Rethinking Remote Sensing CLIP: Leveraging Multimodal Large Language Models for High-Quality Vision-Language Dataset

Yiguo He, Junjie Zhu, Yiying Li, Qiangjuan Huang✉
Zhiyuan Wang✉, and Ke Yang✉

Intelligent Game and Decision Lab, Beijing, China
yangke13@nudt.edu.cn

Abstract. The application of Contrastive Language-Image Pre-training (CLIP) models to remote sensing imagery has garnered significant attention. A key challenge lies in the scarcity of high-quality, large-scale, image-text paired training data. Recently, several works introduced extensive image-text datasets that leverage existing heterogeneous annotated datasets for remote sensing and trained their vision-language foundation models. However, due to the rudimentary methods used for creating text descriptions, the quality of datasets produced by these methods is suboptimal, requiring larger volumes of training data, while only yielding modest performance improvements. In this paper, we primarily propose the employment of Multimodal Large Language Models (MLLMs) to generate higher-quality captions. Specifically, we carefully design an Annotation to Instruction (A2I) module to bridge existing annotations for detection, segmentation, and classification tasks with the input requirements of grounding MLLMs. In addition, we propose a refined rule-based text caption generation method and incorporate 8 classification datasets and 1 multispectral RGB composite image dataset to enhance the diversity of data. Finally, we have created RSM-ITD, a high-quality, large-scale remote sensing image-text dataset, containing approximately 480K image-text pairs. The experimental results suggest that, despite the smaller size of our proposed dataset, the CLIP models trained on it achieve better results than SOTA methods in tasks like zero-shot classification, retrieval, and semantic localization. Dateset, pre-trained models, and codes will be released upon publication.

Keywords: Remote Sensing · Image-Text Paired Dataset · CLIP · Vision Language Foundation Model

1 Introduction

Visual language models (VLMs) such as CLIP [1], pre-trained on large-scale image-text data, demonstrate strong generalization capabilities and can achieve competitive performance across various downstream tasks. Recently, the application of VLMs for remote sensing (RS) and aerial imagery has garnered significant attention [2–4] for its superior capability. Liu *et al.* [2] demonstrated that large CLIP models, trained with extensive pre-training image-text paired RS data, perform expressively on various RS applications. The key to achieving success in this area lies in the high-quality, large-scale RS image-text paired data. Unlike natural images, RS images and their associated text descriptions cannot be effectively sourced from the public internet. Additionally, manually annotating aerial images requires specialized knowledge and is extremely time-consuming [5]. This is more challenging for textual caption annotating, as RS images often lack detailed content, making it difficult even for experts to provide diverse annotations.

To address this gap, Liu *et al.* [2] proposed a data scaling method that converts precise annotations from object detection datasets into English sentences, thus creating the first large-scale RS image-text dataset. Their method offers several advantages. Firstly, it effectively leverages high-quality RS images and precise manual annotations from existing public datasets. Secondly, it rapidly generates a large volume of RS image-text data at a low cost. However, this approach has significant limitations. Firstly, the Box-to-Caption (B2C) algorithm [2] describes only a single category within the image, omitting descriptions for the other categories. This might lead to ambiguity in the construction of positive and negative pairs when training, resulting in poor annotation quality [6,7]. Secondly, the descriptions lack contextual information beyond annotations, such as details about aerial scenes, which can help reduce ambiguity and enhance the alignment between the image and text. Thirdly, the rule-based generated text descriptions tend to be repetitive and lack natural, meaningful, and multi-semantic content [2,3].

Wang *et al.* [3] connected RS images from the Google Earth Engine (GEE) platform with information from the OpenStreetMap (OSM) database using geographic coordinates. They applied a series of rules to convert labels into captions. However, these captions were mechanically assembled and lacked natural, meaningful sentences. Zhang *et al.* [4] proposed RS5M, which employs keyword filtering of natural image-text datasets. However, this method introduces significant noise into the data. Consequently, despite their large scale, this dataset suffers from low quality.

Recently, researchers started to explore the use of Multimodal Large Language Models (MLLMs) for generating captions for RS images. Zhang *et al.* [4] utilized the BLIP-2 model on category-annotated RS datasets, highlighting the potential of MLLMs in this field. However, BLIP-2 [8] can only generate descriptions from class-level label instructions, lacking fine-grained annotation information that can be extracted from large-scale object detection and semantic segmentation datasets. To the best of our knowledge, no attempts have successfully employed MLLMs to create large-scale, high-quality RS image-text datasets.

To leverage MLLMs for higher-quality captions of detection and segmentation datasets, the MLLMs need to accept information such as object locations as input. Fortunately, MLLMs with visual grounding capabilities can perceive bounding box (bbox) annotations that have been proposed, such as Kosmos-2 [9]. However, these grounding MLLMs can not directly accept box or segmentation annotation as inputs. Therefore, we designed an **Annotation to Instruction (A2I)** module to convert classification, detection, and segmentation annotation to the instructions required by grounding MLLMs. By leveraging these instructions, grounding MLLMs can generate much more accurate and detailed captions for RS images.

In addition to using MLLMs for generating image-text datasets, we also improved the RemoteCLIP's [2] Box-to-Caption (B2C) algorithm [2]by taking all categories into account and developed the **Annotation to Caption (A2C)** algorithm. The A2C minimizes image-text category ambiguity and leverages precise manual annotations from existing public datasets. In summary, our proposed dataset construction method includes A2C and MLLMs Generation, as shown in Figure 1.

Our newly introduced dataset, named RSM-ITD (Remote Sensing Multi-source Image-Text Dataset), comprises 210,515 images and 476,342 text captions. It provides natural and meaningful captions, an improvement over the captions in the RemoteCLIP [2] and SkyScript [3] datasets. Distinguished from RS5M [4], RSM-ITD reduces noise and enhances data quality by utilizing object detection boxes and semantic segmentation annotations from existing well-annotated datasets, thus ensuring more accurate and detailed captions.

We employed full fine-tuning to train the CLIP model on RSM-ITD, resulting in RSM-CLIP. Experimental results show that RSM-CLIP significantly enhances performance across various downstream tasks compared to the original CLIP. Furthermore, despite being trained on a much smaller dataset, RSM-CLIP outperforms SkyCLIP [3], RemoteCLIP, and GeoRSCLIP in multiple tasks, highlighting the high quality and effectiveness of our proposed RSM-ITD.

In zero-shot classification (ZSC) tasks, RSM-CLIP achieved a significant average top-1 accuracy improvement of 17.02% over CLIP across three test datasets. In zero-shot retrieval (ZSR) tasks, it showed an increase in mean recall by 8.65% on three benchmarks. In the more challenging task of RS semantic localization (SeLo [10]), Rmi improved by 3.76%. After fine-tuning on downstream datasets, RSM-CLIP demonstrated further enhancements, achieving average mean recall improvements of 1.68% with ViT-B-32 and 2.76% with ViT-L-14 compared to RemoteCLIP on datasets such as RSITMD, RSICD, and UCM.

Remarkably, RSM-CLIP achieved superior performance despite using only one-tenth the training data compared to GeoRSCLIP. It resulted in a 1.43% higher average top-1 accuracy in zero-shot classification (ZSC) tasks, a 2.18% higher average mean recall in zero-shot retrieval (ZSR) tasks, and a 0.47% increase in Rmi for RS semantic localization (SeLo) tasks. Furthermore, RSM-CLIP (ViT-L-14) achieved superior performance compared to the previous SoTA performances on benchmarks like RSITMD, RSICD, and UCM. Our contributions can be summarized as follows.

- We propose a novel method for constructing a RS image-text paired dataset, resulting in a high-quality dataset named RSM-ITD, which comprises 476,342 image-text pairs.

- Based on the RSM-ITD dataset, we develop a RS vision language pre-trained model named RSM-CLIP, which delivers a robust vision language representation for RS applications. The dataset and models will be made publicly available upon publication.
- The effectiveness of RSM-CLIP has been validated across a variety of downstream RS tasks, where it consistently outperforms previous state-of-the-art RS models.

2 Dataset Construction

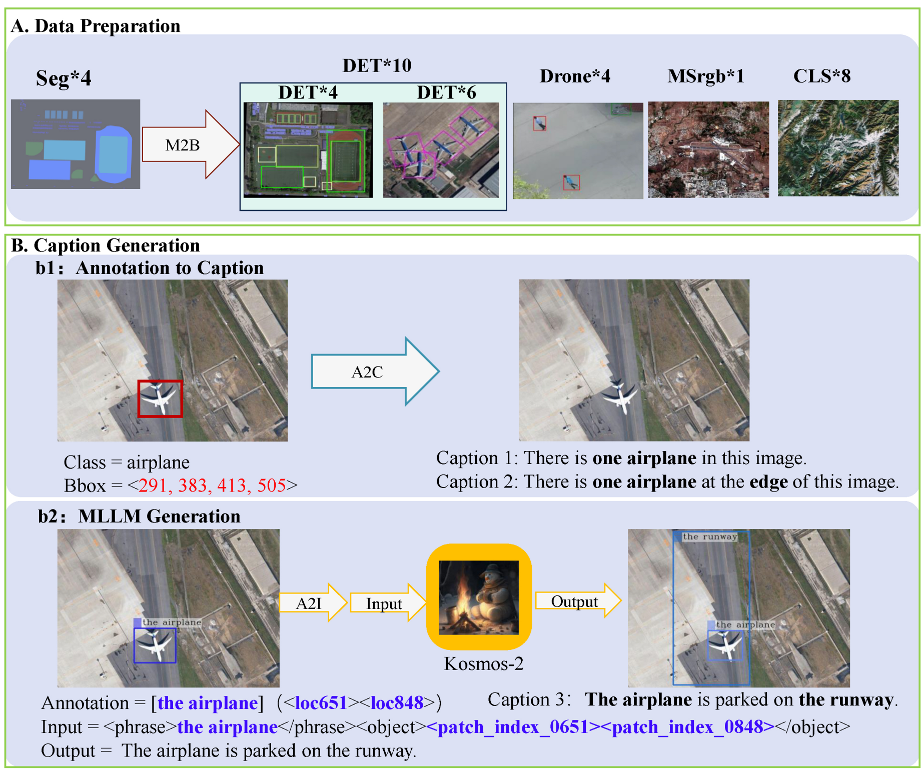

Fig. 1. Pipeline of the RSM-ITD construction. M2B transforms Seg*4 to DET*4. A2C transforms annotations into captions by rule-based method. A2I converts annotations into instructions recognizable by Kosmos-2. <loc651><loc848> represents the bounding box of the airplane in the left image, while Input indicates the actual instruction format which is used to guide Kosmos-2 to generate captions.

2.1 Data Preparation

Before generating captions, we collected 23 datasets and thoroughly cleaned the images and annotations. The datasets are : (1) a multispectral composite image dataset(MSrgb*1): fMoW [11], (2) 4 UAV aerial object detection datasets (Drone*4): AU-AIR [12], CARPK [13], Stanford Drone [14], VisDrone [15], (3) 4 satellite semantic segmentation datasets (Seg*4): iSAID [16], LoveDA [17], Potsdam [18], Vaihingen [19], (4) 6 RS object detection datasets (DET*6): DIOR [20], DOTA [21], HRRSD [22], HRSC [23], LEVIR [24], RSOD [25], and (5) 8 RS scene classification datasets (CLS*8): NWPU-RESISC45 [26], AID [5], RSI-CB128 [27], RSI-CB256 [27], WHURS19 [28], OPTIMAL-31 [29], MLRSNet [30], EuroSAT [31].

After being transformed into DET*4 by the M2B algorithm [2], Seg*4 combined with DET*6 to form DET-10. **We initially gathered the Drone*4, the Seg*4, and DET*6 used in RemoteCLIP to enable a fair comparison of data construction methods. CLS*8 and MSrgb*1 were collected to further increase the diversity of the data.** For the fMoW dataset, we only selected samples from the validation set, as GeoRSCLIP [4] had already generated captions for the training set, thereby avoiding redundancy.

In terms of images, we first removed unannotated images from each dataset. For images that are too large (greater than 4,000,000 pixels), we employ a sliding window approach to partition them into several non-overlapping smaller image patches. A strict deduplication method using p-hash and URLs has been employed to prevent data leakage [2].

Ultimately, about $\frac{1}{5}$ samples were removed. In terms of annotations, denoising was also performed. For example, we removed annotations labeled as "ignored regions" and "others" in the VisDrone dataset. The wording of the original annotations was adjusted where necessary. For example, we changed the annotation for people from "Human" to "person".

2.2 Caption Generation

Annotation to Caption (Rule-based). Firstly, we use Mask-to-Box (M2B) algorithm [2] to extract the coordinates (xmin, ymin, xmax, ymax) of objects annotated in the Seg*4 into bounding boxes, converting them into DET*4. When describing the positions of objects, we define the central area as the rectangular region spanning from 1/4 to 3/4 of the image's width and height, with the remaining area defined as the edge area. Annotation to Caption (A2C) includes the following rules to generate captions for "DET-10":

- **Rule 1**: Describe all objects annotated in the image.
  Example: There are three cars and two trucks in this image.
- **Rule 2**: Describe objects located both in the center and at the edge of this image.
  Example: There are three cars in the center of this image and two trucks at the edge of this image.

The pseudo-code for the A2C algorithm is shown in Algorithm 1. These improved rules ensure that all annotated objects are included in the captions, alleviate the risk of category ambiguity [6], and provide a more complete description. Experimental results demonstrate that our rule refinement strategy is highly effective (Figure 3).

Algorithm 1: Annotation to Caption (A2C)

```
Function Annotation_to_Caption(bbox_list):
    captions ← [ ];
    // Caption 1: Describe all objects annotated in the image.
    class_counts ← count_categories(bbox_list);
    caption_parts ← generate_caption_parts(class_counts);
    captions.append("There are " + join(caption_parts, ", ") + ".");
    // Caption 2: Describe objects located both in the center and at
       the edge of this image.
    center_counts, edge_counts ← count_center_and_edge(bbox_list);
    center_parts ← generate_caption_parts(center_counts);
    edge_parts ← generate_caption_parts(edge_counts);
    captions.append("There are " + join(center_parts, ", ") + " in the center,
     and " + join(edge_parts, ", ") + " at the edge of this image.");
```

MLLMs Caption Generation. Several MLLMs are capable of perceiving annotated information and generating textual descriptions for images, such as chatGPT-4V [32], BLIP-2 [8], and Llava [33]. But chatGPT-4V is closed-sourced and incurs high usage costs. Other MLLMs, like BLIP-2 and Llava, lack visual grounding capabilities. Fortunately, Kosmos-2 is open-sourced and can leverage textual instructions, as well as perceive and link the annotated bounding boxes to the generated captions. It provides more accurate, informative, and comprehensive caption descriptions for images. Kosmos-2 is selected for our MLLMs caption generation.

The instruction format has a significant impact on the performance of MLLMs. To achieve better results, we explored the impact of different instruction templates before formal experiments. We randomly selected 1,000 images from the overall dataset to test the impact of 10 different instruction templates of Kosmos-2. After careful designation and extensive experiments, we adopted the following 3 instruction templates:

① Describe this image with [class] in detail:
② Describe this image with [class + bbox] in detail:
③ Where is/are the [class + bbox]? Answer:

For images with only class-level annotations, we use ① to generate accurate and comprehensive descriptions. For images with one or two bounding boxes, we use both ② and ③ to obtain accurate positional information and comprehensive image descriptions. For images with more than two annotated bboxes, we use ① to achieve comprehensive image descriptions.

Since Kosmos-2 cannot directly accept bbox or semantic segmentation mask as inputs, we propose the Annotation-to-Instruction (A2I) algorithm. It automatically converts categories, object detection bounding boxes, and semantic segmentation masks into instructions that Kosmos-2 can perceive. The pseudocode is shown in Algorithm 2. These instructions, along with the images, are then fed into the Kosmos-2 model to generate corresponding captions. Experimental results demonstrate the high effectiveness of our MLLM generation strategy (Figure 3).

Algorithm 2: Annotation to Instruction (A2I)

```
Function annotation_to_instruction(dataset, dataset_type):
    prompts ← extract_prompts(dataset, dataset_type)
    instructions ← generate_instructions(prompts, dataset_type)
Function extract_prompts(dataset, dataset_type):
    prompts ← []
    for each image in dataset do
        if dataset_type == 'classification' then
            prompts.append(image['category'])
        else if dataset_type == 'object detection' then
            prompts.append(f"obj[category] [location]")
        else if dataset_type == 'segmentation' then
            bbox_list ← M2B(masks) prompts.append(f"obj[category]
              [location]")
    return prompts
Function generate_instructions(prompts, dataset_type):
    instructions ← []
    for each description in prompts do
        if dataset_type == 'classification' then
            Instruction = "Describe image with [prompt] in detail:"
        else
            if object_number <= 2 then
                Instruction1 = "where is/are the [prompt]?"
                Instruction2 = "Describe image with [prompt] in detail:"
            else
                Instruction = "Describe image with [prompt] in detail:"
```

2.3 Dataset Description

Ultimately, we generated 230,766 captions for 115,383 images using the rule-based method and 245,576 captions for 210,515 images using Kosmos-2. The resulting dataset contains a total of 210,515 images and 476,342 image-text pairs. On average, each image description contains 57 words, providing detailed information. The captions in RSM-ITD include rich semantic information, such as image scene, objects, and their positional details.

3 Experiments

3.1 Experiment

Implementation Details. The CLIP ViT-B-32 and CLIP ViT-L-14 models were fully fine-tuned on RSM-ITD. The training code is based on openCLIP[1]. To emphasize the intrinsic effectiveness of our dataset, we did not use any data augmentation techniques or perform specific hyperparameter tuning during training. We randomly selected 10% of the RSM-ITD data as the validation set, with the remaining data used for training. The training process utilized a cosine learning rate scheduler, mixed precision (AMP) mode, and the AdamW optimizer [34]. Modal interaction was conducted using the InfoNCE loss [35]. For ViT-B-32, the learning rate was set to 2e-5, and the batch size to 256. For ViT-L-14, the learning rate was set to 1e-6, and the batch size to 32. Both models had their weight decay set to 1. The training was conducted on a single RTX 4090 24 GB GPU. Ultimately, we obtained RSM-CLIP (ViT-B-32) and RSM-CLIP (ViT-L-14). *Compared to RemoteCLIP's 233.4 hours of training time, our RSM-CLIP (ViT-L-14) training only required approximately 6 hours.* All RS CLIP mod-

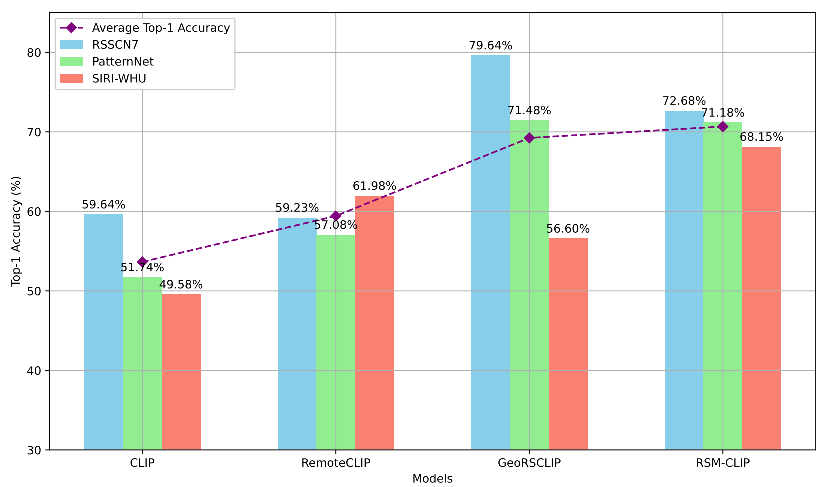

Fig. 2. Comparison of different models' test results on ZSC task

[1] https://github.com/mlfoundations/open_clip

els compared with RSM-CLIP were trained using the fully fine-tuned method. The image backbone of these models is all ViT-B-32 unless we specify. *Experimental results indicate that RSM-CLIP shows superior capabilities while using only 10% of the data of GeoRSCLIP, 60% of SkyCLIP-30's data, and 65% of RemoteCLIP's data.* We evaluated RSM-CLIP on 3 vision-language tasks.

Zero-shot Classification (ZSC). The definition of zero-shot learning in studies such as CLIP, RemoteCLIP, and GeoRSCLIP has been extended from generalizing to unseen object categories to generalizing to unseen datasets, serving as a proxy for performing unseen tasks. Adhering to this definition, we utilized the complete datasets of RSSCN7 [36], SIRI-WHU [37] and PatternNet [6] as test data, evaluating the performance using top-1 accuracy as the metric.

As shown in Figure 2, RSM-CLIP achieved a 17.02% improvement in average top-1 accuracy over the vanilla CLIP, an 11.24% improvement over RemoteCLIP, and a 1.43% improvement over GeoRSCLIP. In addition, compared to GeoRSCLIP, RSM-CLIP demonstrated more robust testing results across 3 different datasets.

RS Cross-modal Text–Image Retrieval (RSCTIR). RSCTIR includes image-to-text retrieval and text-to-image retrieval. RSITMD [38], RSICD [7], and UCM [39] datasets are commonly used for this task. We also define zero-shot retrieval as the ability to generalize to unseen datasets. The evaluation metrics in this paper are recall@1 and mean recall.

Table 1. Results of Zero-shot Retrieval task. The best result is in **bold**.

Test Dataset	Models	Training pairs	I2T R@1	T2I R@1	mean recall
RSITMD	CLIP	-	9.51	8.81	24.19
	SkyCLIP-30	780K	11.73	10.19	30.67
	GeoRSCLIP	5 Million+	19.03	14.16	35.68
	RSM-CLIP	**476,342**	17.7	15.66	36.44
	RSM-CLIP (ViT-L/14)	**476,342**	**23.45**	**16.86**	**39.43**
RSICD	CLIP	-	5.31	5.78	15.74
	SkyCLIP	780K	8.97	5.85	21.83
	GeoRSCLIP	5 Million+	11.53	9.52	26.18
	RSM-CLIP	**476,342**	11.16	9.33	25.32
	RSM-CLIP (ViT-L/14)	**476,342**	**13.17**	**10.23**	**27.69**
UCM	CLIP	-	9.52	8.67	33.13
	GeoRSCLIP	5 Million+	18.57	13.81	47.76
	RSM-CLIP	**476,342**	20.48	14.95	50.25
	RSM-CLIP (ViT-L/14)	**476,342**	**21.90**	**15.52**	**51.89**

As shown in Table 1, the average mean recall for RSM-CLIP was 4.6% higher than SkyCLIP-30 across the RSITMD and RSICD datasets, and 2.18% higher than GeoRSCLIP across the 3 test sets. RSM-CLIP (ViT-L-14) outperformed all ViT-B-32 models in I2T R@1, T2I R@1, and mean recall, indicating that larger models can more effectively leverage the rich information in RSM-ITD.

Table 2. Results of RSCTIR task. The best result is in **bold**.

Image Backbone	Test Dataset	Models	Training Pairs	I2T R@1	T2I R@1	mean recall
ViT-B-32	RSITMD	RemoteCLIP	828,725	27.88	22.17	49.38
		GeoRSCLIP	5 Million+	30.09	23.54	50.10
		RSM-CLIP	**544,907**	**32.08**	**24.07**	**52.12**
	RSICD	RemoteCLIP	828,725	17.02	13.71	35.26
		GeoRSCLIP	5 Million+	**22.14**	15.26	38.00
		RSM-CLIP	**544,907**	21.87	**15.48**	**38.05**
	UCM	RemoteCLIP	828,725	**20.48**	**18.67**	**56.36**
		RSM-CLIP	**544,907**	20.00	18.38	56.05
ViT-L-14	RSITMD	RemoteCLIP	828,725	28.76	23.76	50.52
		RSM-CLIP	**544,907**	**30.53**	**27.21**	**52.29**
	RSICD	RemoteCLIP	828,725	18.39	14.73	36.35
		RSM-CLIP	**544,907**	**22.60**	**17.04**	**39.78**
	UCM	RemoteCLIP	828,725	19.05	17.71	54.68
		RSM-CLIP	**544,907**	**21.9**	**19.43**	**57.71**

To fairly compare with RemoteCLIP and GeoRSCLIP fine-tuned on the downstream test datasets, we fine-tuned our models on the RET-3[2] data provided by RemoteCLIP. As shown in Table 2, After fine-tuning the RSM-CLIP on the RET-3, the average mean recall value of RSM-CLIP is 1.68% higher than RemoteCLIP and 1.03% higher than GeoRSCLIP. The average mean recall value of RSM-CLIP (ViT-L-14) is 2.76% higher than RemoteCLIP (ViT-L-14). It is evident that larger models have better transfer learning capability.

Semantic Localization (SeLo) : SeLo task is considered a more advanced retrieval task than RSCTIR. AIR-SLT is the only semantic localization test set in RS. The evaluation metrics are Rsu, Ras, Rda, and Rmi.

In the SeLo task, RSM-CLIP achieved a 3.76% improvement in the comprehensive metric Rmi compared to CLIP and outperformed GeoRSCLIP by 0.47%. The significant performance improvement of RSM-CLIP in the SeLo task is im-

[2] RET-3 provided by RemoteCLIP includes 68,565 image-text pairs obtained by deduplicating the combined training sets of RSITMD, RSICD, and UCM.

Table 3. Results of SeLo task. The best result is in **bold**.

Method	Training pairs	Rsu ↑	Ras ↓	Rda ↑	Rmi ↑
CLIP	-	0.7188	0.3006	0.6992	0.7071
RemoteCLIP	828,725	0.7365	0.3008	0.6928	0.7125
GeoRSCLIP	5 Million+	**0.7546**	0.2610	0.7180	0.7400
RSM-CLIP	**476,342**	0.7469	**0.2518**	**0.7364**	**0.7447**

portant evidence of the rich semantics, diverse scenes, and spatial relationship information contained in RSM-ITD. The results are shown in Table 3.

3.2 Ablation Study

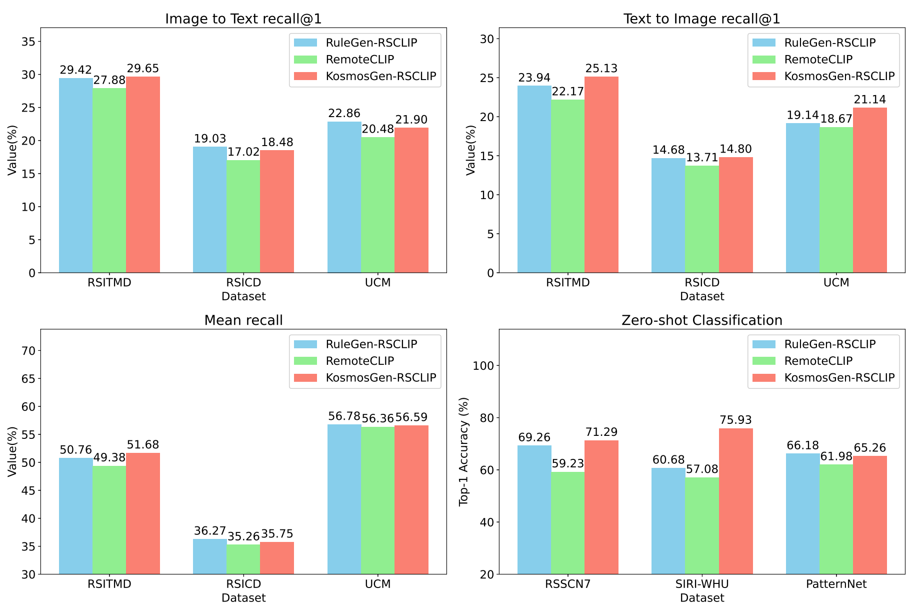

Fig. 3. Influence of Rule-improved and MLLM generation strategies. Compared to the method of RemoteCLIP, our two methods show significant performance improvements across all metrics.

Influence of the Rule-Improved Strategy. To ensure a fair comparison of our rule-based generation method, we used the same image sources as RemoteCLIP and generated captions using our rules. Then, we fine-tuned CLIP in the same manner as RemoteCLIP and referred to the resulting model as

RuleGen-RSCLIP. Experimental results show that RuleGen-RSCLIP outperforms RemoteCLIP across various metrics (Figure 3).

Influence of the MLLM Generation Strategy. We used Kosmos-2 to generate captions for the homologous dataset of RemoteCLIP. After fine-tuning the CLIP model as RemoteCLIP did, we referred to the resulting model as KosmosGen-RSCLIP. Experimental results show that KosmosGen-RSCLIP outperforms RemoteCLIP across various metrics (Figure 3).

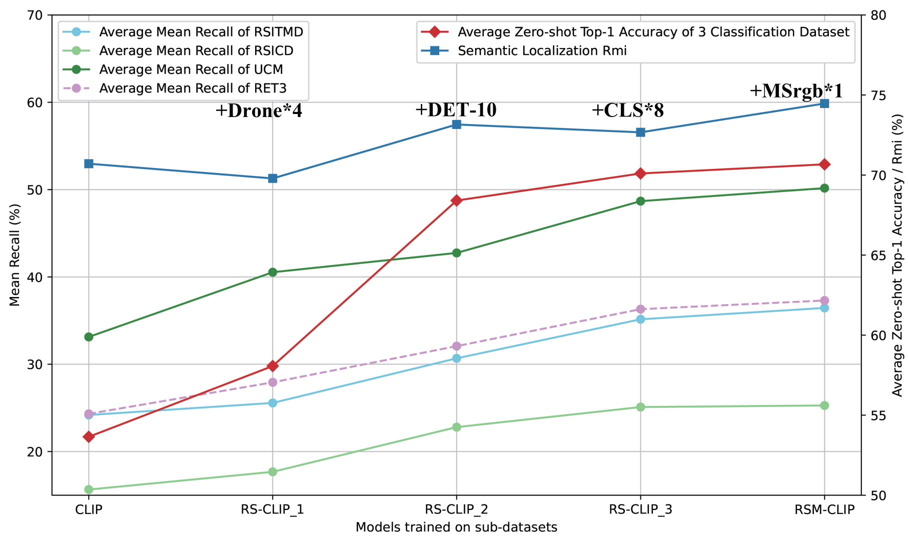

Fig. 4. Influence of Sub-Datasets. All types of datasets enhance CLIP's performance on ZSC and RSCTIR tasks. However, object detection and MSrgb datasets contribute positively to the SeLo task, while drone datasets and RS classification datasets contribute negatively. The 'RET3' in this chart refers to the RSITMD, RSICD, and UCM datasets.

Influence of Different Sub-Datasets. To determine the effect of various sub-datasets, we divided the RSM-ITD into sub-datasets based on their sources. These sub-datasets are Drone*4, DET-10, CLS*8, and MSrgb*1. We first trained the CLIP model using Drone*4, resulting in RS-CLIP_1. Then, we trained it using Drone*4 + DET-10, resulting in RS-CLIP_2. Next, we used Drone*4 + DET-10 + CLS*8, resulting in RS-CLIP_3. Finally, we trained the model using the entire RSM-ITD dataset (after adding MSrgb*1), resulting in RSM-CLIP.

Figure 4 shows the results of three tasks. It indicates that various types of datasets generally enhance the model's performance. However, the impact on the SeLo task varies. Drone*4 negatively affects it, likely due to inconsistencies in data distribution. CLS*8 negatively affects it, likely due to a lack of high intra-class diversity in the images (compared to other types of datasets) and the absence of fine-grained annotations (only class labels). DET-10 and MSrgb*1 significantly improve it, likely due to their bounding box annotations providing fine-grained spatial descriptions. DET-10 yields the most significant performance improvements on all tests, most likely because these object detection datasets provide high-quality, diverse RS images along with detailed fine-grained information.

4 Conclusion

In this paper, we present RSM-ITD, a large-scale, high-quality RS image-text paired dataset. Based on this dataset, we trained RSM-CLIP models to demonstrate the effectiveness of the dataset in various RS tasks. We verify that by carefully designing rules and instruction generation methods, the MLLMs caption generation method is a very efficient method for captioning RS imagery. Training samples of drone aerial images, satellite imagery, and multispectral composite RGB images all enhance the RS classification and retrieval capabilities of the RSM-CLIP. In addition, bounding boxes and segmentation mask annotations of satellite imagery can guide MLLM to generate captions with more fine-grained and location-related information. Our research alleviates the scarcity of large-scale, high-quality RS image-text datasets and advances the perception of RS RGB imagery. Our proposed dataset and pre-trained models can serve as a foundational resource for the RS community to advance research in RS representation.

5 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62006241, 62206307.

References

1. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
2. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE TGRS (2024)
3. Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R.: Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: AAAI. vol. 38, no. 6, pp. 5805–5813 (2024)
4. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. arXiv preprint arXiv:2306.11300 (2023)
5. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE TGRS. vol. 55, no. 7, pp. 3965–3981 (2017)

6. Zhou, W., Newsam, S., Li, C., Shao, Z.: PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS Journal of Photogrammetry and Remote Sensing. vol. 145, pp. 197–209 (2018)
7. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE TGRS. vol. 56, no. 4, pp. 2183–2195 (2017)
8. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742 (2023)
9. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
10. Yuan, Z., Zhang, W., Li, C., Pan, Z., Mao, Y., Chen, J., Li, S., Wang, H., Sun, X.: "Learning to Evaluate Performance of Multi-modal Semantic Localization," *IEEE TGRS*, doi: 10.1109/TGRS.2022.3207171.
11. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: CVPR. pp. 6172–6180 (2018)
12. Vujasinović, S., Becker, S., Breuer, T., Bullinger, S., Scherer-Negenborn, N., Arens, M.: Integration of the 3d environment for UAV onboard visual object tracking. Applied Sciences. vol. 10, no. 21, pp. 7622 (2020)
13. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: ICCV. pp. 4145–4153 (2017)
14. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 549–565 (2016)
15. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.
16. Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.S., Bai, X.: isaid: A large-scale dataset for instance segmentation in aerial images. In: Proceedings of the IEEE/CVF CVPR Workshops. pp. 28–37 (2019)
17. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1, 2021. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf
18. Potsdam dataset. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (2012)
19. Vaihingen dataset. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx (2012)
20. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing. vol. 159, pp. 296–307 (2020)
21. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images. In: CVPR. pp. 3974–3983 (2018)
22. Zhang, Y., Yuan, Y., Feng, Y., Lu, X.: Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. IEEE TGRS. vol. 57, no. 8, pp. 5535–5548 (2019)

23. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: ICPRAM. vol. 2, pp. 324–331 (2017)
24. Chen, H., Shi, Z.: A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing. vol. 12, no. 10, pp. 1662 (2020)
25. Sun, W., Dai, L., Zhang, X., Chang, P., He, X.: RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. Applied Intelligence. pp. 1–16 (2022)
26. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE. vol. 105, no. 10, pp. 1865–1883 (2017)
27. Li, H., Dou, X., Tao, C., Wu, Z., Chen, J., Peng, J., Deng, M., Zhao, L.: RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. Sensors. vol. 20, no. 6, pp. 1594 (2020)
28. Xia, G.S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Maître, H.: Structural high-resolution satellite image indexing. In: ISPRS TC VII Symposium-100 Years ISPRS. vol. 38, pp. 298–303 (2010)
29. Wang, Q., Liu, S., Chanussot, J., Li, X.: Scene classification with recurrent attention of VHR remote sensing images. IEEE TGRS. vol. 57, no. 2, pp. 1155–1167 (2018)
30. Qi, X., Zhu, P., Wang, Y., Zhang, L., Peng, J., Wu, M., Chen, J., Zhao, X., Zang, N., Mathiopoulos, P.T.: MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. ISPRS Journal of Photogrammetry and Remote Sensing. vol. 169, pp. 337–350 (2020)
31. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE JSTARS. vol. 12, no. 7, pp. 2217–2226 (2019)
32. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., *et al.*: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS. vol. 36 (2024)
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *ICLR* (2017). Available: https://api.semanticscholar.org/CorpusID:53592270
35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
36. Zou, Q., Ni, L., Zhang, T., Wang, Q.: Deep learning based feature selection for remote sensing scene classification. IEEE GRSL. vol. 12, no. 11, pp. 2321–2325 (2015)
37. Zhao, B., Zhong, Y., Xia, G.S., Zhang, L.: Dirichlet-derived multiple topic scene classification model fusing heterogeneous features for high spatial resolution remote sensing imagery. IEEE TGRS. vol. 54, no. 4, pp. 2108–2123 (2016)
38. Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X.: Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE TGRS*, vol. 60, pp. 1-19, 2022. doi: 10.1109/TGRS.2021.3078451
39. Qu, B., Li, X., Tao, D., Lu, X.: Deep semantic understanding of high resolution remote sensing image. *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pp. 1-5, 2016. doi: 10.1109/CITS.2016.7546397