Advancing Open-Set Object Detection in Remote Sensing Using Multimodal Large Language Model

Nandini Saini, Ashudeep Dubey, Debasis Das
Indian Institute of Technology Jodhpur, India
{saini.9, dubey.6, debasis}@iitj.ac.in

Chiranjoy Chattopadhyay
FLAME University Pune, India
chiranjoy.chattopadhyay@flame.edu.in

Abstract

In recent years, open-set recognition in remote sensing has attracted significant attention. The goal is to identify unknown objects during inference, extending the generalization of models trained on labeled data for known objects. However, obtaining bounding box annotations for unknown object categories at a large scale is prohibitively expensive. Multimodal large language models (MLLMs) offer a promising alternative, enabling the discovery of unknown object categories without the need for human intervention in labeling novel classes. In this paper, we propose a novel methodology that leverages MLLMs to address the dual challenges of detecting and categorizing unknown objects in remote sensing imagery. By integrating three diverse datasets—DOTA, DIOR, and NWPU VHR-10—we simulate real-world open-set conditions by partitioning object classes into known and unknown categories. The proposed methodology employs a two-step approach: (1) open-set object region detection, where known objects are identified using a model trained on labeled data, while threshold-based region proposal extraction is applied to detect unknown objects; and (2) discovery and semantic labeling of unknown objects using MLLM-based textual annotation. The contextual descriptions generated by the MLLM serve as human-interpretable pseudo-labels, which are further validated using vision-language similarity metrics. Experimental results demonstrate significant improvements in both detection (achieving high recall for unknown objects) and discovery (producing meaningful and accurate categorizations of novel objects). This work highlights the transformative potential of MLLMs for interpreting unknowns and paves the way for more robust open-set object detection in the remote sensing domain.

1. Introduction

The expansion of remote sensing technology has led to the generation of large volumes and diverse types of data at different altitudes and resolutions, increasing the demand for effective techniques in remote sensing image analysis. Remote sensing object detection plays a significant role in Earth observation by simultaneously identifying single or multiple objects in an image and localizing them with bounding boxes and their respective categories. The advancement of deep learning has introduced various object detection models [5, 33, 34], such as DETR [10], Mask R-CNN [16], and YOLOv10 [39], which have achieved remarkable progress under the closed-set paradigm. In this paradigm, models are trained on datasets with fixed, known object class labels. Consequently, these models learn spatial features specific to the training data's object classes. However, this approach often results in incorrect or missed detections when unknown objects are present in the images, as their labels are absent in the training data. Although recent efforts have built large, human-annotated remote sensing object detection datasets [4, 20, 41, 42], these come at a significant annotation cost. Additionally, the cost increases with the long-tailed distribution of objects in remote sensing images due to their extensive overhead views (as shown in Figure 1).

To overcome the limitations of closed-set training in object detection, researchers have introduced a new direction called open-set object recognition [11, 18, 35] and discovery [44]. This approach aims to identify unknown objects alongside known ones and categorize them, providing more comprehensive data for various applications. Emerging methods have addressed this challenge using few-shot learning, RNCDL [13]—which encompasses supervised and semi-supervised techniques—and vision-language models such as CLIP [32], GLIP [24], and Grounding DINO (GDINO) [27] for natural ground-level images. However, the distinct characteristics of remote sensing imagery compared to natural imagery make these models unsuitable for direct application in the remote sensing domain. As remote sensing imagery is increasingly utilized across diverse applications such as environmental monitoring [31], land use mapping [6], and disaster management [14], open-set object recognition and discovery be-

Figure 1. A visual comparison of the object detection task in closed-set and open-set scenarios. In closed-set object detection, the detection model is trained using known class label information and ignores any novel classes. In contrast, open-set object detection and discovery aim to identify and categorize both known and unknown classes within an image, enhancing scene understanding and the learning paradigm.

comes crucial for identifying unknown targets that were not included in training. Therefore, this paper focuses on advancing the open-set object detection task specifically for the remote sensing domain (as shown in Figure 1(b)).

The proposed work harnesses the complementary strengths of foundational multimodal large language models to transition existing closed-set object detectors into open-set object detectors. This approach enables the detection of known classes with predefined labels while also discovering new labels for objects of unknown classes, thereby facilitating intelligent vision-text-based open-set object detection and discovery to handle more realistic use cases.

In this work, we propose a two-step methodology to address open-set object recognition and discovery in remote sensing imagery. Our approach leverages closed-set trained object detector models to identify known object classes and generate region proposals for unknown classes. To discover and assign semantic labels to unknown-class region proposals, we employ foundational multimodal large language models. These textual label descriptions are further validated using the RemoteCLIP model [25], which is trained on remote sensing datasets.

The contributions of this research are as follows:

- **Transitioning Closed-Set Object Detectors to Open-Set Models**: We formalize a novel methodology to extend traditional closed-set object detection models for open-set recognition in remote sensing, enabling the identification and localization of both known and unknown objects.

- **Integration of Multimodal Large Language Models**: We leverage the strengths of pre-trained foundational multimodal large language models to discover and categorize unknown object classes, enhancing the capability of remote sensing object detection systems to handle real-world, dynamic scenarios.

- **Validation with Remote-Specific Adaptation**: We validate the discovered unknown object categories using a domain-adapted validation process with RemoteCLIP. The effectiveness and robustness of our proposed approach are demonstrated on three remote sensing benchmark datasets: NWPU VHR-10, DIOR, and DOTA.

2. Related Work

The goal of our work is to advance open-set object detection in the remote sensing domain by leveraging the rich semantic information generated by large language models, without the need for designing complex models. In this section, we review recent research studies on open-set object

detection and discovery tasks, as well as advancements in multimodal large language models.

2.1. Open Set Object Detection and Discovery

Open-set object detection and discovery is an emerging research area in computer vision that tackles the challenge of identifying and managing objects not included in the training dataset. Unlike traditional closed-set models, which are trained under the assumption that all possible classes are known during training, open-set approaches aim to recognize known categories while effectively managing unseen objects as "unknowns." Foundational insights from the broader domain of open-set recognition, as introduced by [36], establish the theoretical basis for addressing unknown categories by balancing the classification of known classes with the rejection of unknowns. In [18], the authors propose a framework for object detection in an open-world setting, where the system must detect known classes while incrementally identifying and learning new, previously unseen classes. This approach integrates open-set recognition with continual learning to address the challenges of dynamic environments.

Several recent works, particularly in the remote sensing domain, have focused on addressing open-set recognition tasks. For instance, studies leveraging SAR imagery integrate multitask learning and category discovery to identify novel targets effectively [8, 28]. Similarly, [19] presents a learning framework that combines representative and discriminative features for land cover classification in open-set settings. Another work, [7, 43], employs transferability and graph convolutional networks, respectively, to achieve scene classification in open-set remote sensing scenarios, emphasizing efficient handling of novel classes. Furthermore, a comprehensive survey presented by [12] explores broader implications and future directions for integrating open-world frameworks into the remote sensing domain. These related works provide strong motivation for designing an adaptive pipeline to address open-set detection and discovery in complex environments of remote sensing imagery.

2.2. Multimodal Large Language Model

Currently, large language models (LLMs) have garnered significant attention in the field of Natural Language Processing (NLP) [3, 38] due to their impressive performance on text-based downstream tasks such as summarization and language translation. Building on the remarkable capabilities of LLMs, the field of Multimodal Large Language Models (MLLMs) is rapidly evolving, focusing on integrating diverse data modalities such as text, images, and audio to enhance reasoning and interaction capabilities. These models are typically trained on extensive datasets comprising multimodal input-output pairs, enabling superior performance in tasks like Visual Question Answering (VQA) and image understanding. Notable examples of MLLMs include BLIP-2 (a bridge model for vision and language pretraining) [21], OpenAI's GPT-4 Vision (which integrates image understanding alongside text generation) [1], ImageBind by Meta AI (capable of learning across multiple modalities such as text, image, audio, and thermal data) [15], Gemini [37], Flamingo by DeepMind [2], and many others. Similarly, LLaVA [26] demonstrates an end-to-end multimodal system combining a vision encoder with Vicuna [30], enabling advanced general-purpose understanding and interaction comparable to multimodal GPT-4. These innovations collectively push the boundaries of vision and language integration, opening new possibilities for cross-domain understanding and multimodal interaction.

3. Methodology

This section outlines our approach to addressing open-set object detection and discovery, starting with a clear problem definition and a detailed discussion of our methodology using foundation models. The core idea is to fully leverage the capabilities of multimodal large language models (MLLMs) to tackle open-set object detection tasks, thereby reducing human effort in annotating large-scale training datasets and simplifying model design processes.

3.1. Problem Definiton

We aim to advance the task of open-set object detection by developing a model capable of accurately identifying both known and unknown classes in input images, while also providing category names for the unknown classes. The training dataset, denoted as D_{tr}, consists of images containing a set of labeled known object classes, $\mathcal{C}_i$. Each training image is annotated with known class labels $\mathcal{C}_i \in \{1, \ldots, m\}$ and associated bounding box coordinates $\mathcal{B}_i = \{x_i, y_i, w_i, h_i\}$, where x_i, y_i represent the center coordinates and w_i, h_i define the width and height of the bounding box for the labeled object class.

The testing dataset, denoted as D_{te}, contains images with both known and unknown object classes. Initially, the object detection model is trained on the closed-set training dataset. Subsequently, the model identifies regions in the input images that correspond to unknown objects using the threshold-based region proposal generation mechanism and categorizes them into novel object classes.

3.2. Our Approach

As illustrated in Figure 2, our approach primarily consists of four submodules, leveraging off-the-shelf pretrained models: (1) *Initialization*, which extracts feature representations for known and unknown object categories; (2) *Proposal Generation*, which generates proposals based on threshold settings to identify object-class-aware regions;

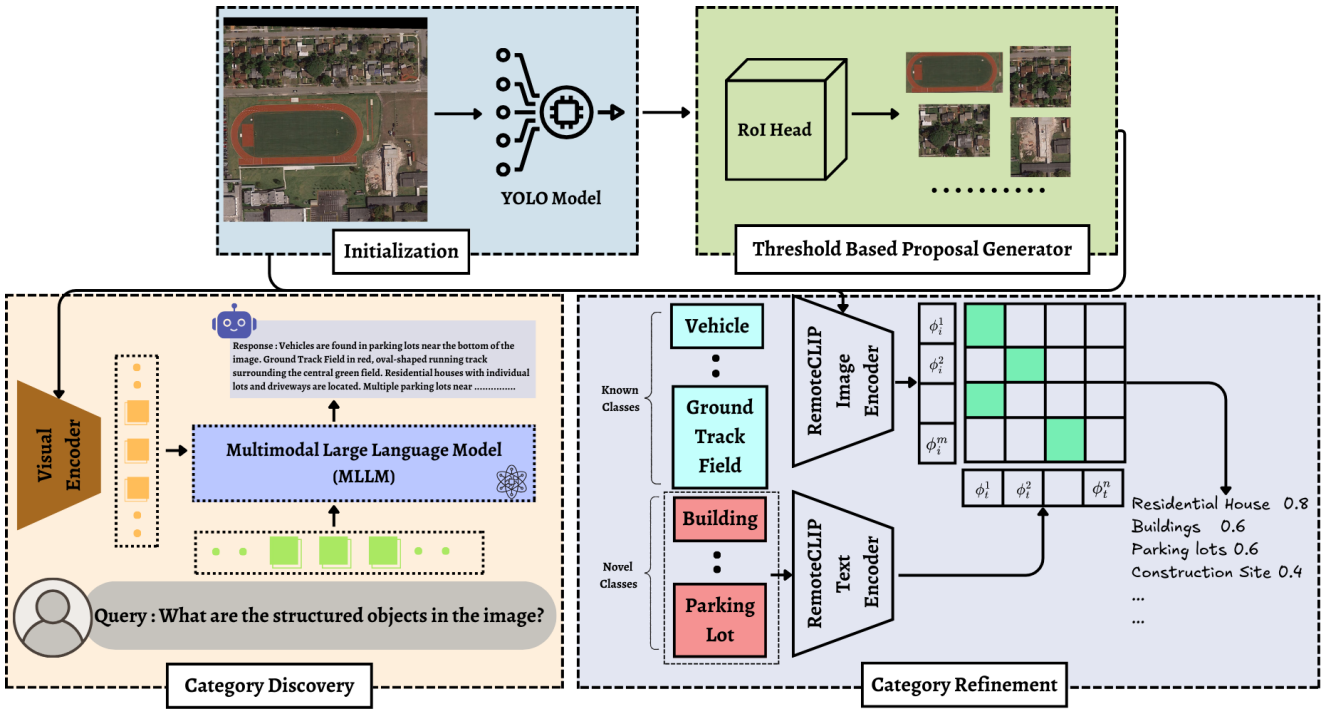

Figure 2. The overall architecture of the proposed pipeline, consisting of four submodules: (1) Initialization to extract feature representations for known and background (unknown) object categories; (2) Proposal Generation to generate region proposals using a threshold-based mechanism; (3) Category Discovery to identify novel class labels from the generated region proposals using MLLM; and (4) Category Refinement to enhance the mapping of novel class labels to the input image. The proposed pipeline leverages current foundation models for open-set object detection and discovery tasks in the remote sensing domain, aiming to achieve strong performance in identifying and classifying objects in complex scenarios.

(3) *Category Discovery*, which identifies novel classes from region proposals using pre-trained multimodal large language models; and (4) *Category Refinement*, which enhances the mapping of novel class discoveries by refining them with visual prototypes using RemoteCLIP [25]. The details of each component are outlined below.

Initialization. In our proposed approach, the first module involves the initialization of a closed-set object detection model to extract unrefined region proposals from the input image. We utilize a one-stage detector, specifically YOLOv8 [17], which does not inherently provide the flexibility to detect objects from an open vocabulary. The YOLO family of object detection models are known as single-stage detectors because they directly predict bounding boxes and class probabilities in a single forward pass through the network. Unlike two-stage detectors, YOLO processes the entire image at once, dividing it into a grid and predicting object locations and their associated classes for each grid cell. In this stage, the YOLOv8 model is trained on a remote sensing dataset with labeled data for specific known classes. The model predicts confidence-based bounding boxes for known class labels, while regions not corresponding to these classes are categorized as the background class, representing unknown objects.

Proposal Generation. In the second component, the YOLOv8 model is employed for region proposal generation, identifying regions in an image corresponding to known object classes while categorizing all other regions as a background class. A threshold-based mechanism is used to extract regions corresponding to both known and background classes. Formally, for a given input image I, the YOLOv8 model outputs bounding boxes for known and background classes, denoted as B_i^{KN} and B_i^{BG}, respectively, with confidence scores $C_i \in [0, 1]$. Since YOLOv8 is trained under closed-set settings, it cannot assign class labels to unknown background region proposals. To transform the closed-set object detector into an open-set object detector, we leverage multimodal large language models to infer and identify unknown classes.

Open Set Category Discovery with MLLM. After obtaining the region proposals from the input image, we utilize the zero-shot capabilities of multimodal large language models (MLLMs) to discover categories for the respective cropped regions in the third submodule. These region proposals are served as inputs to the MLLM, accompanied by a prompt such as *"What are the structured objects in each*

Dataset	Train	Val	Test	Known Class	Unknown Class 1	Unknown Class 2
DOTA	10,554	1696	1696	Airplane, Ship, Storage Tank, Baseball Field, Tennis Court, Ground Track Field, Harbor,Vehicle	Bridge, Basketball Court	Helicopter, Roundabout, Soccer Ball Field, Swimming Pool
DIOR	11,799	1,483	1,489			Expressway Service Area, Expressway Toll Station, Airport, Chimney, Dam, Golf Field, Overpass, Stadium, Train Station, Windmill
NWPU VHR-10	334	113	113			-

Table 1. The distribution of the remote sensing benchmark dataset, detailing the number of samples across each split and category details into three settings: known common classes, unknown class 1, and unknown class 2.

of the proposals?". The MLLM then provides responses in textual form. To extract novel class labels, we use the spaCy library, a robust natural language processing toolkit, to identify noun-based objects from the textual responses. SpaCy efficiently tokenizes the text and applies part-of-speech tagging to extract candidate nouns that represent object labels. As a result, we achieve the final discovery of novel object labels, leveraging the strengths of MLLM for open-set category discovery.

Category Refinement. In the final component of the pipeline, we refine the labels generated by the MLLM more precisely using RemoteCLIP, a model trained on remote sensing image-text pair datasets. RemoteCLIP enables efficient image-text matching and generalization. The input image I and the set of novel class labels, NC_i, where $i \in \{1, 2, \cdots, |C|\}$, are encoded and fed into the pretrained RemoteCLIP model. The cosine similarity between the enriched image and text embeddings is computed to finalize and refine the novel class predictions, providing confidence scores for each prediction.

4. Experiment Setup and Results

In this section, we first introduce the datasets used in the experiments, providing details about the distribution of known and unknown object classes. The qualitative and quantitative performance of our methodology in detecting objects from known classes and discovering novel object categories is discussed in subsequent sections.

4.1. Dataset Description and Settings

In our experiments, to validate the proposed detection methodology, we employed three publicly available remote sensing object detection datasets: DOTA, DIOR [23], and NWPU VHR-10 [9]. The DIOR dataset comprises a total of 23,463 images spanning 20 categories, while the DOTA dataset includes 16 object classes. The NWPU VHR-10 dataset consists of 650 positive images (i.e., images containing objects) across 10 categories. The train-val splits of all three datasets were utilized for closed-set training, while a mix of common and unique classes among the three datasets was used to evaluate the effectiveness of our approach under open-set conditions. The detailed distribution of these datasets is provided in Table 1. We examine our proposed pipeline under three class distribution settings: first, where known classes are common across all benchmark datasets; second, where unknown classes are shared among the datasets; and third, where unknown classes are uniquely defined for each dataset.

We implement the closed-set training as traditional supervised learning, running for 100 epochs using a cosine learning rate schedule and the Adam optimizer. All models are trained and inferred using 2 A100 GPUs and an NVIDIA GeForce RTX 1080 GPU with a dynamic batch size, ensuring 60% utilization of each GPU's memory.

4.2. Quantitative Results

The proposed pipeline advances open-set object detection by leveraging the capabilities of Multimodal Large Language Models (MLLMs). For this empirical study, we utilized four MLLM models: BLIP [22], BLIP-2 (an enhancement of Flamingo) [21], GIT-base by Microsoft [40], and ViT GPT-2 [29]. Initially, we trained a closed-set object detector using YOLOv8, achieving an overall mAP of 85.70 for known classes. This trained detector was subsequently adapted for open-set object detection and discovery tasks, utilizing different MLLMs to evaluate their effectiveness. The overall performance for unknown classes is described using standard metrics, demonstrating that MLLMs significantly enhance open-set object detection and facilitate the discovery of new classes.

Table 2 summarizes the performance of the four MLLM models in terms of Unknown Class Detection Precision (UDP), Unknown Class Detection Recall (UDR), Unknown Class Detection F1-Score (UDF1-Score), and Average Novel Class Detection rate per image. In the quantitative results, GIT outperforms the other models, achieving the highest UDP (82.98), UDF1-Score (83.82), and an average detection of approximately 3 novel classes, showcasing

Model	UDP	UDR	UDF1-Score	Average Novel Class Detection
BLIP	77.36	84.79	80.90	∼2
BLIP-2	72.23	85.58	78.34	∼2
ViT-GPT2	74.94	84.27	79.33	∼2
GIT	82.98	84.67	83.82	∼3

Table 2. The performance comparison of Multiomdal Large Language Models on Unknown Class Detection Metrics

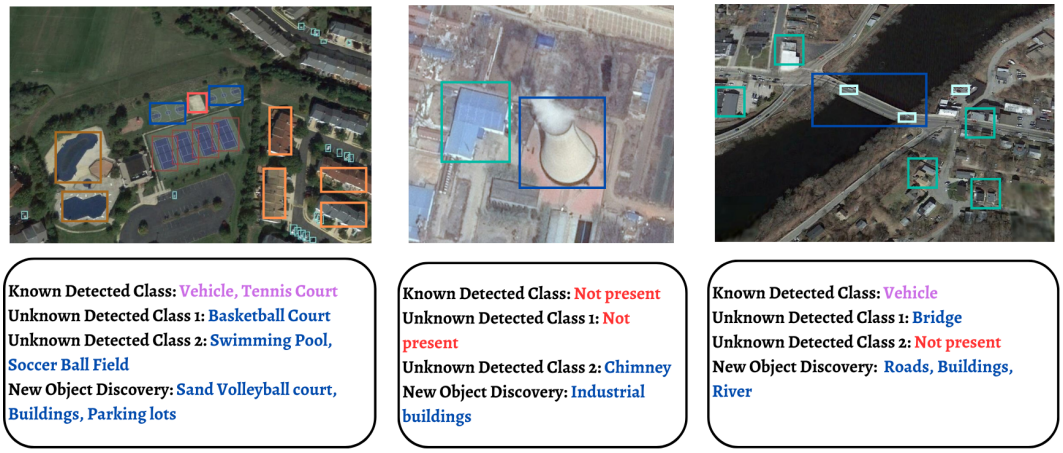

Figure 3. Visulization results of open-set object detection and discovery

its robustness in unknown class detection. BLIP achieves a competitive UDF1-Score of 80.90 but with a relatively lower UDP, indicating reduced precision. ViT-GPT2 exhibits balanced performance across all evaluated metrics. Conversely, BLIP-2 demonstrates the lowest UDF1-Score (78.34) and UDP (72.23), indicating potential areas for further improvement.

4.3. Qualitative Results

In addition to quantitative analysis, a qualitative comparison is presented in Figure 3 to illustrate the advancements in open-set object detection and discovery. It highlights how MLLM models generate richer contextual information from input images and refine categorical labels for unknown objects, assigning labels that are semantically closer to the input image. This analysis underscores the potential of MLLM models in improving the identification and categorization of previously unseen objects, which is crucial for open-set object detection. Moreover, the ability of these models to generalize across diverse object classes without predefined labels demonstrates their utility in expanding object discovery capabilities. This insight supports the use of MLLM models to enhance large-scale dataset annotation processes and tackle real-world scenarios where the model must dynamically adapt to new object categories, all while maintaining a streamlined design.

5. Conclusion & Future Work

In this research, we have presented a methodology for detecting known and unknown objects by leveraging the complementary strengths of foundation multimodal large language models (MLLMs). This methodology provides initial insights into how a closed-set object detector can be extended to perform open-set object detection and discovery tasks, particularly in the remote sensing domain. The proposed approach enhances labeling annotations and facilitates the generation of large-scale training data for multiple downstream tasks. The empirical study, supported by qualitative results, demonstrates that large language models are capable of handling the dynamic scenarios present in remote sensing images, including tasks involving open-set recognition and detection.

Despite the use of MLLMs in the proposed methodology to enhance accuracy and adaptability for open-set recognition tasks, there remain significant opportunities for future research. Some key areas to explore include:

- **Efficiency Improvements**: Developing a framework that optimizes inference speed and reduces computational and memory requirements.
- **Enhancing Detection Accuracy**: The alignment between bounding box detection and novel label discovery can be further improved by incorporating advanced semantic information transformation techniques.
- **Domain Generalization and Multimodal Integration**: Investigating techniques to improve the generalization capabilities of MLLMs across diverse remote sensing datasets (e.g., combining optical, radar, and infrared imagery) and imaging conditions.

By addressing these directions, future research can pave the way for more efficient, robust, and scalable open-set object detection systems, further advancing the application of multimodal large language models in remote sensing.

Acknowledgments: This work is partially supported by the project under the Project Id: IBITF/Note/EIR-PRAYAS/Cohort-03/SanctionLetter/2024-25/0076. The support is gratefully acknowledged.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 3

[4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[6] Vineet Chaturvedi and Walter T de Vries. Machine learning algorithms for urban land use planning: A review. *Urban Science*, 5(3):68, 2021. 1

[7] Jiehu Chen and Xili Wang. Open set few-shot remote sensing scene classification based on a multiorder graph convolutional network and domain adaptation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. 3

[8] Mingyao Chen, Jing-Yuan Xia, Tianpeng Liu, Li Liu, and Yongxiang Liu. Open set recognition and category discovery framework for sar target classification based on k-contrast loss and deep clustering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 3

[9] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016. 5

[10] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2988–2997, 2021. 1

[11] Leyuan Fang, Zhen Yang, Tianlei Ma, Jun Yue, Weiying Xie, Pedram Ghamisi, and Jun Li. Open-world recognition in remote sensing: Concepts, challenges, and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 12(2):8–31, 2024. 1

[12] Leyuan Fang, Zhen Yang, Tianlei Ma, Jun Yue, Weiying Xie, Pedram Ghamisi, and Jun Li. Open-world recognition in remote sensing: Concepts, challenges, and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 12(2):8–31, 2024. 3

[13] Vladimir Fomenko, Ismail Elezi, Deva Ramanan, Laura Leal-Taixé, and Aljosa Osep. Learning to discover and detect objects. *Advances in Neural Information Processing Systems*, 35:8746–8759, 2022. 1

[14] Omid Ghorbanzadeh, Yonghao Xu, Pedram Ghamisi, Michael Kopp, and David Kreil. Landslide4sense: Reference benchmark data and deep learning models for landslide detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. 1

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 4

[18] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021. 1, 3

[19] Razieh Kaviani Baghbaderani, Ying Qu, Hairong Qi, and Craig Stutts. Representative-discriminative learning for open-set land cover classification of satellite imagery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 1–17. Springer, 2020. 3

[20] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018. 1

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3, 5

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[23] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 5

[24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[25] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote

sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 4

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 1

[28] Xiaojie Ma, Kefeng Ji, Linbin Zhang, Sijia Feng, Boli Xiong, and Gangyao Kuang. An open set recognition method for sar targets based on multitask learning. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 3

[29] NLP Connect. vit-gpt2-image-captioning (revision 0e334c7), 2022. 5

[30] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 3

[31] Claudio Persello, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, and Gustau Camps-Valls. Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):172–200, 2022. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[33] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 1

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[35] Hiran Sarkar, Vishal Chudasama, Naoyuki Onoe, Pankaj Wasnik, and Vineeth N Balasubramanian. Open-set object detection by aligning known class representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 219–228, 2024. 1

[36] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 3

[37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[39] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 1

[40] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 5

[41] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 1

[42] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1

[43] Jun Zhang, Jiao Liu, Bin Pan, Zongqing Chen, Xia Xu, and Zhenwei Shi. An open set domain adaptation algorithm via exploring transferability and discriminability for remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021. 3

[44] Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2022. 1