essd-2024-140 — Revision

ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models

By Z. Yuan, Z. Xiong, L. Mou, X. X. Zhu

General remarks to all reviewers and editors:

We sincerely thank the editors and anonymous reviewers for their valuable comments and suggestions. Below, we provide point-by-point responses to the reviewers' comments. The reviewers' comments are in black, and our responses follow in blue. The revised parts are marked in red in the manuscript.

Reviewer #2:

1. In the introduction section, I recommend that the authors include a discussion of recent work, VRSBench, which provides human-verified captions rich in object details.

R: We appreciate the reviewer's comment. We have included the discussion of the excellent recent work VRSBench in the revised introduction section as follows:

"The recent work VRSBench [1] offers a versatile benchmark featuring human-verified captions with detailed object information for remote sensing images."

[1] X. Li, J. Ding, and M. Elhoseiny. "VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding." arXiv preprint arXiv:2406.12384, 2024.

2. In Table 2, for evaluating long captions, traditional translation-based metrics can present challenges and may result in less reliable assessments. I suggest that the authors consider incorporating GPT-based metrics, such as CLAIR, to enable a more semantic-aware evaluation.

R: Thank you for your insightful comment regarding GPT-based metrics like CLAIR for more semantic-aware assessments. We agree that CLAIR is a promising metric for such evaluations. However, its implementation requires access to the OpenAI API, which can be costly and challenging to scale for large benchmark evaluations. While using other open-source LLMs can be an alternative, the inherent limitations of these models may affect the reliability and consistency of the metric. Nevertheless, we acknowledge the potential of CLAIR and plan to explore its application in our future work.

While traditional metrics may have limitations in evaluating long captions, they still provide useful and standardized means to compare the performance of different models. Given the current state of research and available resources, these metrics remain a reasonable choice for benchmarking different models.

As the primary focus of our manuscript is on the creation of the ChatEarthNet dataset, selecting the optimal benchmarking metric falls outside the main scope of our work. However, we appreciate the reviewer's suggestion and will consider incorporating metrics like CLAIR in future studies to enhance the evaluation framework.

Once again, we thank the reviewer for this constructive feedback, which will help guide our future research directions.

3. Table A1 currently lists only remote sensing image caption datasets; while image-text datasets cover more, such as VQA datasets and visual grounding datasets. The table caption should be changed.

R: Thanks for the valuable suggestion. We have revised the caption of Table A1 and the corresponding description. The revised caption is as follows.

Table A1. A summary of the remote sensing ~~mage-text~~ image captioning datasets.

Dataset	#Image-text pairs	Caption Granularity	Caption Generation	Image Data	Geographical Coverage
UCM-Captions (Qu et al., 2016)	10,500	Coarse-grained	Manually Annotated	RGB, UCMerced (Yang and Newsam, 2010)	Regional
Sydney-Captions (Qu et al., 2016)	3,065	Coarse-grained	Manually Annotated	RGB, Sydney (Zhang et al., 2014)	Regional
RSICD (Lu et al., 2017)	54,605	Coarse-grained	Manually Annotated	RGB, Google Earth, Baidu Map	Regional
NWPU-Captions (Cheng et al., 2022)	157,500	Coarse-grained	Manually Annotated	RGB, NWPU-RESISC45 (Cheng et al., 2017)	Regional
RSICap (Hu et al., 2023)	2,585	Fine-grained	Manually Annotated	RGB, DOTA (Xia et al., 2018)	Regional
RS5M (Zhang et al., 2023)	5,000,000	Coarse-grained	Model-generated & multiple datasets	RGB, multiple datasets	Global
SkyScript (Wang et al., 2024)	2,600,000	Coarse-grained	OpenStreetMap	RGB & multispectral, multiple sensors	Global
FIT-RS (Luo et al., 2024)	1,800,851	Fine-grained	STAR & ChatGPT	RGB, STAR (Li et al., 2024)	Global
RemoteCLIP (Liu et al., 2024)	828,725	Coarse-grained	Rule-based	RGB, multiple datasets	Global
ChatEarthNet	173,488	Fine-grained	WorldCover & ChatGPT	RGB&multispectral, Sentinel-2	Global

Reviewer #3:

R: Thank you for recommending our manuscript for acceptance. We are grateful for your acknowledgment of our work.