LARGE LANGUAGE MODELS FOR CAPTIONING AND RETRIEVING REMOTE SENSING IMAGES

João Daniel Silva[♠*] **& João Magalhães**[♢] **& Devis Tuia**[♣] **& Bruno Martins**[♠]
♠INESC-ID, Instituto Superior Técnico, University of Lisbon
♢Faculty of Science and Technology, Universidade NOVA de Lisboa
♣ECEO, Ecole Polytechnique Fédérale de Lausanne
*joao.daniel.silva@tecnico.ulisboa.pt

ABSTRACT

Remote sensing tasks, such as image captioning and cross-modal retrieval, enable non-expert users to extract relevant Earth observation by integrating visual and linguistic information. In this work, we propose RS-CapRet, a Vision and Language model for remote sensing data, in particular image captioning and text-image retrieval tasks. We integrate a large language model together with an image encoder adapted to remote sensing imagery through contrastive language-image pre-training. To bridge together the image encoder and the language decoder, we propose training lightweight linear layers with examples from combining different remote sensing image captioning datasets, keeping the other parameters frozen. RS-CapRet generates descriptions for remote sensing images and retrieves images from textual descriptions, achieving a competitive performance with existing methods.

1 INTRODUCTION

There has been a growing interest in the application of Vision and Language (V&L) models in the remote sensing domain (Wen et al., 2023; Mai et al., 2023), for tasks such as image retrieval (Liu et al., 2024; Rahhal et al., 2022; Mi et al., 2022; Yuan et al., 2022), image captioning (Cheng et al., 2022; Wei et al., 2023; Ramos & Martins, 2022; Shi & Zou, 2017), or visual question answering (Lobry et al., 2020; Silva et al., 2022; Bazi et al., 2022; Zhang et al., 2023). Methods developed for these tasks can enable a wider population of individuals, with different degrees of expertise to interact with Earth observation data (Tuia et al., 2021; Martins & Silva, 2022), supporting the extraction of rich insights from remote sensing images.

Previous methods have adopted deep learning methods for V&L tasks. Despite several recent efforts (Wen et al., 2023; Liu et al., 2024; Hu et al., 2023; Zhan et al., 2024; Kuckreja et al., 2023), the relatively small size of the available datasets of image-text pairs has restrained the application and development of V&L models in the remote sensing domain, contrasting with the trend in general domain images where models are getting increasingly complex and trained with large-scale datasets (Wang et al., 2021; Li et al., 2022; Alayrac et al., 2022). In this domain, many image captioning and visual question asnwering approaches use an encoder-decoder architecture with CNNs as image encoders, emphasizing intricate attention mechanisms for remote sensing imagery (Cheng et al., 2022; Huang et al., 2021; Li et al., 2020; Yuan et al., 2020; Zhao et al., 2021).

Recent advances in natural language processing have been driven by Large Language Models (LLM), known for their zero-shot capabilities, and for exhibiting logical reasoning and commonsense knowledge (Petroni et al., 2019; Brown et al., 2020; Wei et al., 2022a;b; Kojima et al., 2022; Huang & Chang, 2022). These strengths have motivated efforts into the integration of visual information with LLMs, so as to address V&L tasks (Guo et al., 2022; Yang et al., 2022; Alayrac et al., 2022; Tsimpoukelli et al., 2021). However, LLMs face challenges, including high memory demands for inference and the high cost of fine-tuning for downstream tasks.

In this work, we propose RS-CapRet, a model that combines the strengths of a Large Language Model, with an image encoder adapted to the remote sensing domain, using a lightweight training procedure. RS-CapRet generates descriptions for remote sensing images, surpassing methods with

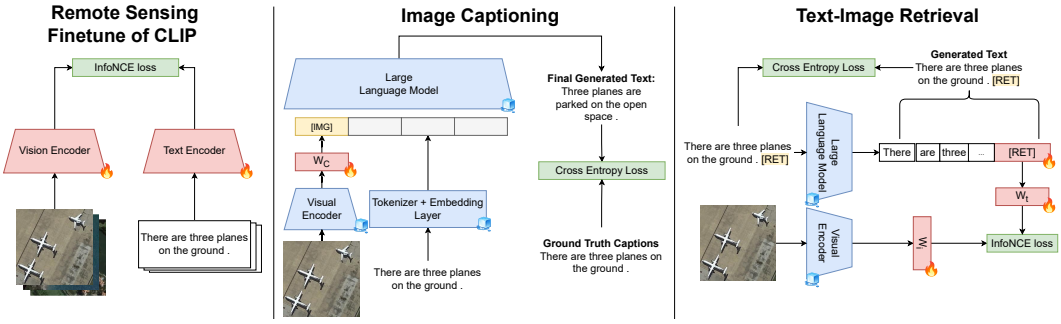

Figure 1: Overview of the method used for training RS-CapRet. Left: CLIP is finetuned to the remote sensing domain with image-text pairs from image captioning datasets. Middle: Image captioning task where image embeddings are obtained via a frozen image encoder and projected with a trainable linear layer to the input embedding space of the frozen large language model, which are then concatenated with the input text. Right: Trainable linear layers are adjusted with contrastive learning between image representations and a special `[RET]` token to address text-image retrieval.

more complex, domain-specific architectural choices. Qualitative results illustrate the ability to describe remote sensing imagery, integrate image and text inputs in dialogue, and use reasoning capabilities.

2 METHOD

RS-CapRet combines the strengths of a Large Language Model (LLM) with an image encoder adapted to remote sensing. Instead of fine-tuning the LLM and the vision encoder, their parameters are kept frozen and only a linear layer to project visual embeddings into the input embedding space of the LLM is trained, enabling the LLM to process visual information as embedding vectors.

In addition to generating image descriptions, RS-CapRet supports image retrieval from textual queries by using a special retrieval token `[RET]`, with its embedding being projected into a common embedding space with the images. Contrastive learning aligns the `[RET]` embedding with corresponding image embeddings, enabling retrieval based on text-image similarity after training.

2.1 THE RS-CAPRET ARCHITECTURE

RS-CapRet consists of the following components: a) a Large Language Model (LLM), b) a vision encoder finetuned to the remote sensing domain, used to obtain image embeddings, c) a linear layer to project the image embeddings to the input space of the LLM, and d) two linear layers to project, respectively, the image embedding, and the text embedding given by the `[RET]` token to a common shared space. A more detailed description of each component is presented next:

a) Pre-trained Language Model. The main and larger component of our model is a text decoder based on the Transformer architecture (Vaswani et al., 2017), which was pre-trained on autoregressive text generation. We used LLamaV2-7B model (Touvron et al., 2023b).

b) Visual Encoder. Given an image $\mathbf{x}$, a vision Transformer encoder based on the CLIP architecture (Radford et al., 2021) is used to obtain the image representation $f_\phi(\mathbf{x}) = \mathbf{v} \in \mathbb{R}^m$, based on the representation of the `[CLS]` token.

c) Projections Between Modalities. The vision and text modalities are bridged in two separate directions for each task with different projection layers. For image captioning, a linear layer projection $\mathbf{W}_c \in \mathbb{R}^{m \times D}$, with D as the LLM hidden dimension, is used to project the vision embeddings into the input embedding space of the LLM, resulting in a visual prefix. For text-image retrieval, the `[RET]` token is appended at the end of each caption, with its embedding at the output of the LLM consisting of an overall representation of the text. A linear layer $\mathbf{W}_t \in \mathbb{R}^{D \times q}$ is used to project this representation while another $\mathbf{W}_i \in \mathbb{R}^{m \times q}$ also projects the visual embedding to a common shared space of dimensionality $q < D$, so that contrastive learning can be applied (with the InfoNCE loss function (van den Oord et al., 2018)).

Table 1: Image Captioning results on the NWPU-Captions, RSICD, UCM, and Sydney datasets.

Evaluation Dataset	Method	Visual Encoder	Text Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
NWPU	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.745	0.624	0.541	0.478	0.337	0.601	1.164	0.285
	RS-CapRet	CLIP-Cap-4	LLamaV2	**0.871**	0.786	0.713	0.650	**0.439**	0.775	1.919	**0.320**
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	**0.871**	**0.787**	**0.717**	**0.656**	0.436	**0.776**	**1.929**	0.311
RSICD	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.757	0.634	0.539	0.461	0.351	0.646	2.356	0.444
	RSGPT (Hu et al., 2023)	EVA-G	Vicuna	0.703	0.542	0.440	0.368	0.301	0.533	1.029	NA
	SkyEyeGPT (Zhan et al., 2024)	EVA-G	LLamaV2-Chat	**0.867**	**0.767**	**0.673**	**0.600**	0.354	0.626	0.837	NA
	RS-CapRet	CLIP-Cap-4	LLamaV2	0.741	0.622	0.529	0.455	**0.376**	**0.649**	**2.605**	**0.484**
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	0.720	0.599	0.506	0.433	0.370	0.633	2.502	0.474
UCM	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.826	0.770	0.717	0.668	0.435	0.772	3.240	0.473
	RSGPT (Hu et al., 2023)	EVA-G	Vicuna	0.861	0.791	0.723	0.657	0.422	0.783	3.332	NA
	SkyEyeGPT (Zhan et al., 2024)	EVA-G	LLamaV2-Chat	**0.907**	**0.857**	**0.816**	**0.784**	0.462	0.795	2.368	NA
	RS-CapRet	CLIP-Cap-4	LLamaV2	0.833	0.760	0.699	0.645	0.447	0.786	3.429	**0.525**
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	0.843	0.779	0.722	0.670	**0.472**	**0.817**	**3.548**	**0.525**
Sydney	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.831	0.742	0.659	0.580	0.390	0.711	2.324	0.409
	RSGPT (Hu et al., 2023)	EVA-G	Vicuna	0.823	0.753	0.686	0.622	0.414	0.748	**2.731**	NA
	SkyEyeGPT (Zhan et al., 2024)	EVA-G	LLamaV2-Chat	**0.919**	**0.856**	**0.809**	**0.774**	**0.466**	**0.777**	1.811	NA
	RS-CapRet	CLIP-Cap-4	LLamaV2	0.782	0.688	0.611	0.545	0.383	0.704	2.390	0.423
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	0.787	0.700	0.628	0.564	0.388	0.707	2.392	**0.434**

2.2 TRAINING PROCEDURE

RS-CapRet is jointly trained with two tasks, i.e. image captioning and image-text retrieval. A graphical depiction of the training process can be seen in the middle and right sections of Figure 1.

Image captioning. The image captioning task is considered as in previous work (Koh et al., 2023; Eichenberg et al., 2022; Tsimpoukelli et al., 2021): conditional generation of a caption $\mathbf{y}$ given an image $\mathbf{x}$. The cross-entropy loss between the generated tokens and the ground-truth caption tokens is used to train the linear projection $\mathbf{W}_c$. To increase the robustness of the model to handle interleaved sequences of images and texts, two captions are concatenated together during training for the image captioning objective.

Image-Text Retrieval. Contrastive learning is also incorporated into the training procedure of RS-CapRet, to train projection layers $\mathbf{W}_i$ and $\mathbf{W}_t$. Specifically, the InfoNCE loss is optimized considering two directions: text-to-image $\mathcal{L}_{\text{t2i}}$ and image-to-text $\mathcal{L}_{\text{i2t}}$, as outlined in Equations 1 and 2, in Appendix B.

The final training loss can be characterized as a weighted sum of both the image captioning and the contrastive learning tasks: $\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r(\mathcal{L}_{\text{t2i}} + \mathcal{L}_{\text{i2t}})$.

3 EXPERIMENTAL SETUP

Datasets. We leverage publicly available datasets used for image captioning and cross-modal retrieval in the remote sensing domain, namely RSICD, UCM-Captions, Sydney-Captions and NWPU-Captions. A detailed description of these datasets is given in Appendix C.1.

Backbone Models. The vision encoder is based on a CLIP vision model (Radford et al., 2021). Specifically, we use a CLIP vision encoder of size large, finetuned with an aggregation of the aforementioned image captioning datasets, with this model being referred in the text as CLIP-Cap-4. See Appendix B.1 for more details regarding the choice of this model. As for the decoder model, we used the LLamaV2-7B language model (Touvron et al., 2023a).

Metrics. Evaluation of image captioning is based on common metrics following previous, namely BLEU, METEOR, ROUGE_L, CIDEr, and SPICE. For image retrieval, the recall at cutoff position top-1 (R@1), top-5 (R@5), and top-10 (R@10), is calculated, also following previous work (Mi et al., 2022; Liu et al., 2024). R@K means the ratio of queries that successfully retrieve the ground truth as one of the first K results. We focus on the text-image retrieval direction, since our model can obtain text by captioning.

Implementation Details. As has been mentioned, the training is lightweight. With a batch size of 64 and leveraging bfloat16 mixed-precision format, model training was conducted in a single NVIDIA A100-40GB.

4 EXPERIMENTAL RESULTS

Image Captioning. The results for the image captioning task are compiled in Table 1. For comparison against previous work, we include the reported results of MLCA-NET (Cheng et al., 2022), i.e. an encoder-decoder method, and also RSGPT (Hu et al., 2023) and SkyEyeGPT (Zhan et al., 2024), which are V&L models for the remote sensing domain.

Table 2: Results for retrieval experiments in the RSICD and UCM datasets. We focus on the text-image retrieval direction, since our model can obtain text by captioning. Models marked with † were evaluated in our setup, otherwise the results are collected from the respective reports.

Dataset	Method	Visual Backbone	Finetune Data	Text-Image Retrieval			
				R@1	R@5	R@10	mR_T2I
RSICD	GaLR (Yuan et al., 2022)	ResNet18	RSICD	4.69	19.48	32.13	18.77
	KCR (Mi et al., 2022)	ResNet101	RSICD	5.40	22.44	37.36	21.73
	CLIP (Radford et al., 2021)†	ViT-B	Zero-shot	5.80	16.85	28.23	16.96
	CLIP (Radford et al., 2021)†	ViT-L	Zero-shot	5.03	19.03	30.25	18.10
	Rahhal et al. (Rahhal et al., 2022)	ViT-B	RSICD	9.14	28.96	44.59	27.56
	CLIP-RSICD (Pal et al., 2021)†	ViT-B	RSICD	11.16	33.25	48.91	31.11
	CLIP-Cap-4†	ViT-L	Cap-4	13.83	39.07	56.05	36.32
	RemoteCLIP (Liu et al., 2024)	ViT-L	RemoteCLIP dataset	**14.73**	**39.93**	**56.58**	**37.08**
	RS-CapRet†	ViT-L	Cap-4	9.83	30.17	47.43	29.14
	RS-CapRet$_{finetuned}$†	ViT-L	Cap-4 + RSICD	10.25	31.62	48.53	30.13
UCM	KCR (Mi et al., 2022)	ResNet101	RSICD	17.43	57.52	80.38	51.78
	CLIP (Radford et al., 2021)†	ViT-B	Zero-shot	8.67	36.48	60.57	35.24
	CLIP (Radford et al., 2021)†	ViT-L	Zero-shot	10.76	46.00	73.33	43.37
	CLIP-RSICD (Pal et al., 2021)†	ViT-B	RSICD	13.81	57.05	91.24	54.03
	CLIP-Cap-4†	ViT-L	Cap-4	16.29	60.57	94.76	57.21
	RemoteCLIP (Liu et al., 2024)	ViT-L	RemoteCLIP dataset	**17.71**	62.19	**93.90**	57.93
	Rahhal et al. (Rahhal et al., 2022)	ViT-B	UCM	19.33	**64.00**	91.42	**58.25**
	RS-CapRet†	ViT-L	Cap-4	15.52	57.24	88.76	53.84
	RS-CapRet$_{finetuned}$†	ViT-L	Cap-4 + UCM	16.10	56.29	90.76	54.38

On the NWPU-Captions dataset, RS-CapRet significantly improves over the previous SOTA model named MLCA-NET (Cheng et al., 2022) (e.g. +0.126 for BLEU-1, +0.755 for CIDEr). Considering the results in the RSICD dataset, RS-CapRet surpasses RSGPT in all metrics (in particular CIDEr by a high amount of +1.576), and for the other models, RS-CapRet has higher scores on all metrics but BLEU. For the UCM-Captions dataset, RS-CapRet has lower results for the BLEU metric, but achieves higher METEOR, CIDEr and SPICE scores, and also a similar ROUGE_L score when compared to RSGPT. Compared to SkyEyeGPT, RS-CapRet surpasses only in CIDEr. Regarding the results in the Sydney-Captions, dataset RS-CapRet achieves higher CIDEr than MLCA-NET and SkyEyeGPT. We also observe that the results can increase with further finetuning, in particular for the smaller datasets UCM and Sydney-Captions (marked with underscript *finetuned* in Tables 1 and 2).

Cross-Modal Retrieval. Retrieval results can be seen in Table 2. To support a comparison against previous state-of-the-art, we include baselines such as GaLR (Yuan et al., 2022), and KCR (Mi et al., 2022), and also CLIP-based models such as Rahhal et al. (2022) and RemoteCLIP (Liu et al., 2024). We collect the results reported for the baselines, while we evaluate the CLIP models in our retrieval setup. RS-CapRet can achieve higher results when compared to GaLR, KCR, (Rahhal et al., 2022), and also zero-shot CLIP models of size Base and Large. However, it cannot surpass larger variants fine-tuned to the remote sensing domain. For the UCM dataset, a similar pattern can be observed.

We note that CLIP is a strong baseline for cross-modal retrieval that improves with domain-specific fine-tuning, as shown in our experiments and also prior work (Liu et al., 2024). Two full Transformer encoders for images and text give an advantage for text-image retrieval over RS-CapRet, which trains only projection layers over embeddings. Nevertheless, RS-CapRet achieves competitive results, with slight performance gains observed after fine-tuning.

We further discuss the results in Appendix D, including model finetuning on each dataset, and an ablation of the vision encoder.

5 Conclusions

We described a new and simple vision and language model for the remote sensing domain named RS-CapRet, which can address the tasks of image captioning and text-image retrieval with a lightweight training procedure. RS-CapRet can also obtain images for specific user requests, e.g. given particular objects or related themes, and it can handle short dialogues about remote sensing images. Appendix E provides more details about these applications. The architecture is highly modular and can be updated with new developments, either related to LLMs or to vision encoders for the remote sensing domain.

ACKNOWLEDGMENTS

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by the Fundação para a Ciência e Tecnologia (FCT), specifically through the project with reference UIDB/50021/2020 (DOI: 10.54499/UIDB/50021/2020), and the project with reference UIDP/04516/2020 (DOI: 10.54499/UIDB/04516/2020).

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Mohamed Lamine Mekhalfi, Mansour Abdulaziz Al Zuair, and Farid Melgani. Bi-modal transformer-based approach for visual question answering in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv:2304.05215*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709*, 2020.

Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 2416–2428, 2022.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot VQA with frozen large language models. *arXiv:2212.10846*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9:1735–1780, 1997.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. RSGPT: A remote sensing vision language model and benchmark. *arXiv:2307.15266*, 2023.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv:2212.10403*, 2022.

Wei Huang, Qi Wang, and Xuelong Li. Denoising-based multiscale feature fusion for remote sensing image captioning. *IEEE Geoscience and Remote Sensing Letters*, 18(3):436–440, 2021.

Johannes Jakubik, Michal Muszynski, Michael Vössing, Niklas Kühl, and Thomas Brunschwiler. Toward foundation models for Earth Monitoring: Generalizable deep learning models for natural hazard segmentation. *arXiv:2301.09318*, 2023.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the International Conference on Machine Learning*, pp. 17283–17300, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. *arXiv:2311.15826*, 2023.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. GEO-bench: Toward foundation models for earth monitoring. In *Advabces in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 36, 2023.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 19730–19742, 2023.

Xuelong Li, Xueting Zhang, Wei Huang, and Qi Wang. Truncation cross entropy loss for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5246–5257, 2021.

Yangyang Li, Shuangkang Fang, Licheng Jiao, Ruijiao Liu, and Ronghua Shang. A multi-level attention model for remote sensing image captions. *Remote Sensing*, 12(6), 2020.

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*, 62, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023b.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 58, 2020. ISSN 0196-2892.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021.

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, 2018.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv:2304.06798*, 2023.

Bruno Martins and João Daniel Silva. Towards natural language interfaces for interacting with remote sensing data. *UC Santa Barbara: Center for Spatial Studies*, 2022.

Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, and Mu Li. GFM: Building geospatial foundation models via continual pretraining. *arXiv:2302.04476*, 2023.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv:2209.15162*, 2022.

Li Mi, Siran Li, Christel Chappuis, and Devis Tuia. Knowledge-aware cross-modal text-image retrieval for remote sensing images. In *Proceedings of the Workshop on Complex Data Challenges in Earth Observation*, volume 3207, pp. 14–20, 2022.

Sujit Pal, Artashes Arutiunian, Goutham Venkatesh, Ritobrata Ghosh, Dev Vidhani, and Mayank Bhaskar. Fine tuning CLIP with Remote Sensing (Satellite) images and captions. *HuggingFace Blog*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv:1909.01066*, 2019.

Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *Proceedings of the International Conference on Computer, Information and Telecommunication Systems*, pp. 1–5, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *Open AI Blog*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Mohamad M. Al Rahhal, Yakoub Bazi, Norah A. Alsharif, Laila Bashmal, Naif Alajlan, and Farid Melgani. Multilanguage transformer for improved text to remote sensing image retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9115–9126, 2022.

Rita Ramos and Bruno Martins. Using Neural Encoder-Decoder Models With Continuous Outputs for Remote Sensing Image Captioning. *IEEE Access*, 10, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

Zhenwei Shi and Zhengxia Zou. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Transactions on Geoscience and Remote Sensing*, 55, 2017.

João Daniel Silva, João Magalhães, Devis Tuia, and Bruno Martins. Remote sensing visual question answering with a self-attention multi-modal encoder. In *Proceedings of the ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pp. 40–49, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

Gencer Sumbul, Sonali Nayak, and Begüm Demir. Sd-rsic: Summarization-driven deep remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6922–6934, 2021.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*, 2023a.

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

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Devis Tuia, Ribana Roscher, Jan Dirk Wegner, Nathan Jacobs, Xiaoxiang Zhu, and Gustau Camps-Valls. Toward a collective agenda on AI for earth science data analysis. *IEEE Geoscience and Remote Sensing Magazine*, 9(2), 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer towards remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *arXiv:2108.10904*, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

Tingting Wei, Weilin Yuan, Junren Luo, Wanpeng Zhang, and Lina Lu. VLCA: Vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning. *Journal of Systems Engineering and Electronics*, 34(1):9–18, 2023.

Congcong Wen, Yuan Hu, Xiang Li, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-language models in remote sensing: Current progress and future trends. *arXiv:2305.05726*, 2023.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3081–3089, 2022.

Zhenghang Yuan, Xuelong Li, and Qi Wang. Exploring multi-level attention and semantic relationship for remote sensing image captioning. *IEEE Access*, 8:2608–2620, 2020.

Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

Valerie Zermatten, Javiera Castillo Navarro, Lloyd Hughes, Tobias Kellenberger, and Devis Tuia. Text as a richer source of supervision in semantic segmentation tasks. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pp. 2219–2222, 2023.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv:2401.09712*, 2024.

Zixiao Zhang, Licheng Jiao, Lingling Li, Xu Liu, Puhua Chen, Fang Liu, Yuxuan Li, and Zhicheng Guo. A spatial hierarchical reasoning network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Rui Zhao, Zhenwei Shi, and Zhengxia Zou. High-resolution remote sensing image captioning based on structured attention. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Usman Zia, M Mohsin Riaz, and Abdul Ghafoor. Transforming remote sensing images to textual descriptions. *International Journal of Applied Earth Observation and Geoinformation*, 108:102741, 2022.

A Related Work

The growing availability of multimodal data in the remote sensing domain has led to a rise in research addressing tasks such as image captioning or cross-modal retrieval. Recent work has introduced foundation models for this particular domain, including vision encoders as well as multimodal vision and language models.

A.1 Image Captioning

Most previous image captioning methods in the remote sensing domain have been based on encoder-decoder architectures (Qu et al., 2016; Lu et al., 2018; Zia et al., 2022; Cheng et al., 2022; Zhao et al., 2021; Yuan et al., 2020; Li et al., 2020; Sumbul et al., 2021; Li et al., 2021; Huang et al., 2021), leveraging CNNs as image encoders and LSTM to generate the caption word-by-word according to weights obtained by an attention component. In this line of research, many studies have proposed specialized mechanisms in the attention component between the encoder and the decoder, to take into account specific characteristics of remote sensing images, such as dealing with visual features at different scales (Huang et al., 2021; Yuan et al., 2020; Cheng et al., 2022). MLCA-Net proposed by Cheng et al. (2022) is an example of one such model that has achieved high performance on remote sensing datasets, by using a VGG backbone (Simonyan & Zisserman, 2014) to extract features at different resolutions, that are combined in multilevel and contextual attention mechanisms, and which are then passed to a LSTM (Hochreiter & Schmidhuber, 1997) to generate a caption. The authors have also created a new dataset for image captioning named NWPU-Captions, which has a higher quantity of data together with more diversity of descriptions and image contents. Zia et al. (2022) proposed an encoder-decoder architecture based on a Transformer (Vaswani et al., 2017), with image features obtained with a CNN developed to get features at multiple stages. The authors also include a topic modeling stage of the captions, as input to the decoder.

Some work has proposed methods incorporating recent V&L methods developed in the general domain for remote sensing. For instance, VLCA (Wei et al., 2023) leverages a CLIP model to obtain image features, and trains a cross-modal network to produce a representation to be used in a cross-attention layer of a GPT-2 decoder (Radford et al., 2019) to generate descriptions of the image.

A.2 Cross-Modal Retrieval

Remote sensing cross-modal retrieval is a task with increasing interest that can be used to evaluate representations of V&L models. Most previous work has obtained image features with CNNs and text features with LSTM or Transformer encoders, with different attention mechanisms proposed. In particular, GaLR (Yuan et al., 2022) introduced a method that leverages both global features from a CNN and local features obtained with a graph convolution network. The authors also apply a post-processing stage with a multivariate reranking algorithm to improve the accuracy without further training. KCR (Mi et al., 2022) proposes the usage of a knowledge graph to incorporate in-domain information about the concepts mentioned in the captions, enriching the textual embeddings extracted with a SentenceBERT model (Reimers & Gurevych, 2019). An attention mechanism is leveraged to combine features extracted at different stages from a CNN , and a triplet loss is used to optimize the model end-to-end.

CLIP (Radford et al., 2021) is a V&L model that has been trained with a contrastive loss such that images and their corresponding captions are close in the embedding space. Due to the high-quality representations of CLIP, it has motivated studies also in the remote sensing domain for cross-modal retrieval. Pal et al. (2021) finetuned CLIP with the RSICD dataset (Lu et al., 2018), studying the impact of different augmentations both for the images and the text, and showed that the resulting model has high-quality representations, particularly for image classification. Rahhal et al. (2022) also fine-tuned CLIP in both single and multi-language contexts, obtaining good results for cross-modal retrieval. RemoteCLIP (Liu et al., 2024) has developed a pipeline to process available datasets of object detection and semantic segmentation, to increase the number of image and text pairs. With this higher amount of data, the authors could train a CLIP model with a ViT-L backbone (Dosovitskiy et al., 2020) and improve the results in cross-modal retrieval, compared to the ViT-B backbone used in previous approaches. TACOSS, proposed by Zermatten et al. (2023), learns a fine-grained

alignment between visual and textual features with a contrastive learning objective, being able to perform semantic segmentation at the pixel level with this method.

A.3 Foundation Models for Remote Sensing

There has been a growing desire to develop foundational models for the remote sensing community (Mai et al., 2023; Wen et al., 2023; Jakubik et al., 2023; Martins & Silva, 2022). A common line of research has mainly focused on furthering the capabilities of vision encoders for the remote sensing domain, by pre-training in a self-supervised manner with different objectives, leveraging the available high quantity of unlabeled remote sensing images. These models are then used as backbones for other methods, e.g. for object detection and semantic segmentation. Wang et al. (2022) pretrained a Vision Transformer with 100M parameters using a Masked Auto Encoder (MAE) objective (He et al., 2022) in the MillionAID dataset (Long et al., 2021). The same authors also propose a new rotated varied-size window attention (RVSA) module with different orientation angles for computing attention. RingMO (Sun et al., 2022) is a ViT model pre-trained with a Masked Image Modeling (MIM) objective, which the authors argue is best to address local features and tiny objects. They propose masking pixels instead of the conventional way of masking captions, to take into account the small size of the objects in RS images. Cha et al. (2023) follow the same pre-training strategy and try to scale to models with over a billion parameters. (Mendieta et al., 2023) combined teacher-student distillation with masked image modeling. The development of better vision encoders for the remote sensing domain has also motivated the development of unified benchmarks for different Earth Observation tasks, such as GEO-Bench (Lacoste et al., 2023).

Regarding vision-and-language models for the remote sensing domain, RSGPT (Hu et al., 2023) is one example that effectively adapts the InstructBLIP (Dai et al., 2023) model with a new dataset named RSICap of high-quality image captioning data. This dataset can be considered as featuring "dense" captioning, as it covers multiple aspects of the image such as theme, image attributes, object attributes (shape, color, quantity, size), and description of the scene. Kuckreja et al. proposed GeoChat (Kuckreja et al., 2023), following the LLaVA1.5 architecture (Liu et al., 2023a) that can tackle different vision-and-language remote sensing tasks in a unified way, accepting both image-level or region-specific queries. This model can also ground objects in the images by referring to their spatial coordinates. Moreover, the authors proposed a multimodal instruction-following dataset built from current RS datasets. Another recently proposed model is SkyEyeGPT (Zhan et al., 2024), which also addresses multiple vision-and-language tasks, including visual question answering, image captioning, and vision grounding.

In this work, we contribute with a vision-and-language foundation model capable of image captioning and also text-image retrieval, addressing tasks in the remote sensing domain with a lightweight training procedure based on aligning the outputs of a visual encoder with the input space of a language model.

B Finetuning CLIP to the Remote Sensing domain

CLIP (Radford et al., 2021) is widely used as a vision encoder for different vision-and-language models leveraging large language models, due to the high-quality embeddings that it obtains (Li et al., 2023; Liu et al., 2023b; Koh et al., 2023). Given that CLIP was mainly trained with ground-level images, and inspired by recent work such as CLIP-RSICD (Pal et al., 2021) and RemoteCLIP (Liu et al., 2024) that improved results in downstream tasks for the remote sensing domain, we also have finetuned CLIP with remote sensing image captioning datasets consisting of image and caption pairs. Specifically, for a dataset of M image-caption pairs $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{M}$, a vision encoder obtains a representation for the image $f_\phi(\mathbf{x}_i) = \mathbf{v}_i \in \mathbf{R}^m$, and a text decoder obtains another for the caption $t_\theta(\mathbf{y}_i) = \mathbf{u}_i \in \mathbf{R}^m$. During training, the InfoNCE loss for both text-to-image $\mathcal{L}_{\text{t2i}}$ and image-to-text $\mathcal{L}_{\text{i2t}}$ are minimized. In a batch of N examples, each pair of images and captions is considered a positive while the other elements in the batch are negatives. Given a learnable parameter τ, both losses can be formalized as:

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}\right), \tag{1}$$

Table 3: Retrieval performance of different CLIP variants in the RSICD dataset, used to motivate the choice of vision encoder. RS-CapRet leverages CLIP-ViT-L/14 finetuned with an aggregation of different remote sensing image captioning datasets, which we refer to as Cap-4.

Method	Visual Backbone	Finetune Data	Image-Text Retrieval			Text-Image Retrieval			
			R@1	R@5	R@10	R@1	R@5	R@10	mR
CLIP (Radford et al., 2021)	ViT-B	Zero-shot	4.58	14.55	23.70	5.80	16.85	28.23	15.62
CLIP (Radford et al., 2021)	ViT-L	Zero-shot	6.04	17.48	27.54	5.03	19.03	30.25	17.56
CLIP-RSICD (Pal et al., 2021)	ViT-B	RSICD	14.09	30.10	43.64	11.16	33.25	48.91	30.19
CLIP-RSICD-L	ViT-L	RSICD	14.27	32.02	46.39	12.11	34.97	50.47	31.70
CLIP-Cap-4	ViT-L	Cap-4	17.02	33.94	47.76	13.83	39.07	56.05	34.61
RemoteCLIP (Liu et al., 2024)	ViT-L	RemoteCLIP dataset	18.39	37.42	51.05	14.73	39.93	56.58	36.35

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\text{sim}(\mathbf{v}_i,\mathbf{u}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{v}_i,\mathbf{u}_i)/\tau)}\right), \tag{2}$$

where the similarity is the cosine similarity given by $\text{sim}(\mathbf{a},\mathbf{b}) = \exp(\mathbf{a}\cdot\mathbf{b})/(\|\mathbf{a}\cdot\mathbf{b}\|\|\mathbf{a}\cdot\mathbf{b}\|)$.

B.1 Choice of CLIP Vision Encoder

We first measured the retrieval performance (in the RSICD dataset (Lu et al., 2018)) of different CLIP model variants as a proxy to motivate our choice of vision encoder for RS-CapRet. The obtained results are compiled in Table 3. The larger variant of CLIP ViT-L/14 obtains higher results when compared to the ViT-B/32 variant. When measuring the results of an open-source version of a CLIP ViT-B/32 fine-tuned on RSICD, named CLIP-RSICD (Pal et al., 2021), the results increased and even surpassing those of the larger variant. We note that the authors of this model (Pal et al., 2021) used extensive augmentation strategies (both for images and caption text), as well as a high batch size, which highly benefits contrastive learning-based approaches (Radford et al., 2021; Chen et al., 2020). Motivated by both these results, we have also fine-tuned a CLIP ViT-L/14 with remote sensing data. When using multiple training datasets (i.e., Cap-4), the results further improved over the aforementioned CLIP variants, on both the RSICD and UCM datasets. From these results, we fix the vision encoder of RS-CapRet to CLIP ViT-L/14 finetuned with Cap-4 data, and illustrate this on the left section of Figure 1. For completeness, we also include the retrieval results for the RemoteCLIP (Liu et al., 2024), which leveraged an automatic procedure to scale the training data over +800k image-text pairs. Despite its better performance, we still used the CLIP-Cap-4 model because we had full control over its training procedure, which was done over only image-captioning datasets.

C Experimental Setup

This section presents a detailed description of the datasets used in the experiments, followed by implementation details.

C.1 Datasets

We leverage different remote sensing image captioning datasets for our setup. These are summarized in Table 4. The RSCID dataset (Lu et al., 2018) contains 10,921 images collected from different sources. These images were manually annotated, but many captions are duplicated to ensure 5 captions per image, due to some images not reaching that count during the annotation process. UCM-Captions and Sydney-Captions were proposed by Qu et al. (2016), which were based on scene classification data and were repurposed for image captioning by manual annotation. NWPU-Captions (Cheng et al., 2022) contains more data, with a total of 31,500 images. Each image has 5 manually annotated captions associated with it, corresponding to a total of 157,500 sentences. The authors intended to increase the category variety of the image and balance between classes. Therefore, the NWPU-Captions contains 45 classes describing different land cover and land use types. The spatial resolution of the images ranges between 0.2 and 30 meters. RS-CapRet was trained by combining all the aforementioned datasets to increase the quantity of available data and the diversity of geographical scenes represented. This aggregated dataset is referenced in the manuscript as Cap-4. In Table 4, we also include the RemoteCLIP dataset (Liu et al., 2024) as a reference. This latter

Table 4: The different remote sensing image captioning datasets used for the experiments.

Dataset	#Images	Image Size	Spatial Resolution	#Total Captions
NWPU-Captions (Cheng et al., 2022)	31,500	256×256	$\sim$30-0.2m	157,500
RSICD (Lu et al., 2018)	10,921	224×224	different resolutions	54,605
Sydney-Captions (Qu et al., 2016)	613	500×500	0.5m	3,065
UCM-Captions (Qu et al., 2016)	2,100	256×256	$\sim$0.3m	10,500
Cap-4	45,134	224×224	different resolutions	225,670
RemoteCLIP	165,745	different sizes	different resolutions	828,725

was generated through a pipeline that utilizes publicly available object detection and segmentation datasets to scale the quantity of image-text pairs.

C.2 Implementation Details

Our experiments were implemented in PyTorch (Paszke et al., 2019), and the model was trained with mixed-precision in bfloat16, to lower the memory requirements. The batch size was set to 64, the learning rate was set to 0.0003, with a warmup of 100 steps, and the Adam optimizer was used. Both the loss weights of the image captioning and contrastive learning tasks were equal to one, $\lambda_c = \lambda_r = 1$. For simplicity, we only considered one token embedding for visual information. The dimensionality of the contrastive learning space was set to $q = 256$. The gradient updates were only made on the parameters of the linear layers that were introduced and the `[RET]` embedding token. As for the input image size, a resizing operation was applied to the resolution of 224×224.

D Further discussion over the Results

In this section, more comments regarding the image captioning results are presented, covering experiments of further finetuning of the model, as well as others ablating the choice of different vision encoders.

D.1 Image Captioning

We extend some comments regarding the performance of RS-CapRet in the Sydney-Captions dataset. We note that this dataset is more specific compared to the other ones that were considered, both in constrained scene diversity (images are taken only over Sydney, divided into only 7 different types of classes), and textual description diversity. RS-CapRet was trained with the Cap-4 dataset, where the number of total captions from the Sydney-Captions dataset corresponds to $\sim 1\%$ of the total captions (see Table 4). Thus, our model does not have much data to align its generation outputs with descriptions that correspond better to the specific format expected for Sydney-Captions; however, RS-CapRet still achieves a relatively high performance. Overall, from the results across the different datasets, it can be concluded that RS-CapRet is a single model that can achieve high performance on heterogeneous remote sensing image captioning datasets.

D.1.1 Finetuning RS-CapRet to Image Captioning datasets

The results from the baselines MLCA-Net (Cheng et al., 2022) and RSGPT (Hu et al., 2023) came from training or fine-tuning with the respective dataset. SkyEyeGPT (Zhan et al., 2024) has also evaluated both the general model and finetuned versions, and observed that while the base general model achieved high results, finetuning would also help to improve them. To experiment with the role of fine-tuning in the dataset on which the evaluation is done, we further fine-tuned RS-CapRet (with a learning rate of 0.0001) for each dataset separately. The results are collected in Table 1, specifically in the row marked as *finetuned* for each dataset. It can be seen that fine-tuning furthers the performance, and the results are particularly expressive for the datasets of smaller size (UCM and Sydney-Captions). For NWPU, most metrics are improved, furthering even more the high performance of RS-CapRet in this dataset. As for RSICD, the results suffered a slight degradation for all metrics. In UCM, RS-CapRet already had higher performance in different metrics compared to the baselines (METEOR, ROUGE_L, CIDEr, SPICE), and with this fine-tuning the BLEU results (BLEU-2, BLEU-4) surpass those from the baselines as well. The experiments show that fine-tuning

Table 5: Comparison of results in image captioning and text-image retrieval, when changing the vision encoder of RS-RetCap from CLIP to one based on MAE (Wang et al., 2022), and considering encoders of different sizes. LLamaV2 was chosen as the language model, as in the original architecture, and the whole model was trained on Cap-4 data, following the training procedure of RS-CapRet.

Dataset	Visual Encoder	Visual Backbone	Image Captioning				Text-Image Retrieval			
			BLEU-1	BLEU-4	CIDEr	SPICE	R@1	R@5	R@10	mR
NWPU-Captions	RS-ViT-B (Wang et al., 2022)	ViT-B	0.810	0.547	1.542	0.269				
	CLIP-RSICD (Pal et al., 2021)	ViT-B	0.826	0.565	1.645	0.276				
	CLIP-Cap-4	ViT-L	0.871	0.650	1.919	0.320				
RSICD	RS-ViT-B (Wang et al., 2022)	ViT-B	0.706	0.410	2.329	0.449	5.00	17.97	30.01	17.66
	CLIP-RSICD (Pal et al., 2021)	ViT-B	0.728	0.439	2.524	0.466	7.47	25.36	40.73	24.52
	CLIP-Cap-4	ViT-L	0.741	0.455	2.605	0.484	9.83	30.17	47.43	29.14
UCM	RS-ViT-B (Wang et al., 2022)	ViT-B	0.699	0.467	2.347	0.374	9.71	41.43	69.81	40.32
	CLIP-RSICD (Pal et al., 2021)	ViT-B	0.810	0.606	2.901	0.439	11.14	48.48	83.33	47.65
	CLIP-Cap-4	ViT-L	0.833	0.645	3.429	0.525	15.52	57.24	88.76	53.84
Sydney-Captions	RS-ViT-B (Wang et al., 2022)	ViT-B	0.757	0.520	2.114	0.408				
	CLIP-RSICD (Pal et al., 2021)	ViT-B	0.772	0.538	2.177	0.406				
	CLIP-Cap-4	ViT-L	0.782	0.545	2.390	0.423				

exclusively on the dataset over which the evaluation is being done helps with the results. However, we argue the resulting models are less interesting, since they are now specific for a given dataset, and do not have the overall performance and generalization abilities that should be aimed for.

D.2 Usage of different Vision Encoders

As mentioned in previous sections, the remote sensing community has already proposed vision encoders finetuned for this specific domain, mainly with models based on masked autoencoder or masked image modeling objectives (Wang et al., 2022; Sun et al., 2022; Cha et al., 2023). We tested the use of a publicly available ViT-B encoder pre-trained on the MillionAID dataset with a MAE objective, which we refer in the text as RS-ViT-B (Wang et al., 2022). We also compare this encoder with another CLIP-based model finetuned to the remote sensing domain, sharing the same ViT-B backbone size, namely CLIP-RSICD (Pal et al., 2021). For this set of experiments, the decoder model is fixed also to LLamaV2-7b, and we also used the Cap-4 dataset for finetuning the connectors. Image captioning and text-image retrieval results for these experiments are reported in Table 5. It can be seen that using a vision encoder based on CLIP leads to better results in image captioning across the different datasets, as well as better results in text-image retrieval. CLIP-Cap-4 of larger size leads to better results on the complete model. These results follow findings from previous work, which argues that the use of vision encoders that were pre-trained with a text supervision signal leads to better performance when the embeddings of these models are integrated to create large vision-and-language models (Koh et al., 2023; Merullo et al., 2022).

E Qualitative Analysis

E.1 Describing Images

Some examples of captions generated by RS-CapRet are illustrated in Figure 2, where examples of images corresponding to different classes of the NWPU-Captions test set were chosen, namely *airplane*, *airport*, *baseball field*, *industrial area*, *medium residential area*, *forest*, *beach*, *freeway*, and *overpass*. An example of a caption of the dataset that is most simlar to the generated caption is included. These examples show that the model can generate short descriptions of remote sensing images, following the format of the captions in the original dataset, by describing a general overview of the image with a general description of the positional relationship between the most relevant objects.

E.2 Retrieving images from textual queries

Examples of image requests from textual queries can be seen in Figure 3. For the different queries, the top-3 images with the most similar embeddings to the `[RET]` token are shown. For the first three examples, a query was given to the model which generated the special `[RET]`, and with the first example generating a short description before it. It can be seen that the model can obtain

Figure 2: Qualitative examples of generated captions given images of different classes of the test set associated to the NWPU-Captions dataset (Cheng et al., 2022).

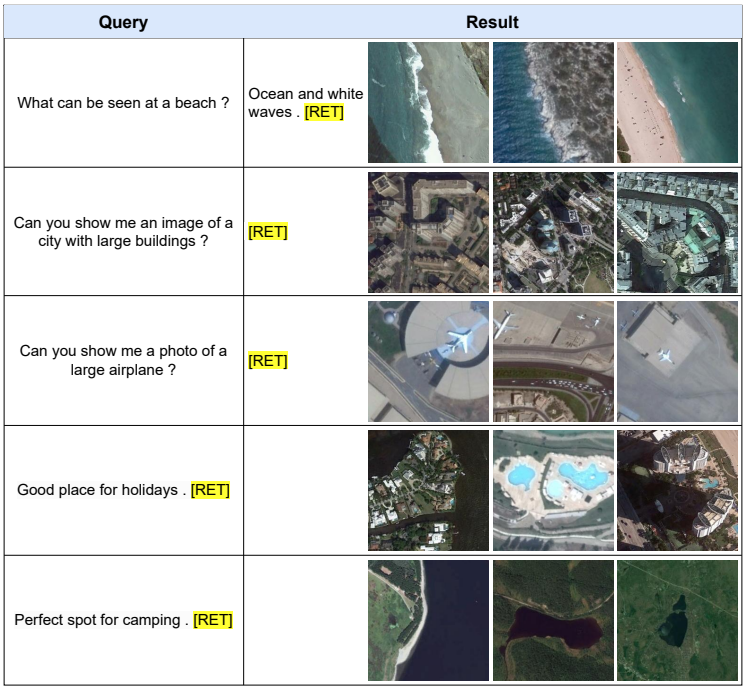

Figure 3: Examples of image retrieval achieved by RS-CapRet, given different requests by the user and considering object features and related topics.

images considering the different types of scenes or objects requested, namely a beach, a city with tall buildings, and a large airplane.

Due to the high performance of the LLM which was used to build RS-CapRet, the model can leverage knowledge between concepts that it has learned beyond our specific training. We experimented with asking the model to obtain images related to different concepts, in particular images related to holidays and camping spots, illustrated in the two bottom examples of Figure 3. For these examples, we directly appended the `[RET]` token to the end of the text, and retrieve the 3 most similar images to the obtained representation. In the example referring holidays, the images retrieved consist of pools and a hotel resort. Regarding a spot for camping, the images obtained depict a beach and a lake in a forest.

E.3 Exploring Dialogues

The ability to integrate visual embeddings into the input space of the LLM allows further combination of multi-modal inputs, with sequences of multiple images along with text and questions. This was further enhanced while training for the image captioning objective, where two image and text sequences were concatenated. We experimented with the ability of the model to handle short dialogues given multimodal inputs. Some dialogue examples are illustrated in Figure 4. In the example on the left, the model receives as input an image together with a question, and it can generate a description that answers the question. When asked to obtain a variation of the initial image to include buildings, without the explicit mention by the user that the scene being described is a golf course ("*Show me one likes this with...*"), the model can obtain an image that fits the desired criteria. A second variant is requested, also obtaining an image that fits the desired criteria (i.e. a golf course near the beach with sand). A second example is also shown on the right in Figure 4, regarding airport scenery and where the model can provide a description of the initial image, obtain a variant of the scenery, and another variant regarding the number of objects. Despite these promising results, we have observed that the model can also lack robustness, due to not having been directly optimized for this type of dialogue task during training. Still, these experiments show promising avenues for conversational agents for the remote sensing domain.

E.4 In-Context Learning

Finally, in Figure 5, we present some examples in which the model demonstrates in-context learning abilities to describe images. In particular, we present some examples where RS-CapRet struggles to adequately describe the image (in the first example there are no "lush woods", while the second example mentions a "palace", probably due to the shapes of the buildings), that can be refined by giving an example of the same class, leading to accurate descriptions of the initial image.

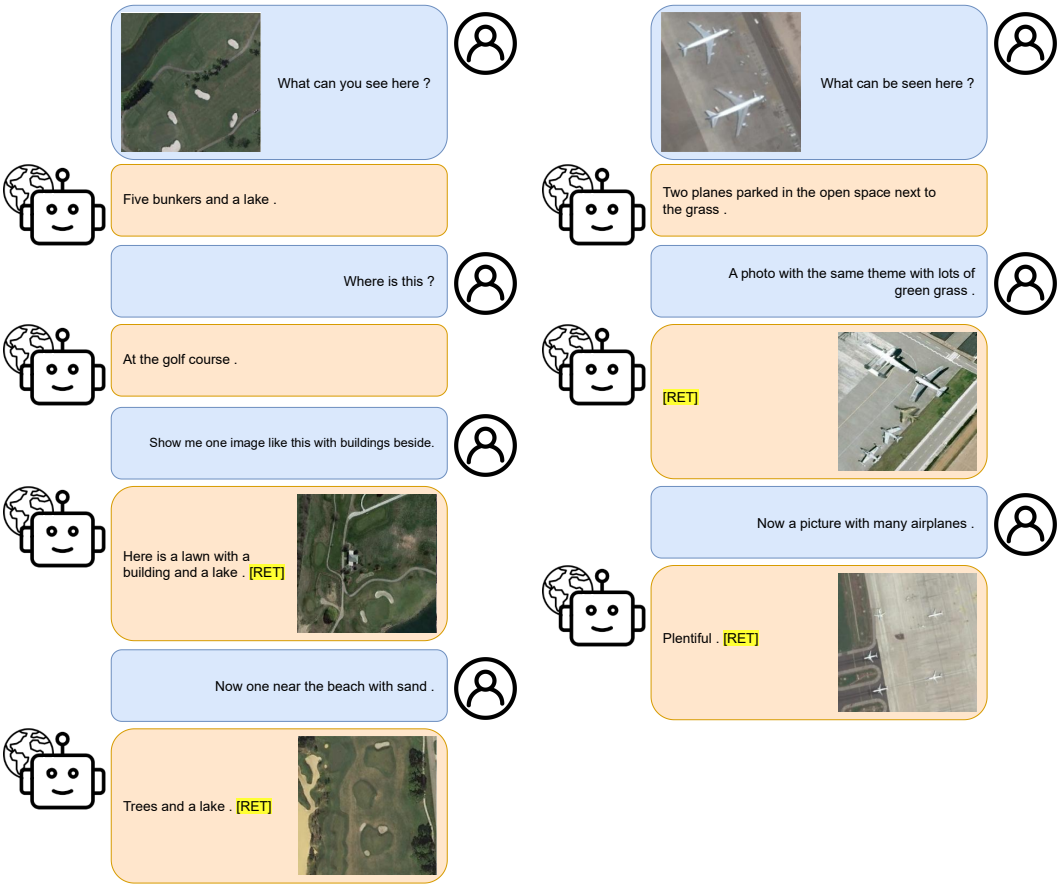

Figure 4: Examples of dialogue with RS-CapRet, showing (a) the ability to handle multi-modal inputs with interleaved sequences of images and text, as well as (b) reasoning abilities given world knowledge.

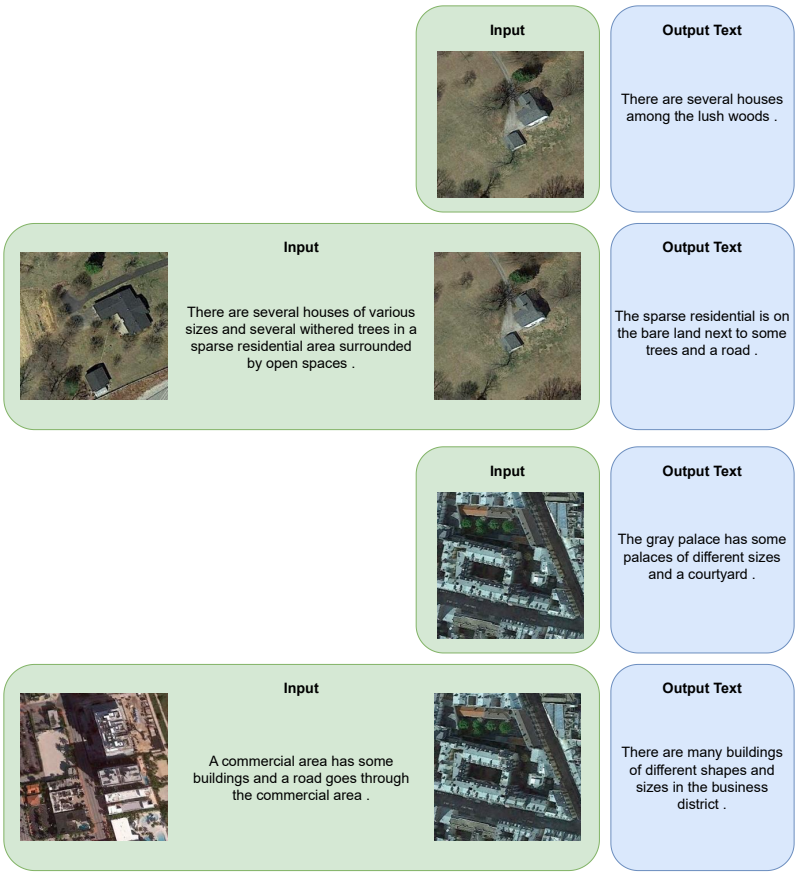

Figure 5: In-context learning ability of RS-CapRet. Given one example of the correct class of the input image, RS-CapRet can generate an accurate description, where it had before failed.