Vision-informed Semantic Text Alignment for Open-set Recognition in Remote Sensing - Supplementary Material

Siddhant Gole[1] Akash Pal[1] Ankit Jha[2] Subhasis Chaudhuri[1] Biplab Banerjee[1]

[1]Indian Institute of Technology, Bombay [2]The LNM Institute of Information Technology (LNMIIT)

Supplementary Material

For completeness, we include additional qualitative and quantitative results in the Supplementary Material to further validate our approach.

1. What is in this supplementary?

- We provide detailed information about the dataset and the known/unknown class split used during training.
- We present an ablation study of the loss components used in the prototype and reconstruction networks, using features extracted from CLIP with a ResNet backbone.
- We evaluate the closed-set classification performance of our method in comparison to existing baselines.
- We analyze the impact of different ViT backbones in CLIP on open-set recognition (OSR) performance.
- We further examine how the size of the prototype network and the reconstruction network affects performance.
- We provide results to evaluate performance of ViSTA-RS with multi-modal features for PatternNet, NWPU-RESISC45 and UC Merced dataset.
- We provide the results for performance of ViSTA-RS on Hard Split of data for PatternNet, NWPU-RESISC45 and UC Merced dataset.
- Analysis of caption quality across multiple generative models.
- We visualize the distribution of weights learned by the reconstruction network across all training datasets.
- Provide details on adapting ViSTA-RS to SAR dataset.
- We provide the algorithm for implementing ViSTA-RS.

1.1. Dataset Details

- **MLRSNet** [8] consists of 46 classes and a total of 109,161 images, with resolution ranging from 0.1 m to 10 m and image size of 256 × 256 pixels.
- **PatternNet** [17] includes 38 classes, each containing 800 images with a resolution 0.06 m to 4.7 m and size 256 × 256 pixels
- **NWPU-RESISC45** [2] has 45 classes, with each class containing 700 images with resolution 0.2 m to 30 m with size 256 × 256 pixels
- **UC-Merced** [15] includes 21 classes, each containing 100 images with a resolution of 0.3 m and size 256 × 256 pixels.

The known/unknown splits for open-set configuration during training for all the datasets is provided in Table 1.

Datasets	Known		Unknown	
	$\mathcal{C}_L$	Train/Test	$\mathcal{C}_U$	Train/Test
MLRSNet	25	35000 / 24599	21	- / 20162
NWPU-RESISC45	30	12000 / 9000	15	- / 10500
PatternNet	19	13300 / 1900	19	- / 1900
UC Merced	15	1200 / 300	6	- / 120

Table 1. Known/unknown splits for the OSR setting across all RS datasets, showing the number of classes. (Train/Test) denotes the number of training and testing samples in each split.

1.2. Evaluating performance using Resnet-CLIP

Given that many existing Open Set Recognition (OSR) methods are built upon CNN-based backbones, we evaluate the generality of our approach by implementing it with a CLIP ResNet encoder. The quantitative results on the MLRSNet dataset are presented in Table 2. To further analyze the learned feature representations, we provide a qualitative comparison using t-SNE visualizations (Figure 1). These visualizations reveal two critical insights. First, the clusters formed by the ViT backbone are substantially more compact and well-separated than those from the ResNet backbone, confirming the superior feature discriminability of the Transformer-based architecture. Second, the visualizations demonstrate that training the prototype network with InfoNCE ($\mathcal{L}_{\text{InfoNCE}}$) loss yields significantly tighter intra-class clusters compared to contrastive loss ($\mathcal{L}_{\text{Con}}$).

1.3. Closed set classification performance

In accordance with the OSR performance we report the Macro-F1 score for closed set classification of our model. Results in Table 3 shows that ViSTA-RS consistently per-

Backbone	MM	f_θ		g_ϕ		AUROC	DTACC
		$\mathcal{L}_{Con}$	$\mathcal{L}_{InfoNCE}$	$\mathcal{L}_{MSE}$	$\mathcal{L}_{CS}$		
ResNet	×	✓	×	✓	×	0.6291	0.6240
	×	✓	×	×	✓	0.6804	0.6718
	×	×	✓	✓	×	0.7418	0.7294
	×	×	✓	×	✓	0.7855	0.7817
	✓	✓	×	✓	×	0.7677	0.7075
	✓	✓	×	×	✓	0.7747	0.7132
	✓	×	✓	✓	×	0.8343	0.7704
	✓	×	✓	×	✓	**0.8349**	**0.7707**

Table 2. Ablation study of ViSTA-RS with varying loss functions and a ResNet backbone for CLIP's image encoder. Here, MM,$\mathcal{L}_{Con}$, $\mathcal{L}_{InfoNCE}$, $\mathcal{L}_{MSE}$, and $\mathcal{L}_{CS}$ represent multimodal features, contrastive loss, InfoNCE loss, mean squared error loss, and cosine similarity loss, respectively.

Method	Dataset	
	MLRSNet	PatternNet
PROSER [16]	0.6781	0.6524
CVAECAP [4]	0.6882	0.6953
ARPL [1]	0.6940	0.7290
DIAS [7]	0.7071	0.7965
ConOSR [14]	0.7177	0.7772
MEDAF [12]	0.7468	0.7686
ViSTA-RS	**0.7736**	**0.8443**

Table 3. F1 scores for closed set classification of ViSTA-RS compared to different baselines.

forms better than other methods in classifying closed set samples.

1.4. Evaluating different ViT backbone on CLIP

To investigate the impact of model scale on feature representation, we extended our analysis from the ViT-B/32 CLIP backbone to include larger variants such as ViT-B/16 and ViT-Large. The results on the MLRSNet dataset, presented in Table 4, demonstrate a clear correlation between increased backbone capacity and improved performance.

CLIP Backbone	AUROC	AUIN	AUOUT
ViT-B/16	0.8336	0.8297	0.7680
ViT-B/32	0.8632	0.8755	0.7872
ViT-Large	0.8875	0.8928	0.8069

Table 4. Performance on various ViT backbone for CLIP on MLRSNet.

1.5. Evaluating performance of Prototype and Reconstruction network

The class prototype network (f_θ) and the reconstruction network (g_ϕ) serve as adapter modules responsible for learning clusters and reconstructing image embeddings, respectively. We evaluate the impact of capacity of these networks and report the corresponding performance metrics on the MLRSNet dataset in Table 5.

f_θ	g_ϕ	AUROC	AUIN	AUOUT
1	1	0.6438	0.6172	0.5682
	2	0.7349	0.6718	0.6879
	3	0.7231	0.6678	0.6610
2	1	0.8018	0.7849	0.7463
	2	0.8382	0.8931	0.7688
	3	0.8559	0.8933	0.7795
3	1	0.8284	0.8678	0.7799
	2	0.8609	0.8945	0.7828
	3	0.8632	0.8960	0.7904

Table 5. Performance across varying network capacity (number of layers) in Class Prototype Network and Reconstruction Network.

1.6. Effectiveness of Multimodal features

To disentangle the contributions of multimodal information from those of model size, we compare our approach with a capacity-matched, vision-only network in Table-6. Our evaluation on the PatternNet, NWPU-RESISC45, and UC Merced datasets confirms that the performance gains stem from the semantic richness of the added modality, not merely from an increase in network parameters.

1.7. OSR Performance on Hard Split of known and unknown classes

We show the performance of ViSTA-RS on Hard split of Known and Unknown classes for PatternNet, NWPU-RESISC45 and UC Merced dataset in Table-7. In Figure-2 we show the captions generated for known and unknown classes for hard split of data on the MLRSNet dataset. The details of known and unknown classes in the hard split for each dataset are provided in Table 8.

1.8. Evaluating Caption Quality

In Figure 5, we present captions generated by both BLIP [6] and GIT[11] alongside the predefined set of descriptions for MLRSNet dataset used for performance evaluation across different caption sources. The figure illustrates that BLIP consistently produces descriptive and accurate captions, whereas GIT occasionally generates incorrect descriptions—for example, it mislabels the class *"Freeway"*

Model	Feature Type	Dimension	PatternNet		NWPU-RESISC45		UC Merced	
			AUROC	DTACC	AUROC	DTACC	AUROC	DTACC
Visual-only	visual	512	0.8612	0.8296	0.8264	0.7967	0.8612	0.8575
Capacity matched	visual + random	1024	0.8711	0.8312	0.8356	0.8089	0.8634	0.8667
ViSTA-RS	visual + text	1024	0.9524	0.9086	0.9121	0.9034	0.9330	0.9121

Table 6. **Ablation Study on the Contribution of Multimodal Features on PatternNet, NWPU-RESISC45 and UC Merced Datasets.** VISTA-RS significantly outperforms both the visual-only and capacity-matched control models, demonstrating that the performance gain is due to the semantic richness of the text, not just the increased model capacity.

Method	MM	PatternNet		NWPU-RESISC45		UC Merced	
		AUROC	DTACC	AUROC	DTACC	AUROC	DTACC
ConOSR [14]	×	0.8269	0.8123	0.8064	0.7990	0.8112	0.8066
MEDAF [12]	×	0.8611	0.8412	0.8186	0.8289	0.8557	0.8385
CLIPN [10]	✓	0.9044	0.8901	0.8562	0.8431	0.8977	0.8702
CLIPScope [3]	✓	0.9124	0.8824	0.8657	0.8534	0.9162	0.8873
ViSTA-RS (unimodal)	×	0.8967	0.8744	0.8467	0.8408	0.8826	0.8678
ViSTA-RS	✓	0.9388	0.9086	0.8922	0.8934	0.9287	0.9006

Table 7. **Performance on Hard-Split Classes:** The substantial performance gap between VISTA-RS and other methods demonstrates that the multimodal framework effectively resolves visual ambiguity by leveraging semantic context.

as *"the bridge over the road"*. Such inaccuracies in GIT-generated captions lead to a noticeable decline in performance, an issue that does not arise when using captions from BLIP.

1.9. Visualising the Reconstruction weights

To better understand the weights assigned to prototypes for known and unknown samples, we visualize the Probability Distribution Function (PDF) of the reconstruction weights for both categories. Our reconstruction network is expected to map known class images closely to their respective class prototypes while struggling to do so for unknown class images, enabling effective rejection. Accordingly, weights for known classes should be high for their corresponding prototypes and low for others, whereas weights for unknown classes should be more uniformly distributed across prototypes. Figure 4 illustrates these distributions on the MLRSNet dataset, showing that known classes exhibit concentrated high weights on their correct prototypes, while unknown classes assign higher weights to multiple prototypes, confirming our hypothesis. Additionally, Figure 3 presents Weibull plots for all the four datasets that help determine suitable rejection threshold.

1.10. Adapting ViSTA-RS to SAR Imagery

To generate SAR-aware feature extractors, we specifically fine-tuned BLIP [6] and CLIP [9] using LoRA [5]. Following the methodology from the SARLANG-1M [13] paper, the fine-tuning protocol used a LoRA rank of 8. This adaptation was performed on the training split of the SARLANG-1M-Cap benchmark. For the OSR protocol on the MSTAR dataset, the 10 MSTAR vehicle classes were partitioned into six known classes (`2S1`, `BRDM-2`, `BTR-70`, `BMP-2`, `D7`, `T-72`) and four unknown classes (`BTR-60`, `T-62`, `ZIL-131`, `ZSU-23-4`).

1.11. Training Procedure:

ViSTA-RS presents a highly robust and flexible framework that can be seamlessly integrated with a wide variety of backbone architectures, enabling broad applicability across diverse remote sensing tasks. Its modular design facilitates effortless adaptation without compromising performance. To elucidate the practical implementation of ViSTA-RS, we provide comprehensive pseudo-code that details the training procedure in algorithm 1.

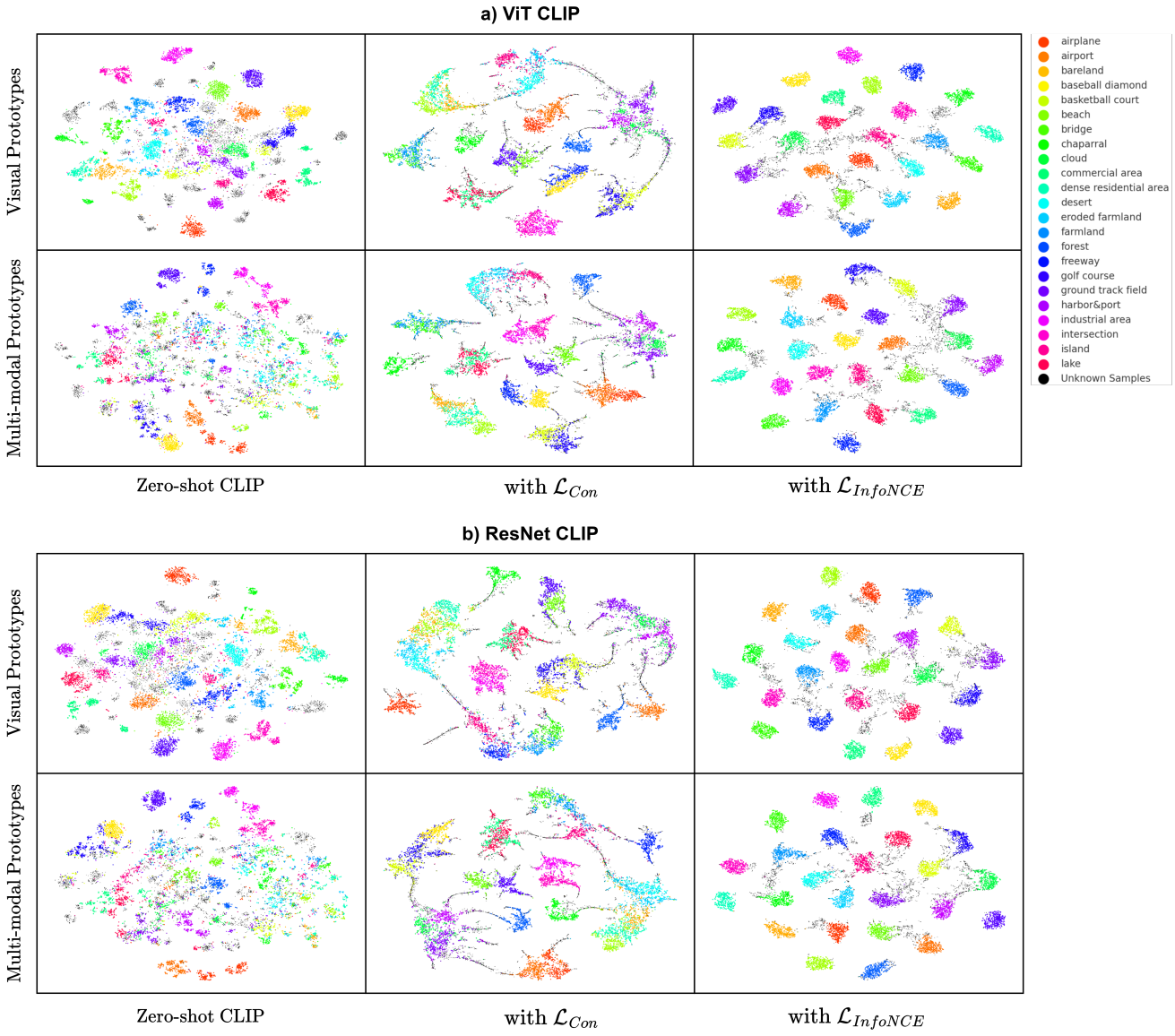
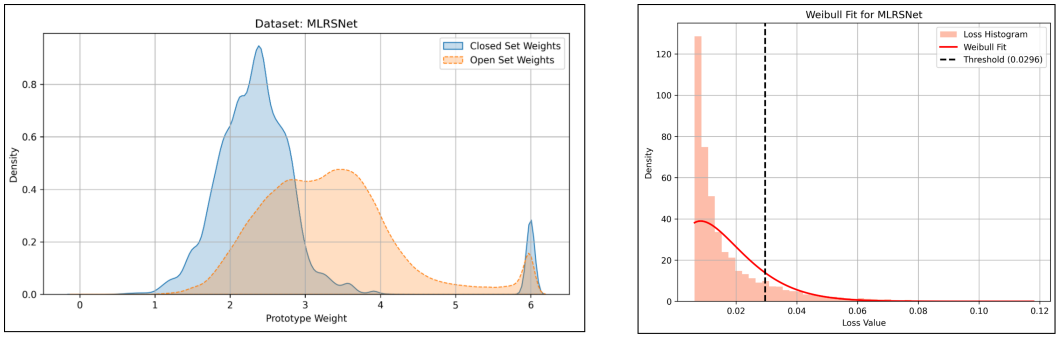
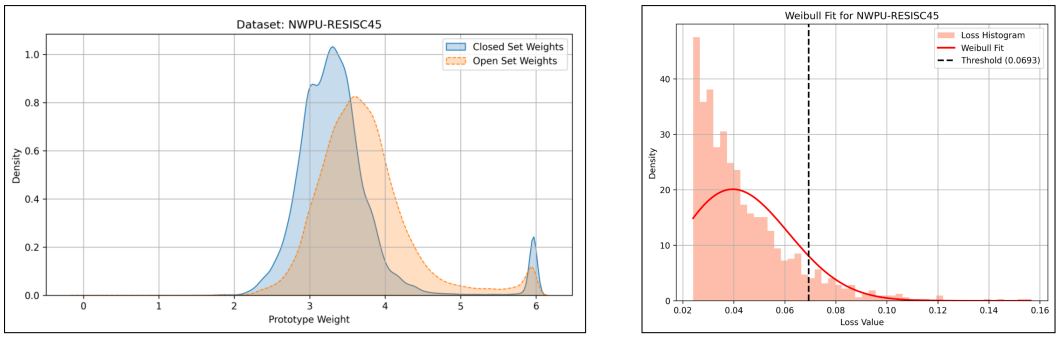


Figure 1. Comparison of t-SNE visualizations of visual and multimodal (visual + text) prototypes obtained from: (i) zero-shot CLIP, (ii) training with $\mathcal{L}_{\text{Con}}$ loss, and (iii) training with $\mathcal{L}_{\text{InfoNCE}}$ loss for both ViT CLIP and ResNet CLIP.

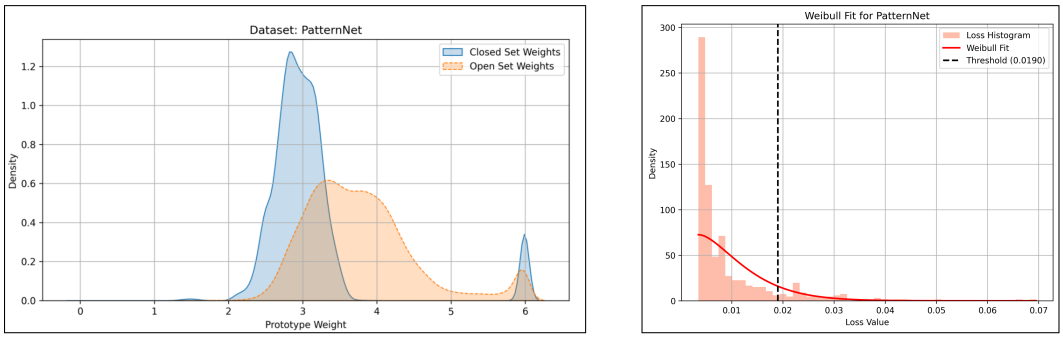
Figure 2. Qualitative examples of captions generated for known and unknown classes on the MLRSNet Hard split. The captions provide crucial semantic cues that mitigate the challenge of high inter-class similarity, enabling effective differentiation between known and unknown categories.



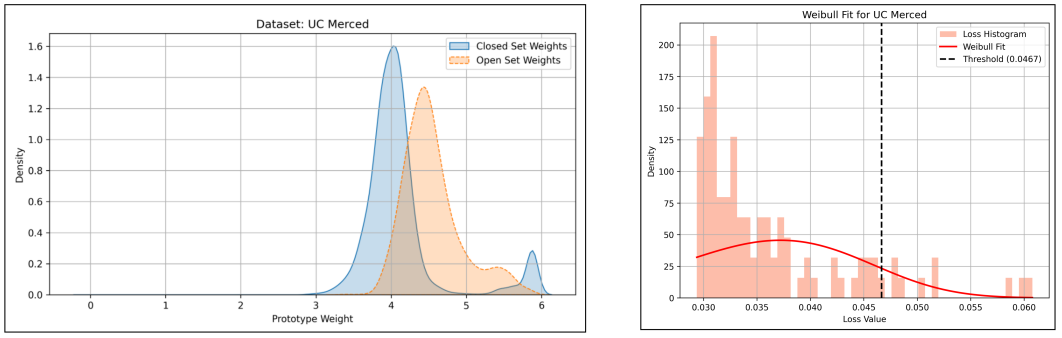
a) MLRSNet



b) NWPU-RESISC45



c) PatternNet



d) UC Merced

Figure 3. (Left) Distributions of prototype weights for known (closed set) and unknown (open set) classes, showing a clear separation between the two. (Right) Visualization of the Weibull distribution fitted to the reconstruction losses of known training data, which is used to automatically determine the rejection threshold (dashed line) for identifying unknowns.

References

[1] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2022. 2

[2] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105:1865–1883, 2017. 1

[3] Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad khorrami. Clipscope: Enhancing zero-shot ood detection with bayesian scoring. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 5346–5355, 2025. 3

[4] Yunrui Guo, Guglielmo Camporese, Wenjing Yang, Alessandro Sperduti, and Lamberto Ballan. Conditional variational capsule network for open set recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 103–111, 2021. 2

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 3

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. 2, 3

[7] WonJun Moon, Junho Park, Hyun Seok Seong, Cheol-Ho Cho, and Jae-Pil Heo. Difficulty-aware simulator for open set recognition. In *European Conference on Computer Vision*. Springer, 2022. 2

[8] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 3

[10] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no, 2023. 3

[11] 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, 2022. 2

[12] Yu Wang, Junxian Mu, Pengfei Zhu, and Qinghua Hu. Exploring diverse representations for open set recognition, 2024. 2, 3

[13] Yimin Wei, Aoran Xiao, Yexian Ren, Yuting Zhu, Hongruixuan Chen, Junshi Xia, and Naoto Yokoya. Sarlang-1m: A benchmark for vision-language modeling in sar image understanding, 2025. 3

[14] Baile Xu, Furao Shen, and Jian Zhao. Contrastive open set recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10546–10556, 2023. 2, 3

[15] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 1

[16] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2021. 2

[17] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 1