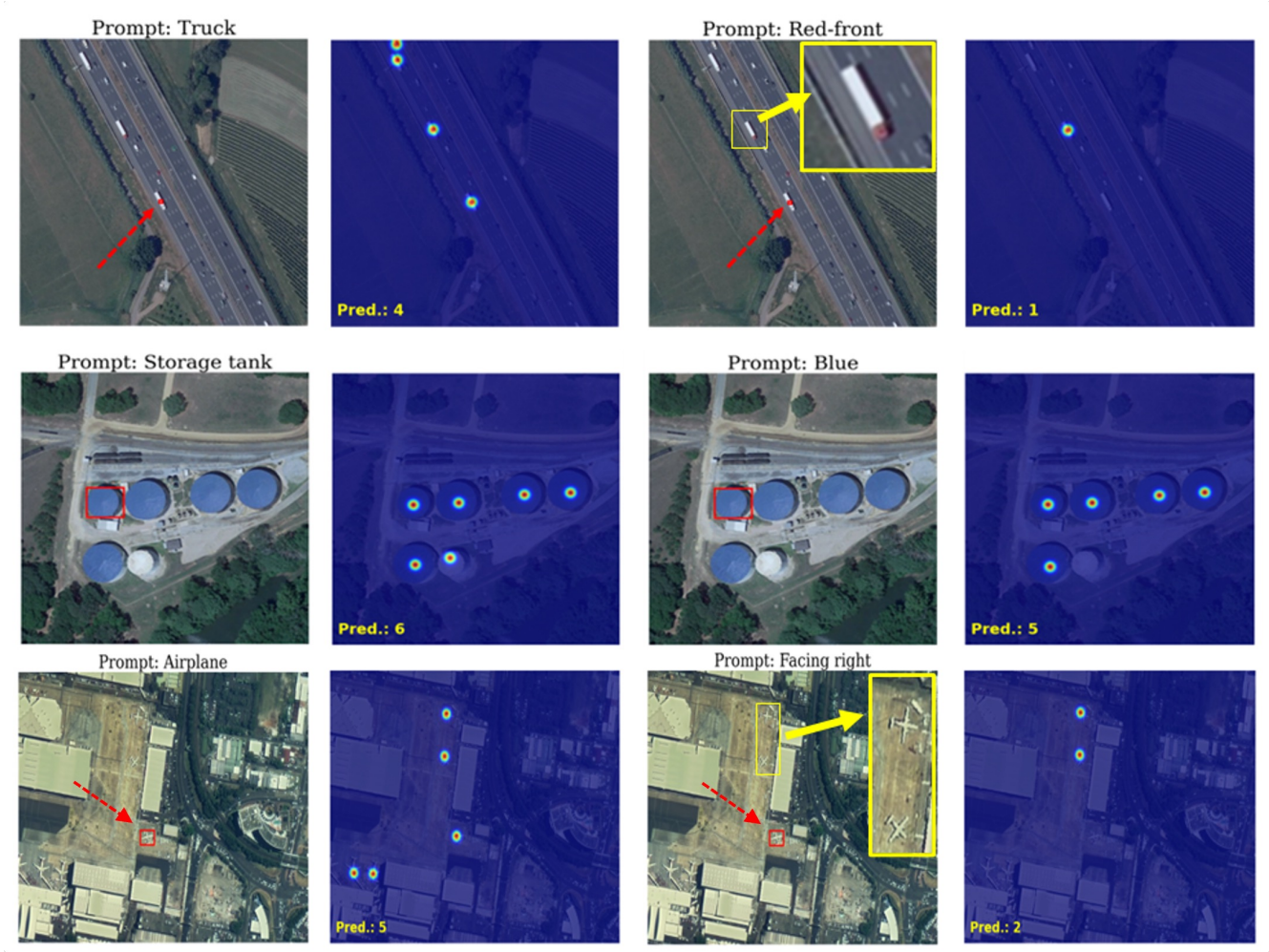
RSOC-Building

CountGD		CountGD (RS-FT)		LAE		RS-OVC (RS-only)		RS-OVC (Ours)	
MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
13.52	17.80	12.38	18.39	15.39	21.84	10.94	15.91	**10.51**	**15.30**

Quantitative Results

- Beyond quantitative benchmark results in counting, we test whether RS-OVC can demonstrate general scene understanding based on visual and textual conditioning.
- We design 3 tests in increasing levels of difficulty:
  1. **Local semantic understanding -** tests scene understanding focus on the **object of interest**.
  2. **Global semantic understanding -** tests concepts that require understanding the **broader scene** and relationships between objects.
  3. **Basic Reasoning -** tests concepts that require **inferring information from context**, rather than directly matching visible features

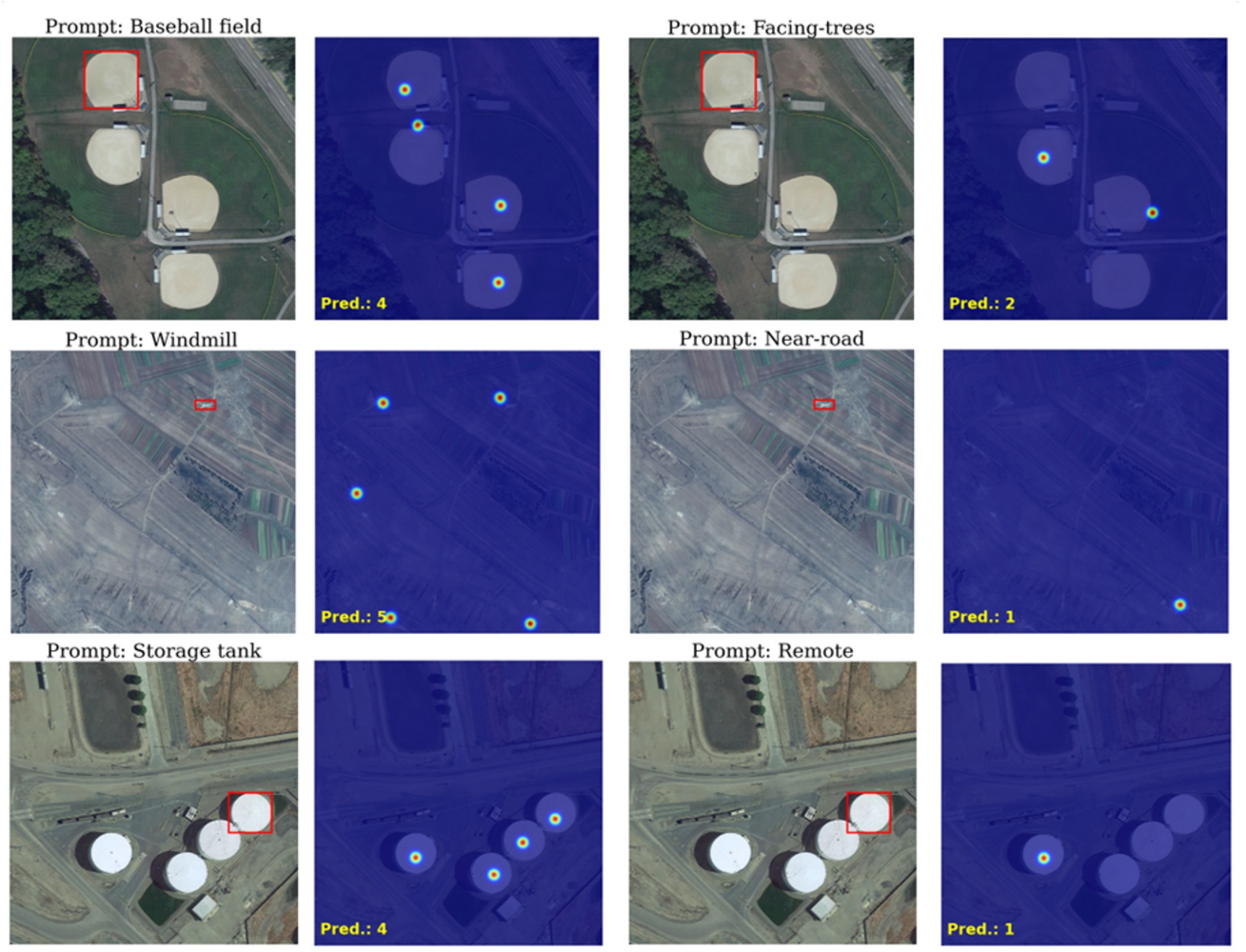
Local Understanding



Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

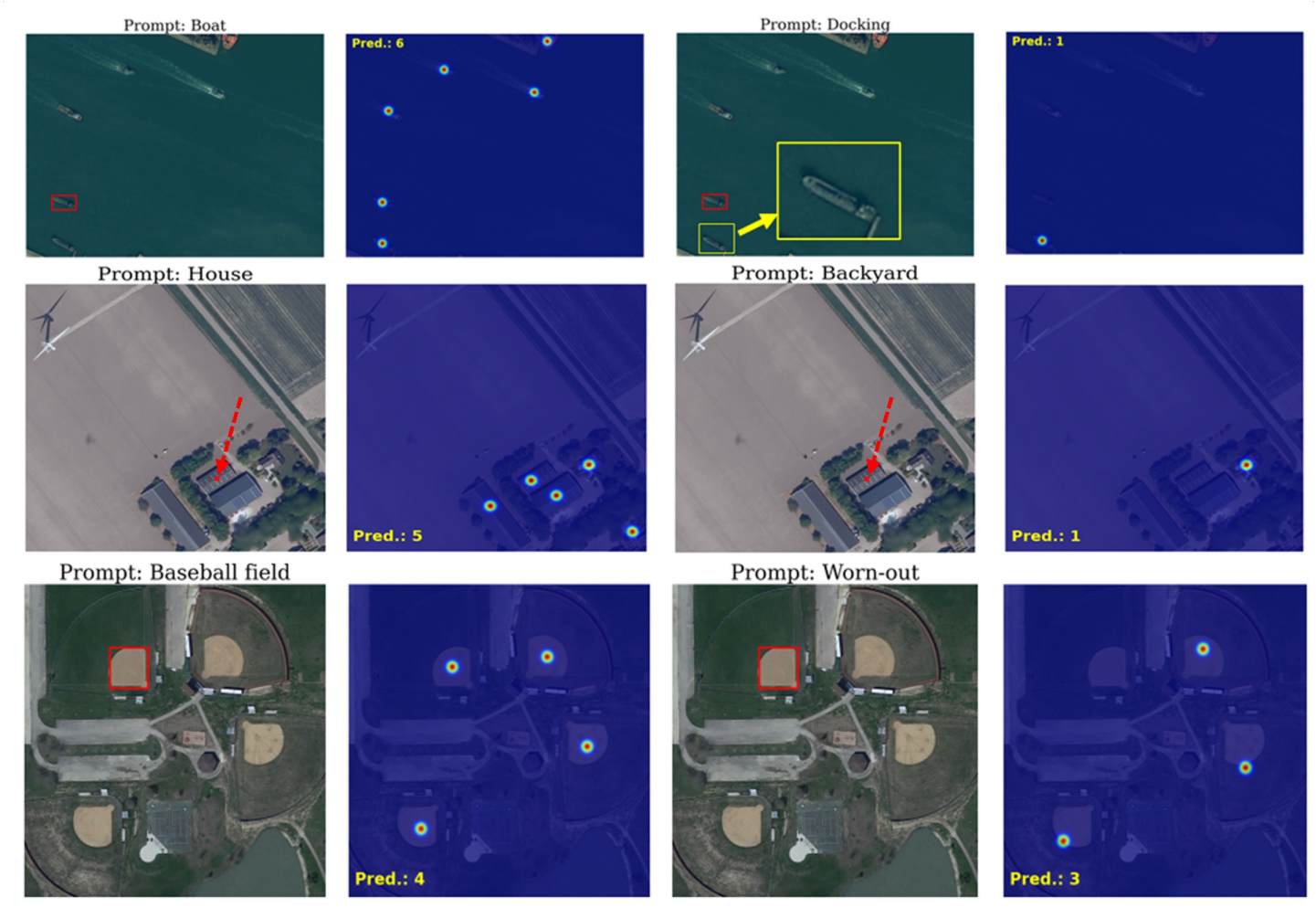
Global Understanding



Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

Basic Reasoning



Google Research

ICPR 2026 International Conference on Pattern Recognition LYON

Conclusion & Future Work

- We presented **RS-OVC** – the first open-vocabulary model for remote-sensing data.
- We used RS **Feature Injection** and consolidated a **curated RS OVC dataset** to alleviate data scarcity in RS.
- Our model shows substantial improvements over existing RS OVD and OVC solutions for general CV.
- **Future work** should explore a better balance between the dense and sparse scene performance, and also alleviate manual threshold tuning required for RS-OVC (inherited from CountGD).