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 #1:

The authors propose a land cover dataset, ChatEarthNet, built by pairing Sentinel-2 patches with their corresponding WorldCover masks, which contain 12 land cover classes.

The originality comes from providing the land cover data, not directly a a bitmap, but as a textual description extracted from the WorldCover map by means of a large language model (LLM).

Specifically, they use two different models: ChatGPT-3.5, an LLM that can only receive text as input, and ChatGPT-4V, a vision LLM (VLLM) that is able to understand both text and images. Due to cost, they provide 163k images with captions generated by GPT-3.5 and 10k by GPT-4V.

The Sentinel-2 patches are obtained from the dataset SatlasPretrain.

Main comments:

1. The paper describes the prompting process, which differs for GPT-3.5 and -4V. Although the prompt is provided, some details are missing in relation to the exact construction of the outputs of algorithms 1 to 3, since the exact wording of the prompt produced by these algorithms is not given.

R: We appreciate the reviewer's comment. We have included the exact wording produced by algorithms 1 to 3 in Appendix A. Below are examples of the outputs generated by these algorithms:

An example of the exact wording of the prompt generated by Algorithm 1:

An example of prompt output by Algorithm 1

grass; tree; developed area; crop; water; bare land.
The **top left** mainly contains the following land cover types, in descending order of content:
grass (medium part), tree (medium amount), and developed area (medium amount).
The **top right** mainly contains the following land cover types, in descending order of content:
tree (medium quantity), grass (medium amount), and developed area (small amount).
The **bottom left** mainly contains the following land cover types, in descending order of content:
grass (medium amount), tree (medium fraction), and developed area (medium fraction).
The **bottom right** mainly contains the following land cover types, in descending order of content:
crop (medium part), grass (medium portion), and developed area (small portion).
The **middle** mainly contains the following land cover types, in descending order of content:
tree (medium part), grass (medium portion), and developed area (medium fraction).

An example of the exact wording of the prompt generated by Algorithm 2:

An example of prompt output by Algorithm 2

crop: top left: 72.27% top right: 43.16% bottom left: 41.78% bottom right: 58.15% middle: 39.85%
grass: top left: 10.02% top right: 26.73% bottom left: 36.18% bottom right: 16.22% middle: 27.62%
developed: top left: 14.67% top right: 24.27% bottom left: 15.21% bottom right: 21.97% middle: 23.75%
water: top left: 1.68% top right: 3.22% bottom left: 2.73% bottom right: 0.21% middle: 3.45%
bare: top left: 0.12% top right: 0.21% bottom left: 0.11% bottom right: 0.29% middle: 0.28%
tree: top left: 1.04% top right: 1.06% bottom left: 3.98% bottom right: 2.92% middle: 4.46%

An example of the exact wording of the prompt generated by Algorithm 3:

An example of prompt output by Algorithm 3

top left distribution: crop: 0.72; developed: 0.15; grass: 0.10; water: 0.02; tree: 0.01; bare: 0.00;
top right distribution: crop: 0.43; grass: 0.27; developed: 0.24; water: 0.03; tree: 0.01; bare: 0.00;
bottom left distribution: crop: 0.42; grass: 0.36; developed: 0.15; tree: 0.04; water: 0.03; bare: 0.00;
bottom right distribution: crop: 0.58; developed: 0.22; grass: 0.16; tree: 0.03; bare: 0.00; water: 0.00;
middle distribution: crop: 0.40; grass: 0.28; developed: 0.24; tree: 0.04; water: 0.03; bare: 0.00;

Bold and underline are used to improve readability. Note that in Algorithm 2, we calculate the percentage of a specific land cover in each patch, not the percentage of one land cover in the entire image. Therefore, the sum of the percentages is not 1. These examples are added to Appendix A to provide a clear understanding of how the exact outputs of the algorithms are used to construct the prompts.

2. Section 2.5 briefly mentions that manual verification is applied in order to check that the LLM correctly followed the prompt instructions. However, it is not clear how many times the prompt had to be modified, and the kind of modifications that were required.

R: Thanks for the comment. To clarify the potential confusion, the manual verification process described in Section 2.5 aims to ensure the quality and correctness of the generated captions, not modify the prompts. Once the prompts were finalized, we did not further modify them. Instead, our manual verification process focuses on reviewing and correcting the generated captions to ensure they meet our quality standards.

Unfortunately, we did not track the number of samples where modifications were made for captions, making it difficult to provide exact answers. However, we would like to emphasize the reason for manual verification: when using ChatGPT-4V, we combine four images into a single API request. We adopt this approach due to the constraints on API usage for ChatGPT-4V (with a tier 1 limit of 500 requests per day by February 2024) and to enhance the efficiency of text generation. Despite providing specific instructions for ChatGPT-4V to treat each image individually, it occasionally makes mistakes by describing comparisons between images, which is not our intention. In such cases, manual corrections are necessary. In contrast, when using the ChatGPT-3.5 model (with a tier 1 limit of 10k requests per day by February 2024), each image is processed through individual API requests.

3. Although the authors claim that "10k high-quality image-text pairs using ChatGPT-4V are sufficient for fine-tuning large vision-language models", they do not provide any evidence for this. There is not evaluation of the properties of a model trained with the proposed dataset, making it impossible to judge the quality of the representation that can be learned with it, in comparison with a model trained directly for land cover mapping using the WorldCover data.

R: Thanks for the insightful comments. We appreciate the opportunity to provide further evidence and clarification regarding the quality of our dataset. Our claim that the dataset consists of high-quality image-text pairs is grounded in the following key factors:

1) Our dataset provides rich descriptions of the Sentinel-2 images, which contain information about shapes, spatial relationships, distributions, and the main theme of the image, which can be used to train or fine-tune multimodal large language models (MLLMs). Regarding the comparison with traditional segmentation models trained using WorldCover data, a key difference is evident: traditional segmentation models trained on WorldCover data lack the ability to generate rich linguistic descriptions of shapes, spatial relationships, and distributions of land cover types. This limitation indicates the superiority of our dataset in capturing and conveying complex geospatial information through natural language.

2) To further support our claim, we provide experiments to fine-tune MLLMs to prove that the proposed dataset can be used to enhance the development of large vision language models. As shown in Table 2, compared with existing MLLMs in the zero-shot setting, like LLaVA-v1.5 [1], MiniGPT-v2 [2], MiniGPT-4 [3], and GeoChat [4], the fine-tuned models using ChatEarthNet (ChatGPT-4V version) can achieve clearly better performance. The results indicate that the proposed ChatEarthNet dataset is

not only useful for downstream applications but also effective as a benchmark to evaluate different MLLMs. Please refer to *Comment #4* for more details.

3) **It is worth noting that the rich descriptions with natural language on Sentinel-2 images in our dataset provide opportunity to non-expert users to understand the sentinel images, who may have difficulty understanding WorldCover labels.**

Table 2. Performance comparison of different models on the ChatEarthNet (ChatGPT-4V Version) test set.

Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	METEOR	ROUGE_L	SPICE
LLaVA-v1.5	0.285	0.116	0.040	0.014	0.012	0.104	0.186	0.093
MiniGPT-v2	0.279	0.116	0.041	0.015	0.009	0.104	0.180	0.091
MiniGPT-4	0.175	0.072	0.023	0.008	0.000	0.116	0.180	0.079
GeoChat	0.199	0.088	0.034	0.011	0.005	0.067	0.126	0.083
MiniGPT-4 (ChatEarthNet)	0.310	0.184	0.113	0.071	0.001	**0.209**	0.254	0.186
GeoChat (ChatEarthNet)	**0.445**	**0.269**	**0.170**	**0.109**	**0.094**	0.208	**0.286**	**0.211**

4. The authors conclude that "ChatEarthNet is a valuable resource for training and evaluating vision-language geo-foundation models for remote sensing". However, it is not fully clear how this evaluation would work. To be able to conclude this, I suggest the authors do use the dataset to evaluate existing models, such as RemoteCLIP [1], RSCLIP [2] and others.

R: We thank the reviewer for the valuable suggestion. We agree that demonstrating ChatEarthNet's utility for evaluating vision-language geo-foundation models is crucial. In response to the suggestion, we have conducted benchmarking experiments using various existing models on the proposed dataset.

Regarding the reviewer's recommendations to include RemoteCLIP [1] and RSCLIP [2], we appreciate the suggestion. However, the pretrained RSCLIP model is not publicly available at this time. Reproducing its training process would require significant computational resources, which presents substantial challenges. As such, direct evaluation of RSCLIP is currently not feasible.

As for RemoteCLIP, while it is a CLIP-based model suitable for vision tasks, applying it directly to ChatEarthNet, which contains long and detailed descriptions, would require extensive alignment with large language models through training connectors on a sizable dataset. This process is resource-intensive and beyond the scope of this paper. We thank the reviewer for pointing out this, and we have added explanations in the revised version, as presented below.

To address the reviewer's concerns, we have conducted benchmarking experiments using widely established MLLMs. Specifically, we evaluated several MLLMs, including LLaVA-v1.5 [3], MiniGPT-v2 [4], MiniGPT-4 [5], and GeoChat [6]. These evaluations further support our conclusion that ChatEarthNet is a valuable resource for training and evaluating

vision-language geo-foundation models for remote sensing. We have added the experimental part to the revised paper as follows.

"To demonstrate the effectiveness of ChatEarthNet in evaluating multimodal large language models, we conduct benchmarking experiments using a range of existing models. Given that ChatEarthNet includes long and detailed descriptions, it is not well-suited for evaluating CLIP-based vision-language models like RemoteCLIP [1] and RS-CLIP [2]. Therefore, we focus on evaluating widely used multimodal large language models, including LLaVA-v1.5 [3], MiniGPT-v2 [4], MiniGPT-4 [5], and GeoChat [6]. All experiments are performed using the ChatGPT-4V version of our dataset, which allows us to conduct extensive evaluations across multiple models while significantly reducing computational resource requirements.

Table 2 summarizes the results of these evaluations, detailing the models' performance across several widely used metrics: BLEU, CIDEr, METEOR, ROUGE-L, and SPICE. We evaluate these models in two experimental settings. The first is a zero-shot transfer setting, where pre-trained models are used to generate captions without any additional training or fine-tuning on the ChatEarthNet dataset. The first four rows in Table 2 present the results of this zero-shot transfer setting. The performance is suboptimal due to the domain gap between the models' original training datasets and our test dataset. In addition to zero-shot testing, we fine-tune some of these models on the ChatEarthNet dataset (ChatGPT-4V version) and report their performance. The results clearly show that fine-tuning on our proposed dataset significantly improves image captioning performance in the context of remote sensing data. These findings strongly suggest that ChatEarthNet is a valuable resource for both training and evaluating vision-language geo-foundation models in the remote sensing domain."

Table 2. Performance comparison of different models on the ChatEarthNet (ChatGPT-4V Version) test set.

Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	CIDEr	METEOR	ROUGE_L	SPICE
LLaVA-v1.5	0.285	0.116	0.040	0.014	0.012	0.104	0.186	0.093
MiniGPT-v2	0.279	0.116	0.041	0.015	0.009	0.104	0.180	0.091
MiniGPT-4	0.175	0.072	0.023	0.008	0.000	0.116	0.180	0.079
GeoChat	0.199	0.088	0.034	0.011	0.005	0.067	0.126	0.083
MiniGPT-4 (ChatEarthNet)	0.310	0.184	0.113	0.071	0.001	**0.209**	0.254	0.186
GeoChat (ChatEarthNet)	**0.445**	**0.269**	**0.170**	**0.109**	**0.094**	0.208	**0.286**	**0.211**

[1] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "RemoteCLIP: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[2] X. Li, C. Wen, Y. Hu, and N. Zhou, "RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision," *International Journal of Applied Earth Observation and Geoinformation*, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems 36*, 2023.

[4] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
[5] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
[6] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Minor comments:

1. In Section 3.2, the authors write that "ChatGPT-3.5 is more dense, covering a wider range of areas". However, aren't both datasets obtained by randomly sampling SatlasPretain? Shouldn't they therefore have roughly the same distribution? If I understand it well, the only difference should be the number of images.

R: Thank you for your question. You are correct. Due to the cost and access limitations of ChatGPT-4V, the number of images used in ChatGPT-4V is significantly lower compared to the number used in ChatGPT-3.5. Regarding the geographical coverage, they basically have roughly the same distribution. The only difference is the density of coverage.

2. In Section 3.3, they authors explore word frequency in the generated captions.

R: The word frequency analysis in Section 3.3 provides valuable insights into the linguistic characteristics of the generated captions.

3. In line 50, "few pairs in the website" should be "few pairs on the web" or "online".

R: Thank the reviewer for pointing out this. We have revised the relevant sentence to: "However, few pairs on the web provide detailed descriptions for satellite images." Please kindly check out the revised version.

4. The authors often refer to "land covers", although may be "land cover types" or "classses" would be more appropriate.

R: We appreciate the reviewer's suggestions on this point. We agree that "land cover types" or "classes" are more appropriate terms. We have revised the relevant terms in the manuscript.