A RECIPE FOR IMPROVING REMOTE SENSING VLM ZERO SHOT GENERALIZATION

Aviad Barzilai[*], Yotam Gigi[*], Amr Helmy[*], Vered Silverman[*], Yehonathan Refael[*], Bolous Jaber, Tomer Shekel, George Leifman, Genady Beryozkin
Google Research

ABSTRACT

Foundation models have had a significant impact across various AI applications, enabling use cases that were previously impossible. Contrastive Visual Language Models (VLMs), in particular, have outperformed other techniques in many tasks. However, their prevalence in remote sensing (RS) is still limited, due to the scarcity of diverse remote-sensing visual-language datasets. In this work we introduce two novel image-caption datasets for training of remote sensing foundation models. The first dataset pairs aerial and satellite imagery with captions generated by Gemini using landmarks extracted from Google Maps. The second dataset utilizes public web images and their corresponding alt-text, filtered for the remote sensing domain, resulting in a diverse dataset with greater breadth in image styles and subject matter. These datasets are used to pre-train the MaMMUT (Kuo et al., 2023) VLM architecture, resulting in state-of-the-art generalization performance in zero-shot cross-modal retrieval on well-known public benchmarks. Finally, we present our ongoing research to distill image-level knowledge gained in the VLM contrastive training procedure to enhance the model's localization ability. Specifically, we iteratively generate pseudo-labels for image regions based on the model's attention maps and use these labels for further training. To mitigate noisy attention maps and create robust segmentation masks, we introduce a novel attention-pooling mechanism called the Smooth-Attention-Operation.

1 INTRODUCTION

Foundation models have demonstrated exceptional performance serving as a basis for diverse downstream tasks by training on large-scale datasets. Yet, applying them to remote sensing tasks presents unique challenges, as remote sensing data exhibits distinct characteristics, including complex spatial relationships from orbital viewpoints and lower spatial resolution compared to standard ground-level imagery. These differences hinder existing models optimized for ground-level perspectives and high-resolution images. Furthermore, the remote sensing domain faces labeled data scarcity, including limited availability of paired text-image examples. The paucity of datasets with rich textual descriptions accompanying remote sensing images limits the ability to build models that comprehend the context and nuances of the visual data.

Self-supervised learning techniques, such as contrastive spatial pre-training explored in (Mai et al., 2023), offer a promising avenue to mitigate this issue by leveraging the inherent spatial information within readily available unlabeled remote sensing data. As the study in Bourcier et al. (2024) demonstrates, leveraging metadata supervision offers a promising avenue to overcome this limitation, learning effective representations even when paired text-image data is scarce. These challenges highlight the need for tailored solutions that address the specific nature of remote sensing data. Similarly, a novel VLM, specifically designed for remote sensing is needed. Such a model will enable more accurate and insightful analysis for a wide range of applications using remote sensing imagery, including open-vocabulary object detection, zero-shot image segmentation, text-to-image generation and editing, and so on. Recent research like GeoCLIP (Vivanco Cepeda et al., 2023) and SatCLIP (Klemmer et al., 2024) exemplify the growing recognition of this need, exploring CLIP-inspired methods and location embeddings specifically for geospatial understanding. These efforts, while

[*]Equal Contributors. Correspondence to: George Leifman gleifman@google.com.

promising, also underscore the ongoing need for more generalized and robust VLMs in the remote sensing domain, capable of handling the unique challenges of this data.

Recent advances in VLMs have shown substantial benefits across numerous natural images tasks. CLIP (Radford et al., 2021) was the first to utilize contrastive loss, aligning images with their corresponding textual descriptions in a shared embedding space. BLIP (Li et al., 2022) introduced a bootstrapping mechanism to refine noisy captions, further enhancing performance in vision-language tasks. SigLip (Zhai et al., 2023) proposed a sigmoid loss function for language-image pre-training, improving alignment and robustness, particularly in handling diverse and noisy datasets. MaMMUT (Kuo et al., 2023) presented a streamlined architecture with a vision encoder and text decoder, achieving state-of-the-art performance across diverse multi-modal tasks, including image-text retrieval, video question answering, and open-vocabulary detection. SigLiT (Zhai et al., 2022) (Sigmoid Loss with Locked-Image Tuning) uses a pretrained Vision Transformer (ViT) from ImageNet-22K and fine-tunes only the text encoder on image-caption pairs. It employs a simplified loss function for efficient training, achieving strong zero-shot performance.

Several studies have been exploring the application of VLMs on the remote-sensing domain. RemoteCLIP (Liu et al., 2023) adapts CLIP for remote sensing, leveraging image-text pairs derived from existing remote-sensing datasets. SkyScript (Wang et al., 2023) aligns OpenStreetMap features on remote sensing images to create a template-based RS image captioning dataset. Several works used LLM-modified RS-datasets to train RS-VLMs (e.g, LHRS bot by Muhtar et al. (2024), GeoChat by Kuckreja et al. (2023), Geo-RSCLIP by Zhang et al. (2024)). However, the advances in remote-sensing foundation models demonstrated by these studies remain constrained by the scale and semantic diversity of these datasets, leaving a large margin to improve generalized VLMs for remote sensing.

2 METHODOLOGY

This section introduces two methods for creating unique remote sensing vision-language datasets to address the absence of rich textual descriptions in remote sensing images and the scarcity of labeled data. These datasets are then used to train a compact yet robust 800M-parameter VLM for remote sensing, designed with the potential to run efficiently on resource-constrained devices. We leverage the pre-trained foundation VLM, MaMMUT (Kuo et al., 2023), which is a model pre-trained with a contrastive learning approach. To fine-tune it on our two novel datasets, namely RS-WebLI and RS-Landmarks, we use the shape-optimized 400M-parameter ViT as a vision encoder (Dosovitskiy et al., 2021), while the language model has additional 400M parameters.

2.1 REMOTE-SENSING DATASET CREATION

RS-Landmarks Dataset. The RS-Landmarks dataset is a novel remote-sensing dataset comprised of 18 million images with high-quality textual descriptions generated with Gemini 1.5 Pro (Gemini Team Google, 2023) as a teacher model. Initially, the satellite and aerial images are aligned with locations and footprints of places and landmarks extracted from Google Maps[1], found within the associated images. We feed this information, alongside with the image, to the Gemini model, and use a tailored, curated prompt to generate concise captions for each image. This process yields high-quality and informative captions, describing a diverse set of object categories.

RS-WebLI Dataset. The RS-WebLI dataset is comprised of aerial and satellite imagery taken from the WebLI dataset (Chen et al., 2023). We generated RS-WebLI by training aerial and overhead classifiers and thereafter using the classifiers to filter the WebLI dataset. First, we manually classified several hundred RS images in the WebLI dataset, using a simple caption heuristic and manual inspection of the dataset. Second, using the manually labeled sample, we trained an image classifier for remote sensing images and generated a dataset consisting of 40K images. Third, we harnessed the capabilities of crowd computing and initiated a labeling task wherein participants classified images as overhead aerial or satellite imagery, angled aerial imagery, or none of the above. The results were combined with random negatives, yielding a 60K labeled dataset. Lastly, using the large dataset, we trained new aerial and overhead classifiers and applied them across the entirety of WebLI, filtering the data of which we chose 3 million clean images, creating the RS-WebLI dataset.

[1]Google Maps data used with permission from Google

2.2 TRAINING

Following Zhai et al. (2023), Kuo et al. (2023) and Chen et al. (2022), we start with the MaMMUT model, pre-trained for 500K steps on the WebLI dataset, denoted as the *MT-WebLI* model. We then train two baseline models *MT-RSWebLI* and *MT-RSLandmarks* on the RS-WebLI and on RS-Landmarks datasets respectively, each for 20K steps. Finally, we train the combined *MT-RSWebLI-RSLandmarks* model in an optimized tri-step curriculum: (i) for 20K steps over the RS-WebLI dataset, (ii) for additional 20K steps over the RS-Landmarks dataset, and (iii) for another 5K steps on a mix of the two. All training steps (including the pre-training) were done with a batch size of 16K, using a Sharded Adafactor optimizer with Adam learning rate decay. The learning rates used are 1e-3, 1e-6, 5e-6 and 1e-7, for the WebLI pre-training, RS-WebLI training, RS-Landmarks, and mix phases correspondingly. In the mix training phase, we use a mix of 97% RS-Landmarks and 3% of RS-WebLI.

2.3 ZERO-SHOT EVALUATIONS

We evaluated the zero-shot retrieval performance of our proposed model, as presented in Table 1. We used a nearest neighbor approach on the output embeddings of the model to match each image to a class. Following standard practice, we report the average top-1, top-5, and top-10 recall scores. If a dataset provided multiple captions for a single image, we considered the retrieval a success if any of the correct captions were selected. The table shows that MT-RSWebli-RSLandmarks outperforms all other public models. Moreover, using both datasets is essential to obtain optimal results. We demonstrate the benefit of utilizing our remote sensing datasets to train a model which is

Table 1: Average of top-1/5/10 of zero-shot retrieval results for image to text (I2T) and text to image (T2I). Our result are compared to PIR-ITR Pan et al. (2024), SkyScript Wang et al. (2023), and Zhang et al. (2023) (Geo-RSClip). Our base model was trained on the WebLI dataset.

Model Name	**RSICD**		**UCM Cap.**		**NWPU**		**RSITMD**	
	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
PIR-ITR	24.43	25.77	-	-	-	-	38.64	39.85
SkyScript SkyCLIP-30	23.70	19.97	72.22	59.33	-	-	30.75	30.58
Geo-RSClip+RS5M	26.41	25.96	-	-	-	-	33.33	38.02
MT-WebLI	23.88	24.17	69.21	66.50	20.33	23.18	28.83	32.70
MT-RSWebLI	26.68	25.09	72.38	65.96	24.62	22.60	32.15	35.57
MT-RSLandmarks	33.12	32.98	72.38	70.91	40.10	32.15	41.96	42.30
MT-RSWebli-RSLandmarks	**33.33**	**33.59**	**74.76**	**71.79**	**41.44**	**32.28**	**42.63**	**42.58**

capable of generalizing on the remote-sensing domain. This result is further substantiated by showing that the enhanced accuracy over the non-RS baseline is kept even when testing on categories which *were not* explicitly presented during training. To illustrate this, we created *RS-Landmarks-89*, an image classification benchmark dataset, comprising 89 manually chosen landmark categories from the RS-Landmarks dataset, where each image is centred on one of the landmark types. Complementarily, the *RS-Landmarks-89-holdout* image-text dataset was created by removing images from the RS-Landmarks dataset that contained the previously mentioned categories. To evaluate the zero-shot classification abilities on unseen categories, we fine-tune the MT-WebLI model with the *RS-Landmarks-89-holdout* dataset. We evaluate the resulting model on the RS-Landmarks-89 classification dataset using simple nearest-neighbor approach.

Unsurprisingly, Table 2 shows that a model tuned on the RS-Landmarks dataset outperforms the baseline, general purpose MT-WebLI model on the RS-LandMarks-89 dataset. The ability to generalize on the remote sensing domain is evaluated by the model tuned on the RSLandMarks-89-holdout dataset (2nd row). The performance of this model, which was not trained on any of the 89 categories, is significantly higher than the MT-WebLI baseline, and is close to that of the full MT-RSLandmarks.

3 SELF-SUPERVISED ZERO-SHOT LOCALIZATION VIA ITERATIVE REFINEMENT

In this section, we present the ongoing effort to distill image-level knowledge gained in the VLM contrastive training procedure to enhance the model's localization ability. Enhanced localization is

Table 2: **Generalization to unseen classes**. Nearest neighbour classification accuracy on the hold-out dataset. The 89 classes in the *RS-Landmarks-89* are excluded from the *RS-Landmarks-89-holdout* training-set.

Model Name	**RS-Landmarks-89**
MT-WebLI	13.78
MT-WebLI tuned on RSLandmarks-89-holdout	21.38
MT-WebLI tuned on RSLandmarks	25.42

essential for performance in downstream vision tasks, notably segmentation. Applying the methodology described in the previous section, with an additional attention pooling head similarly to SigLIP[2], reveals an important insight, that despite being explicitly trained for image-level on broad captions, the model appears to classify individual patches at the attention-pooling layer. This phenomenon is illustrated in Figure 1, where the similarity is captured at the patch-level.

We build upon this phenomenon to present a pseudo labeling algorithm that yields segmentation masks for a query of interest. Our algorithm begins with the model receiving the text query along with a stack of images. The query is paired with the images during inference, where for each image we create a similarity-based attention map (Figure 1). This map segments the associated patches that match the text query within every image. A pre-defined adaptive similarity threshold is then applied to identify patches whose similarity to the query exceeds the threshold. We propose to use these pseudo-labeled samples to fine-tune the model. This self-supervised process is repeated iteratively, with the adaptive similarity threshold gradually increasing as the model achieves higher accuracy in producing robust segmentation masks for arbitrary text queries.

To mitigate the noise in the segmentation maps, we introduce a novel approach called *Smooth-Attention-Operation*. This method differs from standard attention mechanisms by applying the attention pooling layer, not to the entire input key, but rather to smaller, sliding window blocks of the key. The sliding window traverses the image, pooling attention values within its boundaries. Smaller windows capture highly localized attention, while larger windows incorporate more contextual information from neighboring regions. This approach balances local detail with broader context, offering a trade-off between per-patch similarity and full-image attention.

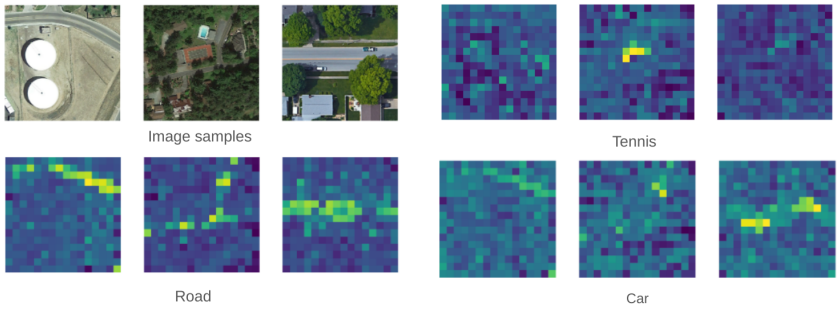

Figure 1: Attention similarity maps for text queries (*'Tennis', 'Road', 'Car'*) in different images, selected from the DIOR dataset (Zhan et al., 2023).

4 Discussion

We tackled the challenges of applying Vision-Language Models (VLMs) in remote sensing by introducing two new datasets: RS-WebLI and RS-Landmarks. RS-WebLI is a curated large-scale collection of aerial and satellite imagery, while RS-Landmarks includes high-quality captions generated using Google Maps data by the Gemini teacher model. These datasets enable VLM training optimized for remote sensing, improving generalization. Our VLM foundation model demonstrated state-of-the-art cross-modal retrieval performance on public benchmarks. Additionally, we developed a self-supervised zero-shot retrieval scheme for pseudo-labeling, allowing segment-level classification in the attention-pooling layer.

[2]Constructed by the shape-optimized version, SoViT-400m architecture, as in Alabdulmohsin et al. (2024)

REFERENCES

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design, 2024. URL https://arxiv.org/abs/2305.13035.

Jules Bourcier, Gohar Dashyan, Karteek Alahari, and Jocelyn Chanussot. Learning representations of satellite images from metadata supervision. In *European Conference on Computer Vision*, pp. 54–71. Springer, 2024.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. URL https://arxiv.org/abs/2209.06794.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Gemini Team Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery, 2024. URL https://arxiv.org/abs/2311.17179.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing, 2023. URL https://arxiv.org/abs/2311.15826.

Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Xiyang Luo, Ben Caine, Wei Li, Abhijit Ogale, Luowei Zhou, Andrew Dai, Zhifeng Chen, Claire Cui, and Anelia Angelova. Mammut: A simple architecture for joint learning for multimodal tasks, 2023. URL https://arxiv.org/abs/2303.16839.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023.

Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, pp. 23498–23515. PMLR, 2023.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model, 2024. URL https://arxiv.org/abs/2402.02544.

Jiancheng Pan, Muyuan Ma, Qing Ma, Cong Bai, and Shengyong Chen. Pir: Remote sensing image-text retrieval with prior instruction representation learning. *arXiv preprint arXiv:2405.10160*, 2024.

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*, pp. 8748–8763. PMLR, 2021.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8690–8701. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b57aaddf85ab01a2445a79c9edc1f4b-Paper-Conference.pdf.

Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing, 2023. URL https://arxiv.org/abs/2312.12856.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022. URL https://arxiv.org/abs/2111.07991.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023.

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*, 62:1–23, 2024. ISSN 1558-0644. doi: 10.1109/tgrs.2024.3449154. URL http://dx.doi.org/10.1109/TGRS.2024.3449154.

A Appendix

A.1 Zero-shot classification evaluation

We evaluate the model's zero-shot classification performance on remote sensing images without prior training on those specific classes using several remote sensing image classification datasets, presented in Table A.1. For each dataset, we created a set of descriptive sentences in the format "An aerial image of <class name>." Then, using a nearest neighbor algorithm, we determined the best matching class for each image based on these sentences.

Table 3: Comparison of the Top-1 accuracy zero-shot classification performance, results taken from papers.

Model Name	**FMOW**	**SkyScript**	**RESISC45**	**UCM Class.**	**AID**
SkyScript	28.04	(70.89)	70.94	-	
RS-CLIP	-	68.84	71.35	74.28 / 78.00	70.51
GeoRSCLIP-VitL	-	-	71.89	-	76.33
GeoRSCLIP-VitH	-	-	73.83	-	73.72
RemoteCLIP	-	-	79.84	-	91.3
GeoChat (7B)	-	-	-	84.43	72.03
LHRS-Bot (7B)	-	-	-	-	91.26
MT-WebLI	37.58	58.66	66.93	76.52	71.46
MT-RSWebLI	42.73	65.16	70.91	83.52	75.78
MT-RSLandmarks	42.93	66.31	68.55	77.67	73.15
MT-RSWebli+RSLandmarks	47.24	69.46	72.31	80.29	71.96