

SARVLM: A Vision Language Foundation Model for Semantic Understanding and Target Recognition in SAR Imagery

Qiwei Ma¹, Zhiyu Wang¹, Wang Liu¹, Xukun Lu¹, Bin Deng¹

Puhong Duan¹, Xudong Kang^{1*}, Shutao Li^{1*}

¹ The School of Artificial Intelligence and Robotics, Hunan University, China

{maqiwei, shutao_li}@hnu.edu.cn {xudong_kang}@163.com

Abstract

*Synthetic Aperture Radar (SAR) is a crucial imaging modality thanks to its all-weather capability. Although recent advances in self-supervised learning and masked image modeling (MIM) have enabled SAR foundation models, these methods largely emphasize low-level visual features and often overlook multimodal alignment and zero-shot target recognition in SAR imagery. To address this, we construct **SARVLM-1M**, a large-scale vision-language dataset with over one million image-text pairs aggregated from existing datasets. We further propose a domain transfer training strategy to mitigate the large gap between natural and SAR imagery. Building on this, we develop **SARVLM**, the first vision language foundation model (VLM) tailored to SAR, comprising **SARCLIP** and **SARCap**. SARVLM is trained with a vision-language contrastive objective under the proposed domain transfer strategy, bridging SAR imagery and textual descriptions. Extensive experiments on image text retrieval, zero-shot classification, semantic localization, and imagery captioning demonstrate that SARVLM delivers superior feature extraction and interpretation, outperforming state-of-the-art VLMs and advancing SAR semantic understanding. Code and datasets will be released soon.*

1. Introduction

SAR is a high-resolution imaging technology with all-weather, day-night operability and strong penetration capabilities, widely applied in military, environmental, maritime, and disaster monitoring tasks. Compared with optical imagery, SAR images are characterized by speckle noise, geometric distortions, and limited semantic textures, which pose significant challenges for downstream tasks such as ship detection [48], aircraft recognition [14], object detection [50], semantic segmentation [24], and image classification [9]. These challenges highlight the need for robust and

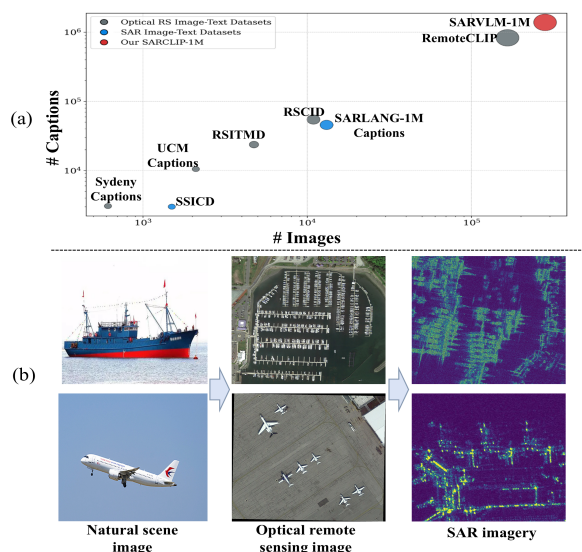


Figure 1. Comparison of image-text datasets in remote sensing. Gray circles represent datasets in the optical image domain, blue circles denote those in the SAR domain, and the red circle indicates our proposed SARVLM-1M. In the SAR domain, SARVLM-1M significantly surpasses existing datasets in both image and text volume.

generalizable feature representations specifically tailored to SAR data. Recent developments in vision foundation models (VFMs) have led to promising generalization capabilities across domains. As illustrated in Figure 2, existing VFMs can be broadly categorized into three paradigms: (a) contrastive learning (CL-based) methods such as SimCLR [5], which aim to learn discriminative representations by aligning augmented views of the same image; (b) masked image modeling (MIM-based) methods that reconstruct occluded image regions to learn spatial representations; and (c) CLIP-based approaches [33] that align images and texts via contrastive loss on large-scale paired data, showing remarkable transferability in cross-modal tasks.

The self-supervised training paradigm has recently driven the development of several vision foundation mod-

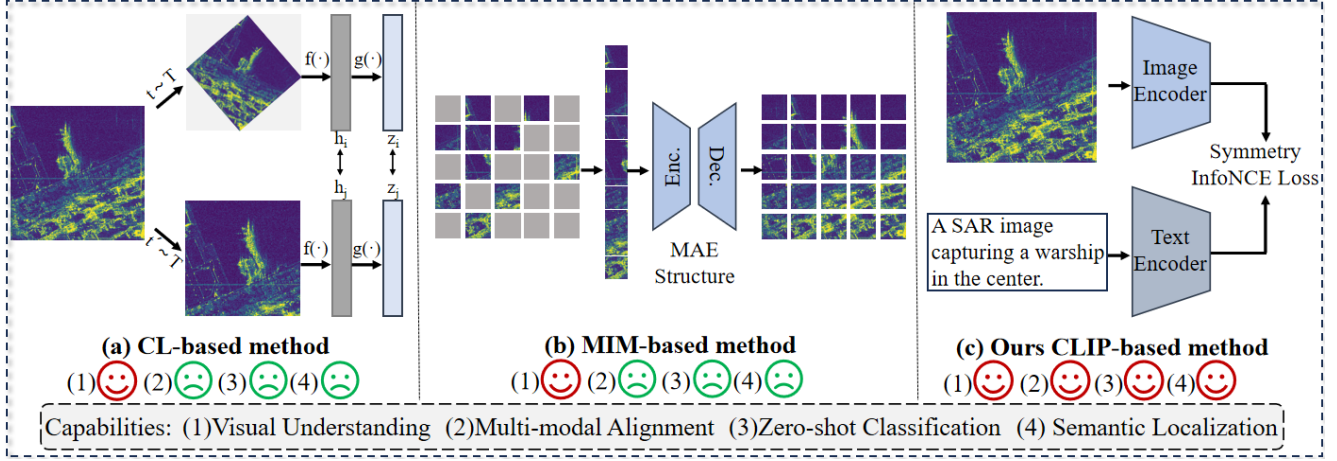


Figure 2. The paradigm for foundation model: (a) CL-based methods, (b) MIM-based methods, (c) CLIP-based methods.

els in remote sensing, including RS-BYOL [13], ScaleMAE [34], and Cross-scale MAE [37]. Notably, ScaleMAE [34] focuses on learning multi-scale representations by reconstructing both low- and high-frequency components across defined spatial scales. Similarly, Cross-scale MAE [37] enforces consistency across scales using a combination of contrastive and generative losses to enhance self-supervised remote sensing representations. However, these methods predominantly address the extraction of visual features. Inspired by CLIP [33], RemoteCLIP [22] pioneered the first vision language foundation model (VLFM) in remote sensing, aiming to capture semantic features by aligning image-text pairs. Subsequent works like RS5M [51] and SkyScript [40] have further explored this aspect. However, these models are pre-trained on optical image, leaving the SAR modality underexplored due to its distinct characteristics and lack of large-scale image-text datasets in SAR domain.

In recent years, the growing interest in SAR modality has led to the creation of various large-scale SAR datasets [20, 21, 25, 43]. Specifically, SARDet-100K consolidates existing SAR datasets to form a large collection encompassing ships, aircraft, cars, bridges, tanks, and harbors. Building upon this, SARATR-X extends SARDet-100K by integrating additional classification datasets to create the larger SARDet-180K. Both SARDet-100K [21] and SARATR-X [20] then employ MIM training strategy to establish their respective foundation models. Furthermore, recent advanced methods have to address the application of multi-modal LLMs in the SAR field. Particularly, SARChat-2M [27] introduces a substantial benchmark with two million multimodal dialogues, facilitating intelligent interpretation of SAR imagery through LLM-based conversational paradigms. A comparable initiative has been undertaken with SARLANG-1M [41]. However, despite their impressive performance, existing SAR vision foundation models and multimodal LLMs remain limited by the lack of tex-

tual annotations in current datasets, hindering their ability to capture rich semantic information fully.

To bridge this gap, we propose **SARVLM** consisting of SARCLIP and SARCap, the first vision-language foundation model designed specifically for SAR imagery. As shown in Figure 1, we construct **SARVLM-1M**, a large-scale SAR image-text dataset comprising 1.7 million pairs across diverse object categories and land cover types. These pairs are generated by leveraging domain knowledge, spatial rules, and templated text synthesis strategies. Based on SARVLM-1M dataset, we adopt a two-stage domain strategy to transfer knowledge from optical to SAR domain, enabling semantic alignment between modalities. As shown in Figure 3, this approach enables the extraction of comprehensive general features from SAR images and significantly enhances their semantic understanding, thereby elevating performance for SAR interpretation.

The key contributions of our SARVLM are summarized as follows:

- We propose SARVLM-1M, a novel large-scale vision-language dataset containing 1.7 million image-text pairs, encompassing various object and land cover types.
- To the best of our knowledge, SARVLM is the first VLFM tailored for SAR imagery, enabling SAR-specific cross-modal alignment, target recognition, zero-shot classification, semantic localization, and imagery captioning.
- Extensive experiments demonstrate that SARVLM consistently surpasses existing state-of-the-art VLFMs on multiple downstream tasks, highlighting its strong generalization and semantic representation capabilities in the SAR domain.

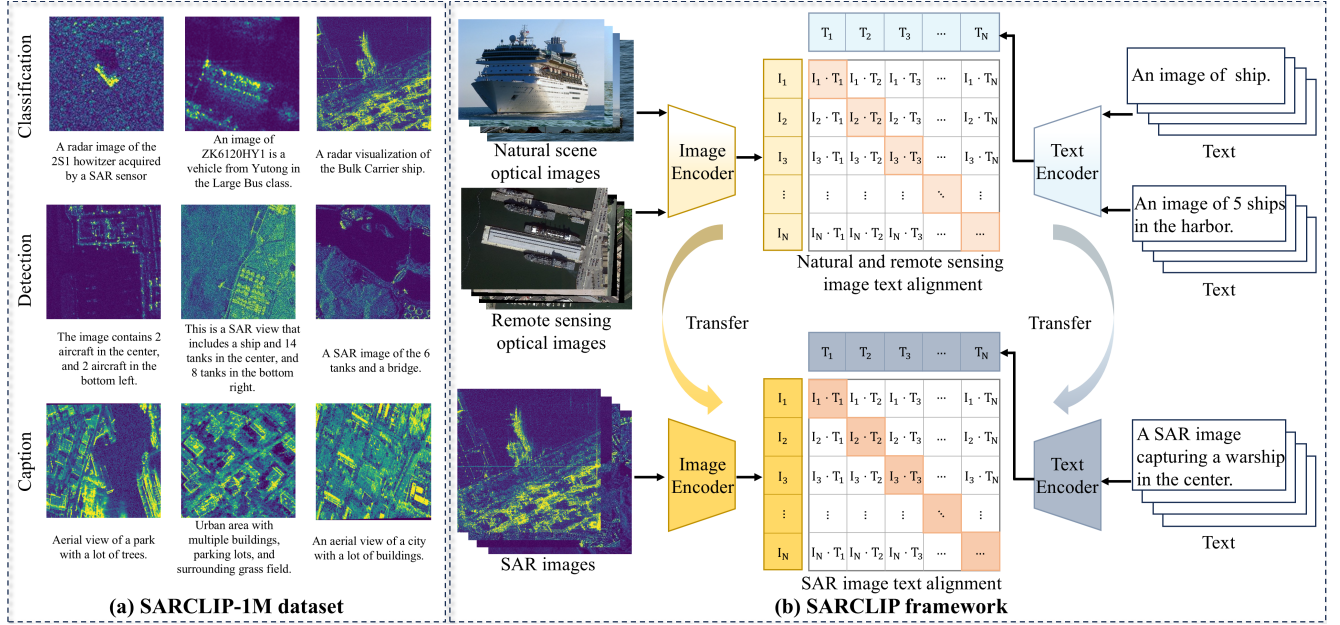


Figure 3. (a) Examples from the SARVLM-1M dataset; (b) Two-stage domain transfer training strategy for SARCLIP.

2. Related Work

2.1. Vision Language Model for Remote Sensing

Vision-language models have significantly advanced remote sensing through the development of multi-modal techniques. These methods broadly fall into two categories: contrastive methods and generative methods. Among contrastive approaches, CLIP [33] stands out as a pioneering work, employing a two-tower architecture to align visual and language features through extensive data from Internet. Inspired by CLIP, several studies have adapted this paradigm to remote sensing, yielding models such as RemoteCLIP [22], GeoRSCLIP [51], SkyScript [40], RSM-CLIP [10] and Mall et al. [29]. Notably, RemoteCLIP [22] enhances pre-training by converting detection and segmentation annotations into image captions, facilitating CLIP-style contrastive learning for image-text alignment in remote sensing. In the realm of generative methods, works like GeoChat [15], EarthGPT [49], and LHRS-Bot [31] implement auto-regressive large language model (LLM) architectures for vision-text alignment. For instance, GeoChat [15] introduces a multimodal model with LLaVa [23] architecture in remote sensing, achieving multi-granularity alignment through parameter-efficient fine-tuning. However, most of these existing methods primarily focus on the optical modality in remote sensing, largely overlooking investigations in the SAR domain.

2.2. SAR Foundation Model

Foundation models, by pretraining on large-scale data, are capable of capturing generalizable visual representations

that effectively support a wide array of downstream tasks. In remote sensing, recent advancements have seen the utilization of self-supervised learning and MIM techniques, as demonstrated by approaches such as RingMo [36], SatMAE [8], ScaleMAE [34], OmniSat [2], and Cross-scale MAE [37]. These models have been widely applied in tasks like aerial object detection and target recognition.

With the increasing availability of SAR imagery, a multitude of datasets have emerged, including SARDet-100K [21], SARATR-X [20], SAR-JEPA [19], FAIR-CSAR [43], and ATRNet-STAR [25]. Specifically, SARDet-100K constructs SAR foundation model through MIM training, initially pretraining on aerial view images before transferring to SAR imagery. Building upon this, SARATR-X leverages various classification datasets to build the SARDet-180K dataset, subsequently yielding a MIM-based SAR foundation model. Furthermore, SARLANG-1M [41] utilizes LLMs for SAR image interpretation, while SARChat [27] supports key tasks such as visual understanding and object detection in SAR imagery. Nevertheless, a common limitation among these existing SAR foundation models is their primary focus on low-level image features, often failing to capture deeper semantic information and realize multi-modal alignment within SAR images.

3. Methodology

This section introduce the paradigm of our framework, dataset construction approach and training strategy for SARCLIP and SARCap.

Table 1. Illustration of SARVLM-1M, which includes 8 SAR datasets. Cls.:Classification. Det.:Detection. Cap.:Caption. # Train Imgs.: Number of training images. # Train Caps.: Number of training captions. # Val Imgs.: Number of validation images. # Val Caps.: Number of validation captions. # Test Pairs: Number of image-text pairs in testing set.

Datasets	Year	Task	# Train Imgs.	# Train Caps.	# Val Imgs.	# Val Caps.	# Test Pairs.
MSTAR [1]	1995	Cls.	3,046	15,230	9,855	49,275	180
SARSim [16, 30]	2017	Cls.	21,168	105,840	—	—	—
OpenSARShip [18]	2017	Cls.	26,679	133,395	—	—	—
SAMPLE [17]	2019	Cls.	5,380	26,900	—	—	—
ATRNet-STAR [25]	2025	Cls.&Det.	68,091	340,455	29,284	146,420	6,667
SARDet-100K [21]	2024	Det.	94,493	472,465	10,492	52,460	2,783
FAIR-CSAR [43]	2024	Det.	51,948	259,740	11,790	58,950	7,096
SARLANG-1M-Captions [41]	2025	Cap.	9,191	31,968	3,939	13,682	3,902
SARVLM-1M(ours)	2025	Cap.	279,996	1,385,993	65,360	320,787	20,628

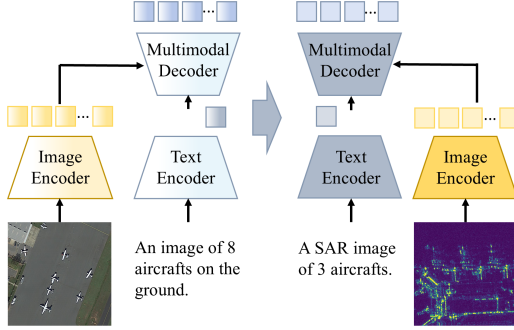


Figure 4. The framework of SARCap method.

3.1. Problem Definition

In this section, we investigate the paradigm of learning joint representations from SAR images and their corresponding textual descriptions. Specifically, we construct SARVLM-1M dataset $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{T}_i)\}_{i=1}^M$ consisting of SAR images $\mathbf{I}_i \in \mathcal{R}^{H \times W}$ with the corresponding descriptions $\mathbf{T}_i \in \mathcal{T}$. As shown in Figure 3, our objective is to learn a pair of modality-specific encoders that project both SAR images and text into a shared semantic space. Specifically, we define a visual encoder $f_v : \mathcal{R}^{H \times W} \rightarrow \mathcal{R}^d$ that maps the input SAR image to a d -dimensional visual feature embedding $z_v^i = f_v(\mathbf{I}_i)$, and a textual encoder $f_t : \mathcal{R}^T \rightarrow \mathcal{R}^d$ that maps the textual input to a corresponding textual embedding $z_t^i = f_t(\mathbf{T}_i)$. The goal is to align the embeddings z_v^i and z_t^i of matched image-text pairs in a common representation space, such that semantically similar inputs across modalities are embedded close to each other. This formulation enables the model to bridge the modality gap between SAR images and natural language, thereby supporting a wide range of downstream tasks such as cross-modal retrieval, target recognition, and zero-shot classification in SAR domains.

3.2. SARVLM-1M Dataset Construction

To address the aforementioned challenges, we build a large-scale vision-language dataset SARVLM-1M collected from

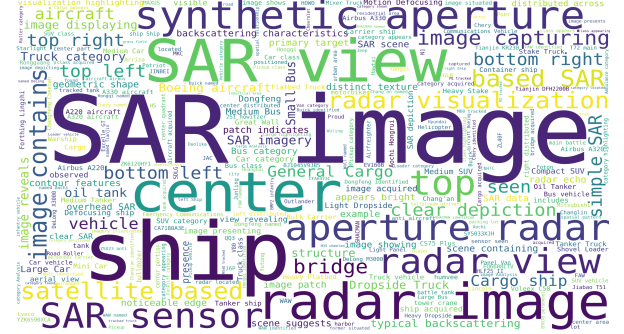


Figure 5. The word cloud of SARVLM-1M dataset.

existing classification, detection and caption datasets in SAR domain. As illustrated in Table 1, several large datasets have released in the past two year. For classification datasets, we employ 10 simple and complex description template to generate image captions. As for detection, we additionally design template to describe object absolute position and relative spatial position between targets in the images. The details for template design method are as follows:

(1) **General descriptions.** We utilize simple templates such as "A SAR image of the [class]" where [class] is replaced with a category name from classification datasets, or object type and quantity from detection datasets.

(2) **Complex descriptions.** We employ complex templates like "A SAR image reveals the distinct texture and structure of the [class]." where the usage of template is the same as general. These complex templates can enhance the diversity of descriptions and the robustness of models

(3) **Absolute region descriptions.** The image is divided into five regions: upper left, upper right, bottom left, bottom right and center region. We calculate the IoU between the box annotations of target and each region to determine its location. Templates like "A SAR image of [classes] located in the [location] of the image." are utilized.

(4) **Relative region descriptions.** Templates are utilized to describe spatial relationships between targets. Relative templates like "In this SAR image, the [class1] in the [lo-

cation1] are positioned [relative_direction] the [class2] in the [location2].” In these template, “[relative_direction]” includes above, below, left and right.

Based on the above templates design, we obtain caption for each SAR imagery. Then, A large language model is utilized to verify the fluency and grammatical correctness of text. Finally, we build our SARVLM-1M dataset contains 279,996 images and 1,385,993 captions in training set, and 20,628 image-text pairs in testing set. As shown in Figure 5, the dataset includes diverse target types such as ship, vehicle, aircraft, bridge and other land cover types.

3.3. Model Training

3.3.1. SARCLIP training

To further enhance the semantic understanding ability of the model for SAR image, we utilize two stages training strategy to train our SARCLIP. Firstly, we train vision large model on natural remote sensing dataset like RemoteCLIP which contains LoveDA [39], DOTA [44], RSCID [26] and others. By this strategy, our model is capable of better understanding and recognizing the unique characteristics of remote sensing imagery, particularly in identifying small scale and densely distributed objects. Secondly, we fine-tune this pretrained model on our SARVLM-1M dataset to transfer knowledge from natural to SAR domain.

In both two stage, we train our model using a contrastive learning framework that encourages positive pairs (z_v^i, z_t^i) to be more similar than negative pairs. The objective is to minimize the following symmetric InfoNCE-based contrastive loss which can be define as follows:

$$\mathcal{L}_{\text{clip}} = -\frac{1}{N} \sum_{i=1}^N \left\{ \left[\log \frac{\exp(\text{sim}(z_v^i, z_t^i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(z_v^i, z_t^j)/\tau)} + \log \frac{\exp(\text{sim}(z_t^i, z_v^i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(z_t^i, z_v^j)/\tau)} \right] / 2 \right\} \quad (1)$$

where $\text{sim}(a, b) = \frac{a^\top b}{\|a\| \|b\|}$ is cosine similarity, τ denotes a temperature hyperparameter, and N is the batch size.

3.3.2. SARCap training

For the SAR image captioning task, we use the CoCa [45] architecture, which combines contrastive loss and captioning loss. The captioning loss minimizes the difference between the predicted and ground-truth captions:

$$\mathcal{L}_{\text{caption}} = -\sum_{t=1}^T \log P(w_t | w_1, \dots, w_{t-1}, z_v) \quad (2)$$

where T is the caption length, w_t is the t -th token, and $P(\cdot)$ denotes the predicted probability of w_t conditioned on the previous tokens and the visual feature.

3.3.3. Downstream fine-tuning

To adapt our pretrained SARCLIP model to the SAR target recognition task, a lightweight fine-tuning is designed. Specifically, we freeze the parameters of the visual encoder and train a task-specific classification head $f_h : \mathcal{R}^d \rightarrow \mathcal{R}^C$ on the top of frozen visual features, the formula are as follows:

$$z_v = f_v(I) \quad (3)$$

$$\hat{y} = f_h(z_v) \quad (4)$$

where I is the input SAR image, C denotes the number of target categories, and $z_v \in \mathcal{R}^C$ is the visual feature from SARCLIP. The model is optimized using the standard cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\log \frac{\exp(\hat{y}_y)}{\sum_{c=1}^C \exp(\hat{y}_c)} \quad (5)$$

where y is the ground-truth label of the input image.

4. Experiment

4.1. Datasets

To evaluate the superiority of our proposed SARCLIP, we utilize downstream dataset as follows: **SARVLM-1M test**. This SAR retrieval dataset comprises 20,628 image-text pairs compiled from existing datasets. The test set consists of 180 images from MSTAR, 6,667 images from ATRNet-STAR, 2,783 images from SARDet-100K, 7,096 images from FAIR-CSAR, and 3,902 images from SARLANG-1M-Captions. Each image is uniquely captioned.

MSTAR-SOC. This SAR target recognition benchmark acquired by an X-band HH radar at 0.3 m resolution, containing 10 military vehicle classes and four standard experimental settings [4]. In the SOC setting, all 10 classes appear in 2,747 training images from 17 aspect angles and 2,425 test images from 15 angles.

FUSAR-Ship. This ship recognition dataset comprising 15 main ship classes, 98 subclasses, and diverse non-ship maritime targets [20]. It is built from 126 Gaofen-3 ultrafine-resolution images (1.124×1.728 m, dual-pol DH/DV), covering open sea, coastal waters, rivers, islands, and land backgrounds.

SAR-VSA. This SAR target recognition dataset comprises 25 fine-grained categories, integrating data from MSTAR [1], FUSAR-ship [12], and SAR-ACD [35]. It contains 11,045 training images and 8,161 testing images, as constructed by [20].

AIR-SLT. This is a remote sensing semantic localization benchmark with 22 large-scale images and 59 text-query test cases, each paired with ground-truth bounding boxes [46, 47].

Table 2. Retrieval performance on the SARVLM-1M test set and improvements over the SARCLIP baseline(%). LAION [6] is an optical image dataset in the natural scene domain, while RS5M, Skyscript, and R3+D10+S4 are optical image datasets in the remote sensing domain. [†], [§], and [‡] denote models pretrained on RS5M [51], Skyscript [40], and R3+D10+S4 [22], respectively.

Method	Image Backbone	Pretrain Data	Tune On	Image to Text			Text to Image			Mean Recall
				R@1	R@5	R@10	R@1	R@5	R@10	
OpenCLIP	ResNet-50	LAION	–	0.01	0.02	0.07	0.01	0.04	0.13	0.04
OpenCLIP	ViT-B-32	LAION	–	0.02	0.05	0.10	0.02	0.12	0.20	0.08
OpenCLIP	ViT-L-14	LAION	–	0.01	0.07	0.13	0.05	0.16	0.30	0.12
GeoRSCLIP	ViT-B-32	LAION	RS5M	0.03	0.09	0.18	0.06	0.19	0.31	0.14
GeoRSCLIP	ViT-L-14	LAION	RS5M	0.05	0.15	0.24	0.09	0.32	0.51	0.22
GeoRSCLIP	ViT-H-14	LAION	RS5M	0.01	0.06	0.12	0.08	0.32	0.51	0.18
SkyCLIP	ViT-B-32	LAION	SkyScript	0.01	0.08	0.15	0.05	0.23	0.37	0.14
SkyCLIP	ViT-L-14	LAION	SkyScript	0.03	0.10	0.20	0.14	0.39	0.65	0.25
RemoteCLIP	ResNet-50	LAION	R3+D10+S4	0.01	0.07	0.16	0.03	0.12	0.21	0.10
RemoteCLIP	ViT-B-32	LAION	R3+D10+S4	0.02	0.07	0.13	0.03	0.14	0.24	0.10
RemoteCLIP	ViT-L-14	LAION	R3+D10+S4	0.01	0.08	0.16	0.05	0.17	0.33	0.13
SARCLIP	ResNet-50	LAION	SARVLM-1M	2.71	10.82	17.57	3.07	11.76	19.03	10.49
SARCLIP [‡]	ResNet-50	R3+D10+S4	SARVLM-1M	2.98	10.79	17.35	3.21	12.02	19.39	10.95 (+4.38)
HarMA	ViT-B-32	LAION	SARVLM-1M	1.15	4.80	9.00	1.13	5.05	9.21	5.06
SARCLIP	ViT-B-32	LAION	SARVLM-1M	3.42	12.82	20.32	3.63	13.72	21.40	12.55
SARCLIP [†]	ViT-B-32	RS5M	SARVLM-1M	3.73	13.33	20.61	3.95	13.99	21.77	12.89
SARCLIP [§]	ViT-B-32	SkyScript	SARVLM-1M	3.79	13.23	20.31	3.72	13.80	21.46	12.72
SARCLIP [‡]	ViT-B-32	R3+D10+S4	SARVLM-1M	3.58	13.25	20.89	3.79	13.64	21.52	12.77 (+1.75)
SARCLIP	ViT-L-14	LAION	SARVLM-1M	4.73	15.53	23.81	4.75	16.09	24.51	14.90
SARCLIP [†]	ViT-L-14	RS5M	SARVLM-1M	4.81	15.60	23.57	4.77	16.56	25.26	15.09
SARCLIP [§]	ViT-L-14	SkyScript	SARVLM-1M	4.69	15.70	23.84	4.67	16.34	24.65	14.98
SARCLIP [‡]	ViT-L-14	R3+D10+S4	SARVLM-1M	4.69	15.87	24.04	4.90	16.62	25.42	15.25 (+2.34)

4.2. Evaluation Metrics

We employ image text retrieval and target recognition as our downstream task to demonstrate the effectiveness of our framework. For the retrieval task, we adopt Recall at top- K ($R@K$), where $K \in \{1, 5, 10\}$. The formula can be defined as:

$$R@K = \frac{1}{N} \sum_{i=1}^N I[y_i \in \text{TopK}(q_i)] \quad (6)$$

where N is the number of queries, y_i is the ground-truth item for query q_i , and $I[\cdot]$ is the indicator function. For the target recognition task, we utilize accuracy (ACC) as metric which can be defined as:

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^N I[\hat{y}_i = y_i] \quad (7)$$

where $\hat{y}_i$ is the predicted result and y_i is the ground-truth label. For the semantic localization task, we adopt Rsu, Rda, Ras, and Rmi as evaluation metrics, following [47]. For the image captioning task, we employ BLEU [32], ROUGE-L [7], METEOR [3], and CIDEr [38].

4.3. Implementation Details

We develop SARCLIP based on the OpenCLIP framework. Automatic mixed-precision (AMP) training is employed to

reduce memory usage. We adopt ResNet-50, ViT-B-32, and ViT-L-14 as image backbones, with learning rates set to $5e-4$, $5e-5$, and $5e-5$, respectively. ResNet-50 is trained for 30 epochs, while ViT-B-32 and ViT-L-14 are trained for 10 epochs. The batch size is set to 256. Training is accelerated using the Adam optimizer, combined with a linear warm-up and cosine learning rate schedule. Downstream experiments are conducted using the wise-ft [42] and SLM[47] framework. For the downstream target recognition task, we freeze the image backbone and fine-tune only the linear classification layers for 10,000 epochs. All experiments are conducted on two 80GB NVIDIA H100 GPUs. For SAR imagery caption task, we employ optical image caption dataset HQRS-210K [11] as pretrained data.

4.4. Comparing with other Methods

In this section, we compare our method with other state-of-the-art CLIP-based and CoCa-based method on downstream tasks.

4.4.1. Retrieval results on SARVLM-1M test dataset

As presented in Table 2, we compare our SARCLIP methods with several state-of-the-art VLFMs designed for remote sensing. The results clearly indicate that existing models exhibit poor performance in SAR image-text retrieval tasks. In contrast to standard CLIP models (e.g., Open-

Table 3. Recognition results on MSTAR-SOC and SAR-VSA dataset and improvements over the OpenCLIP(%).

Method	Backbone	Param.	dataset	
			MSTAR-SOC	SAR-VSA
OpenCLIP	ResNet-50	38	70.76	71.25
RemoteCLIP	ResNet-50	38	66.60	69.83
SARCLIP	ResNet-50	38	99.75	78.75
SARCLIP [‡]	ResNet-50	38	99.54(+40.67)	81.08(+13.79)
OpenCLIP	ViT-B-32	87	61.89	71.60
GeoRSCLIP	ViT-B-32	87	76.00	74.85
RemoteCLIP	ViT-B-32	87	69.97	73.56
SARCLIP	ViT-B-32	87	99.54	87.30
SARCLIP [†]	ViT-B-32	87	99.58	86.93
SARCLIP [§]	ViT-B-32	87	99.62	87.56
SARCLIP [‡]	ViT-B-32	87	99.58(+60.89)	87.77(+22.58)
OpenCLIP	ViT-L-14	304	78.80	81.58
GeoRSCLIP	ViT-L-14	304	82.26	80.07
GeoRSCLIP	ViT-H-14	632	79.84	78.39
RemoteCLIP	ViT-L-14	304	78.80	80.02
SARCLIP	ViT-L-14	304	99.54	88.92
SARCLIP [†]	ViT-L-14	304	99.62	89.75
SARCLIP [§]	ViT-L-14	304	99.71	88.56
SARCLIP [‡]	ViT-L-14	304	99.83(+26.68)	88.83(+8.88)

Table 4. Zero-shot classification results on FUSAR-ship and improvements over the RemoteCLIP baseline (%).

Method	Backbone	Param.	FUSAR-ship
SARCLIP	ResNet-50	38	15.51
SARCLIP [‡]	ResNet-50	38	14.69
RemoteCLIP	ViT-B-32	87	1.19
SARCLIP	ViT-B-32	87	8.97
SARCLIP [†]	ViT-B-32	87	11.25
SARCLIP [§]	ViT-B-32	87	3.12
SARCLIP [‡]	ViT-B-32	87	0.51
RemoteCLIP	ViT-L-14	304	9.72
SARCLIP	ViT-L-14	304	6.32
SARCLIP [†]	ViT-L-14	304	8.16
SARCLIP [§]	ViT-L-14	304	11.12
SARCLIP [‡]	ViT-L-14	304	12.41(+27.67)

Table 5. Semantic Localization results on AIR-SLT dataset.

Method	Backbone	Rsu $\uparrow$	Rda $\uparrow$	Ras $\downarrow$	Rmi $\uparrow$
RemoteCLIP	ViT-L-14	0.7707	0.6722	0.2876	0.7257
SARCLIP	ViT-L-14	0.5643	0.4830	0.5631	0.5000
SARCLIP [†]	ViT-L-14	0.5654	0.4531	0.5601	0.4889
SARCLIP [§]	ViT-L-14	0.5675	0.4792	0.5677	0.4981
SARCLIP [‡]	ViT-L-14	0.5741	0.4173	0.5731	0.4834

CLIP), our training strategy effectively transfers knowledge from the optical domain (e.g. RS5M) to the SAR domain (SARVLM-1M). Specifically, SARCLIP[‡] achieves a 15.25% mean recall while SARCLIP at 14.90%, demonstrating a significant improvement and highlighting the efficacy of our approach in multi-modal alignment in SAR

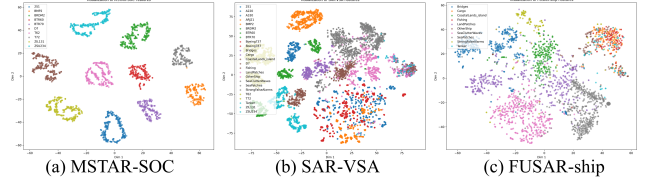


Figure 6. Feature space visualization of SARCLIP[†] image encoder on three downstream datasets (ViT-L-14).

domain.

4.4.2. Target recognition results on MSTAR-SOC and SAR-VSA dataset

Table 3 presents the performance of our SARCLIP models on downstream target recognition tasks after fine-tuning a linear layer, thereby demonstrating their visual understanding capabilities. The experimental results show that our SARCLIP series models consistently achieve superior performance. On the MSTAR-SOC dataset, SARCLIP[‡] achieves an impressive 99.83% accuracy, significantly outperforming optical VLFMs which typically hover around 80%. Similarly, on the SAR-VSA dataset, our SARCLIP[†] reaches 89.75% accuracy, notably surpassing OpenCLIP’s 81.58%. These results underscore the strong visual feature extraction our SARCLIP models for SAR imagery.

4.4.3. Zero-shot classification on FUSAR-ship dataset

Table 4 summarizes zero-shot classification on FUSAR-ship, evaluating SARCLIP without fine-tuning. RemoteCLIP performs poorly (e.g., 1.19% for ViT-B-32), whereas SARCLIP variants show much stronger zero-shot recognition. In particular, SARCLIP[‡] with a ViT-L-14 backbone reaches 12.41%, indicating better transfer to unseen SAR ship categories. This demonstrates that pre-training on the large-scale SARVLM-1M SAR vision-language corpus greatly enhances zero-shot generalization over models trained mainly on visible-light data.

4.4.4. Semantic localization on AIR-SLT dataset

Table 5 reports semantic localization results on the optical AIR-SLT dataset under a cross-domain setting. Our SARCLIP achieves 0.5000 Rmi compared to 0.7257 for RemoteCLIP, showing that SARCLIP retains a certain level of generalization to optical scenes. As illustrated in Figure 7, given query captions such as “four white planes” and “white airport”, our model successfully captures the plane features but misses the “white airport” region, which is mainly attributed to the intrinsic differences between SAR and optical imagery.

4.4.5. SAR imagery caption on SARVLM-1M

Table 8 presents the caption result from our SARCap framework. After high-quality optical remote sensing fine-tuning, CoCa-HQRS performs better than CoCa-Vanilla. We utilize a progressive training strategy to obtain our SAR imagery

Table 6. Image Caption results on SARVLM-1M test dataset with image encoder(ViT-L-14)(%).

Method	Image Encoder	Pretrain Data	Tune On	BLUE-1	BLUE-2	BLUE-3	BLUE-4	METEOR	ROUGE_L	CIDEr
CoCa-Vanilla	ViT-L-14	LAION	–	8.75	3.21	1.14	0.32	10.54	10.67	4.17
CoCa-HQRS	ViT-L-14	LAION	HQRS	10.79	4.02	1.14	0.39	8.21	13.11	7.47
SARCap	ViT-L-14	–	SARVLM-1M	22.42	14.44	10.45	7.79	19.23	21.76	24.32
SARCap *	ViT-L-14	LAION+HQRS	SARVLM-1M	26.10	17.13	11.93	8.47	22.18	26.15	29.58

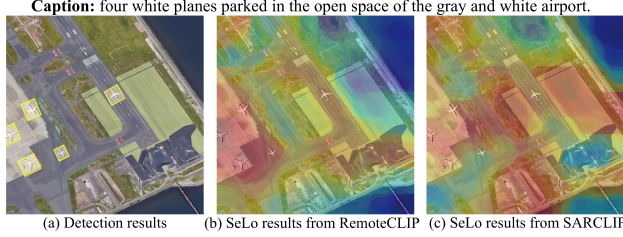


Figure 7. Visualization of results on semantic localization task from SARCLIP (ViT-L-14).

captioning model SARCap* which outperforms SARCap from scripts. As shown in Figure 8, our SARCap* generate fluent and meaningful caption for different SAR sences.

4.5. Ablation Studies

4.5.1. Effect of downmain transferring strategy

There is a large perspective gap between natural and optical remote sensing images, but a smaller gap between optical remote sensing and SAR scenes, making the transfer to SAR more challenging. Validation on tasks like image-text retrieval shows the effectiveness of our progressive knowledge transfer strategy.

4.5.2. Effect of the Training Encoder

This ablation study examines the impact of image and text encoders on downstream tasks. Table 7 shows that the text encoder alone achieves poor performance, with 5.23% mean recall on SARVLM-1M and 1.49% accuracy on FUSAR. The image encoder alone improves performance to 9.57% mean recall and 99.62% accuracy on MSTAR-SOC. Combining both encoders yields the best results, with 15.25% mean recall on SARVLM-1M and 99.83% accuracy on SOC, while zero-shot accuracy on FUSAR rises to 12.41%. This confirms that cross modal contrastive learning is key for better generalization.

4.5.3. Effect of the training layer

As shown in Figure 9, we evaluate the impact of varying the number of activated layers in the image and text encoders on the SARVLM-1M test set. The results reveal a consistent trend: model performance improves as more layers are activated. This highlights the importance of deep, expressive representations in both modalities for effective SAR image-text understanding.

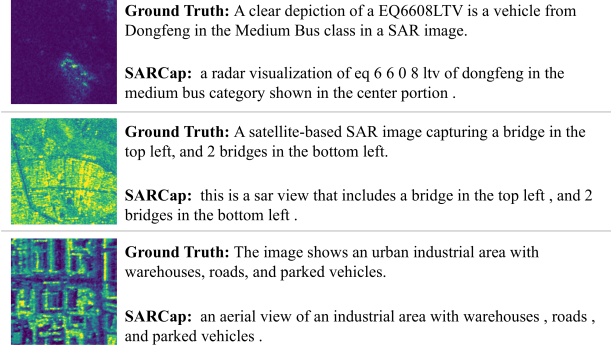


Figure 8. Visualization of SAR imagery caption results from SARCap.

	0	7	14	
	0.13	4.93	5.23	0
Image Encoder Layers	8.98	14.20	14.26	13
	9.57	15.26	15.25	26
	Text Encoder Layers			

Figure 9. Ablation study on training layers of SARCLIP[‡] on the SARVLM-1M test set (Mean Recall %).

Table 7. Ablation study result on several downstream tasks. Img. Enc. and Text Enc. denote the image encoder and text decoder respectively.(%)

Model Architecture	SARVLM-1M test	SOC	VSA	FUSAR
Img. Enc. Text Enc.	Mean Recall	ACC	ACC	ACC
✓	5.23	50.72	78.48	1.49
✓	9.57	99.62	88.94	10.85
✓	15.25	99.83	88.83	12.41

4.6. Visualization

As shown in Figure 6, we use t-SNE [28] to visualize SARCLIP[‡] (ViT-L-14) features on three downstream datasets. The well-separated clusters indicate strong feature discrimination and meaningful semantic representations.

5. Conclusion

We present SARVLM-1M, a large-scale SAR image-text dataset, and propose SARCLVLM, the first vision-language foundation model for SAR. Trained via a two-stage domain transfer strategy, SARCLIP effectively transfers knowledge

from optical remote sensing data to SAR imagery, achieving strong multi-modal alignment, target recognition and zero-shot and semantic localization capability while SARCap performs well in SAR imagery captioning. In the future, we will explore integrating multi-modal large language models and agent techniques for SAR image interpretation.

References

- [1] AFR Lab. The air force moving and stationary target recognition database. <https://www.sdms.afrl.af.mil/index.php?collection=mstar>, 1995. 4, 5
- [2] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024. 3
- [3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6
- [4] Sizhe Chen, Haipeng Wang, Feng Xu, and Ya-Qiu Jin. Target classification using the deep convolutional networks for sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4806–4817, 2016. 5
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1
- [6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 6
- [7] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, 2004, 2004. 6
- [8] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 3
- [9] Bin Deng, Puhong Duan, Xukun Lu, Zihao Wang, and Xudong Kang. Hyperspectral and sar image classification via graph convolutional fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [10] Yiguo He, Junjie Zhu, Yiyang Li, Qiangjuan Huang, Zhiyuan Wang, and Ke Yang. Rethinking remote sensing clip: Leveraging multimodal large language models for high-quality vision-language dataset. In *International Conference on Neural Information Processing*, pages 417–431. Springer, 2024. 3
- [11] Yiguo He, Junjie Zhu, Yiyang Li, Xiaoyu Zhang, Chunping Qiu, Jun Wang, Qiangjuan Huang, and Ke Yang. Enhancing remote sensing vision-language models through mllm and llm-based high-quality image-text dataset generation. *arXiv preprint arXiv:2507.16716*, 2025. 6
- [12] Xiyue Hou, Wei Ao, Qian Song, Jian Lai, Haipeng Wang, and Feng Xu. Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, 63(4):140303, 2020. 5
- [13] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Self-supervised learning for invariant representations from multi-spectral and sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:7797–7808, 2022. 2
- [14] Yuzhuo Kang, Zhirui Wang, Jiamei Fu, Xian Sun, and Kun Fu. Sfr-net: Scattering feature relation network for aircraft detection in complex sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021. 1
- [15] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 3
- [16] Anders Kusk, Adili Abulaitijiang, and Jorgen Dall. Synthetic sar image generation using sensor, terrain and target models. In *Proceedings of EUSAR 2016: 11th European Conference on Synthetic Aperture Radar*, pages 1–5. VDE, 2016. 4
- [17] Benjamin Lewis, Theresa Scarnati, Elizabeth Sudkamp, John Nehrbass, Stephen Rosencrantz, and Edmund Zelnio. A sar dataset for atr development: the synthetic and measured paired labeled experiment (sample). In *Algorithms for Synthetic Aperture Radar Imagery XXVI*, pages 39–54. SPIE, 2019. 4
- [18] Boying Li, Bin Liu, Lanqing Huang, Weiwei Guo, Zenghui Zhang, and Wenxian Yu. Opensarship 2.0: A large-volume dataset for deeper interpretation of ship targets in sentinel-1 imagery. In *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSAR DATA)*, pages 1–5. IEEE, 2017. 4
- [19] Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. 3
- [20] Weijie Li, Wei Yang, Yuenan Hou, Li Liu, Yongxiang Liu, and Xiang Li. Saratr-x: Towards building a foundation model for sar target recognition. *IEEE Transactions on Image Processing*, 2025. 2, 3, 5
- [21] Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *arXiv preprint arXiv:2403.06534*, 2024. 2, 3, 4
- [22] 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, 3, 6
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee.

- Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023. 3
- [24] Wang Liu, Zhiyu Wang, Xin Guo, Puhong Duan, Xudong Kang, and Shutao Li. Learning from noisy pseudo-labels for all-weather land cover mapping. *arXiv preprint arXiv:2504.13458*, 2025. 1
- [25] Yongxiang Liu, Weijie Li, Li Liu, Jie Zhou, Bowen Peng, Yafei Song, Xuying Xiong, Wei Yang, Tianpeng Liu, Zhen Liu, and Xiang Li. Atrnet-star: A large dataset and benchmark towards remote sensing object recognition in the wild, 2025. 2, 3, 4
- [26] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. 5
- [27] Zhiming Ma, Xiayang Xiao, Sihao Dong, Peidong Wang, HaiPeng Wang, and Qingyun Pan. Sarchat-bench-2m: A multi-task vision-language benchmark for sar image interpretation. *arXiv preprint arXiv:2502.08168*, 2025. 2, 3
- [28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (Nov):2579–2605, 2008. 8
- [29] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. *arXiv preprint arXiv:2312.06960*, 2023. 3
- [30] David Malmgren-Hansen, Anders Kusk, Jørgen Dall, Allan Aasbjerg Nielsen, Rasmus Engholm, and Henning Skriver. Improving sar automatic target recognition models with transfer learning from simulated data. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1484–1488, 2017. 4
- [31] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024. 3
- [32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6
- [33] 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, 2, 3
- [34] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 2, 3
- [35] Xian Sun, Yixuan Lv, Zhirui Wang, and Kun Fu. Scan: Scattering characteristics analysis network for few-shot aircraft classification in high-resolution sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. 5
- [36] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022. 3
- [37] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066, 2023. 2, 3
- [38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6
- [39] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 5
- [40] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 2, 3, 6
- [41] 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. *arXiv preprint arXiv:2504.03254*, 2025. 2, 3, 4
- [42] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 6
- [43] Youming Wu, Yuxi Suo, Qingbiao Meng, Wei Dai, Tiao Miao, Wenchao Zhao, Zhiyuan Yan, Wenhui Diao, Guocun Xie, Qingyang Ke, et al. Fair-csar: A benchmark dataset for fine-grained object detection and recognition based on single look complex sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3, 4
- [44] 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. 5
- [45] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 5
- [46] Miao Yu, Heqiang Yuan, Jialiang Chen, Chongyang Hao, Zhe Wang, Zhiqiang Yuan, and Bin Lu. Selo v2: Toward for higher and faster semantic localization. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 5
- [47] Zhiqiang Yuan, Wenkai Zhang, Chongyang Li, Zhaoying Pan, Yongqiang Mao, Jialiang Chen, Shuo Li, Hongqi

- Wang, and Xian Sun. Learning to evaluate performance of multimodal semantic localization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. [5](#), [6](#)
- [48] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021. [1](#)
- [49] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. [3](#)
- [50] Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, and Xiang Li. Rsar: Restricted state angle resolver and rotated sar benchmark. *arXiv preprint arXiv:2501.04440*, 2025. [1](#)
- [51] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. [2](#), [3](#), [6](#)