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 #3:

The authors propose an image-text dataset for remote sensing vision-language geo-foundation models. In detail, the image source is from Sentinel-2 data, and the descriptions of land covers is obtained from the semantic segmentation labels of the European Space Agency's WorldCover project. Moreover, ChatGPT and the manual verification process are introduced to enhance the dataset. The presented work focus on considerable data collection and processing, however the experimentation could be further improved. The reviewer has the following comments:

Main comments:

1. In this work, a global image-text dataset is presented in the field of remote sensing. There are some existing image-text datasets, and the authors are encouraged to specifically compare the proposed dataset with those that exist, such as SkySenseGPT (https://arxiv.org/pdf/2406.10100), SkyScript (https://ojs.aaai.org/index.php/AAAI/article/view/28393), and RemoteCLIP (https://ieeexplore.ieee.org/document/10504785).

R: Thanks for this valuable suggestion. We have added comparisons with the mentioned datasets in Appendix A. Please kindly check out the table as follows.

Table A1. A summary of the remote sensing image-text 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 M	Coarse-grained	Model-generated & multiple datasets	RGB, multiple datasets	Global
SkyScript (Wang et al., 2024)	2.6 M	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

Although the number of FIT-RS dataset proposed in the SkySenseGPT paper is greater than that in ChatEarthNet, this work was submitted to arXiv in June 2024, which is four months later than the submission of ChatEarthNet to arXiv in February 2024.

2. For the designed dataset, how to consider the imbalance between foreground and background in the remote sensing segmentation task?

R: Thank you for your question. In our dataset, we use land cover maps to generate detailed descriptions of all land cover types present in the images. As a result, there is no explicit "background" in the sense. Each region in the image is represented by a specific land cover type, as shown in Figs A1, A2, and A3 in Appendix A. Consequently, there is no imbalance issue between foreground and background, as all land cover types are treated equally in the descriptions.

3. The authors are advised to explain the reasons for choosing the land cover maps from WorldCover. In addition, how to measure the accuracy of labelling in these land cover maps?

R: Thanks for the valuable comment. WorldCover is selected for our dataset due to its high accuracy and comprehensive land cover types when compared to other available products. How to measure the accuracy of the global land cover products is challenging. To address this issue, Xu et al. conducted a comparative independent validation of recent 10m global land cover maps. As demonstrated in the study by Xu et al. [1], WorldCover outperforms other alternatives such as Dynamic World [2] and ESRI LULC [3], offering more accurate labels and more land cover types. These factors inspire us to choose WorldCover for constructing our dataset instead of others.

[1] P. Xu, N. E. Tsendbazar, M. Herold, S. de Bruin, M. Koopmans, T. Birch, S. Carter, S. Fritz, M. Lesiv, E. Mazur, A. Pickens, P. Potapov, F. Stolle, A. Tyukavina, R. Van De Kerchove, and D. Zanaga, "Comparative validation of recent 10 m-resolution global land cover maps," *Remote Sensing of Environment*, 2024.
[2] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, K. Schwehr, M. Weisse, F. Stolle,

C. Hanson, O. Guinan, R. Moore, and A. M. Tait, "Dynamic World, near real-time global 10 m land use land cover mapping," *Scientific Data*, 2022.
[3] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, "Global land use / land cover with Sentinel 2 and deep learning," *IEEE International Geoscience and Remote Sensing Symposium*, 2021.

4. The authors mentioned that the proposed dataset has many high-quality and detailed descriptions, and is it validated by quantitative comparison experiments with other datasets?

R: We appreciate the reviewer's comment on this point. We provided a table to compare the proposed dataset with existing ones as follows.

Table A1. A summary of the remote sensing image-text 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 M	Coarse-grained	Model-generated & multiple datasets	RGB, multiple datasets	Global
SkyScript (Wang et al., 2024)	2.6 M	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

We also conduct experiments to evaluate widely established multimodal large language models (MLLMs). Specifically, we evaluate several MLLMs, including LLaVA [4], MiniGPT-v2 [5], MiniGPT-4 [6], and GeoChat [7]. These evaluations further support our conclusion that ChatEarthNet is a valuable resource for training and evaluating vision-language geo-foundation models for remote sensing. Please kindly check out the revised version 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 [8] and RS-CLIP [9]. Therefore, we focus on evaluating widely used multimodal large language models, including LLaVA-v1.5 [4], MiniGPT-v2 [5], MiniGPT-4 [6], and GeoChat [7]. 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**

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems 36*, 2023.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.

5. The authors are encouraged to discuss which attributes are more important for the multimodal vision-language learning than for the vision representation, e.g., relative size or relative position described in the text.

R: Thanks for the insightful comment. In multimodal vision-language learning, attributes like relative size and relative position are important because they provide contextual information that bridges the gap between visual data and natural language. While vision-

only representations are good at capturing visual features such as color, texture, and shape, they may fall short in conveying spatial relationships and comparative attributes inherent in complex scenes like satellite imagery. For instance, understanding that "a small lake is nestled beside a large forest" requires integrating both visual cues and linguistic descriptions to fully comprehend the scene.

ChatEarthNet emphasizes these attributes in the generated descriptions. By employing detailed prompts and leveraging semantic segmentation labels from the WorldCover project, we ensure that the natural language descriptions include rich details about relative sizes and positions. This enriches the dataset, making it more suitable for training models that need to understand and generate descriptions involving spatial relationships and comparative sizes.

We believe that highlighting these attributes enhances the performance of multimodal models in tasks such as image captioning, scene understanding, and geospatial analysis. It allows models to develop a more comprehensive understanding of the scene by aligning visual features with corresponding textual descriptions that capture both absolute and relative attributes. In contrast to vision-only models, which might detect objects without understanding their spatial relationships, multimodal models can interpret and describe how different elements in an image relate to one another, leading to more informative and accurate outputs.

We appreciate the reviewer's insightful comment. However, this manuscript mainly focuses on the construction and analysis of the dataset. In the future work, we plan to conduct experiments to quantify the impact of these attributes on model performance.

6. It seems that there is a lot of textual information described in the proposed dataset, does this introduce interfering information? How to avoid negative learning due to interfering information?

R: Thank you for your insightful comment. We acknowledge that incorporating extensive textual information can introduce challenges, such as noise or irrelevant details that might negatively impact model training. However, with careful prompt design and semantic guidance, we can mitigate these concerns, ensuring that the dataset enhances learning rather than hinders it.

We utilize detailed and carefully designed prompts to guide ChatGPT in generating descriptions that are both informative and relevant, avoiding inaccuracies, redundancies, or irrelevant details that could introduce noise. For example, we only focus on three main land cover types in Algorithm 1 instead of all land cover types. Moreover, by incorporating semantic segmentation labels from the WorldCover project, we ensure that the descriptions focus on land cover types and spatial relationships in the image. This semantic guidance helps filter out irrelevant information and emphasizes attributes that are crucial for understanding and interpreting remote sensing data.

From the perspective of dataset construction, we have implemented several strategies to enhance quality. However, mitigating negative learning mainly depends on model design,

which is beyond the scope of this paper focusing on introducing the dataset. Nevertheless, we believe this is a valuable research direction and intend to pursue it in our future work.

Minors

1. Please rephrase the description of "Image-Text Dataset", could the proposed dataset be used with other vision-language tasks, such as, image-to-text and text-to-image synthesis?

R: Thank you for your valuable comment and question. Yes, the ChatEarthNet dataset can indeed be readily used for other generative tasks, such as image-to-text and text-to-image synthesis. In addition, the dataset can also be easily extended to visual question answering by leveraging the capabilities of current large language models.

This versatility is why we refer to it as an "image-text dataset," a high-level term that captures its potential for a range of tasks. We have added further clarification in the revised manuscript as follows:

"It is worth noting that the proposed ChatEarthNet dataset can be readily used for other tasks, including image-to-text and text-to-image synthesis. Moreover, leveraging the capabilities of large language models, it can also be extended to visual question answering by prompting large language models for questions and answers based on rich descriptions. This versatility enhances the dataset's value to the community."

2. The "2.5 Manual verification" section is suggested to add details of manual adjustments, such as under what circumstances manual verification are required and what information is adjusted. An example is visual representation.

R: Thank you for your valuable suggestion. We have added more details to section "2.5 Manual verification" to better present the manual adjustment process. Please kindly check out the revised version as follows.

"To avoid unexpected descriptions on comparisons between different images, we design prompts like "Generate the four descriptions separately; do not add connections between them" to guide the description generation process. Despite providing specific instructions for ChatGPT-4V to treat each image individually, it occasionally make mistakes by describing comparisons between images. For instance, phrases such as "similar to other images" and "compared with previous images," need to be revised to eliminate comparisons. We therefore manually check all captions and refine comparison-related captions."

During the manual verification process, the old captions are overwritten, making it impossible to retrieve the precise before-and-after states for comparison. However, to illustrate the general concept, we can provide a hypothetical example that demonstrates the essence of the process:

Original description: "The fourth image features a noticeable spread of developed areas, with a larger extent than in the other images, especially strong in the top left and middle regions, indicative of a significant human footprint. Grassland areas are uniformly

distributed throughout, suggesting a balance between natural landscapes and developed spaces. Crops are situated in the lower quadrants, forming large agricultural plots. Similarly, the depiction of this landscape suggests a balance between urban development and agricultural uses with some remaining grassland regions."

Corrected description: "This image features a noticeable spread of developed areas, especially strong in the top left and middle regions, indicative of a significant human footprint. Grassland areas are uniformly distributed throughout, suggesting a balance between natural landscapes and developed spaces. Crops are situated in the lower quadrants, forming large agricultural plots. The depiction of this landscape suggests a balance between urban development and agricultural uses with some remaining grassland regions."

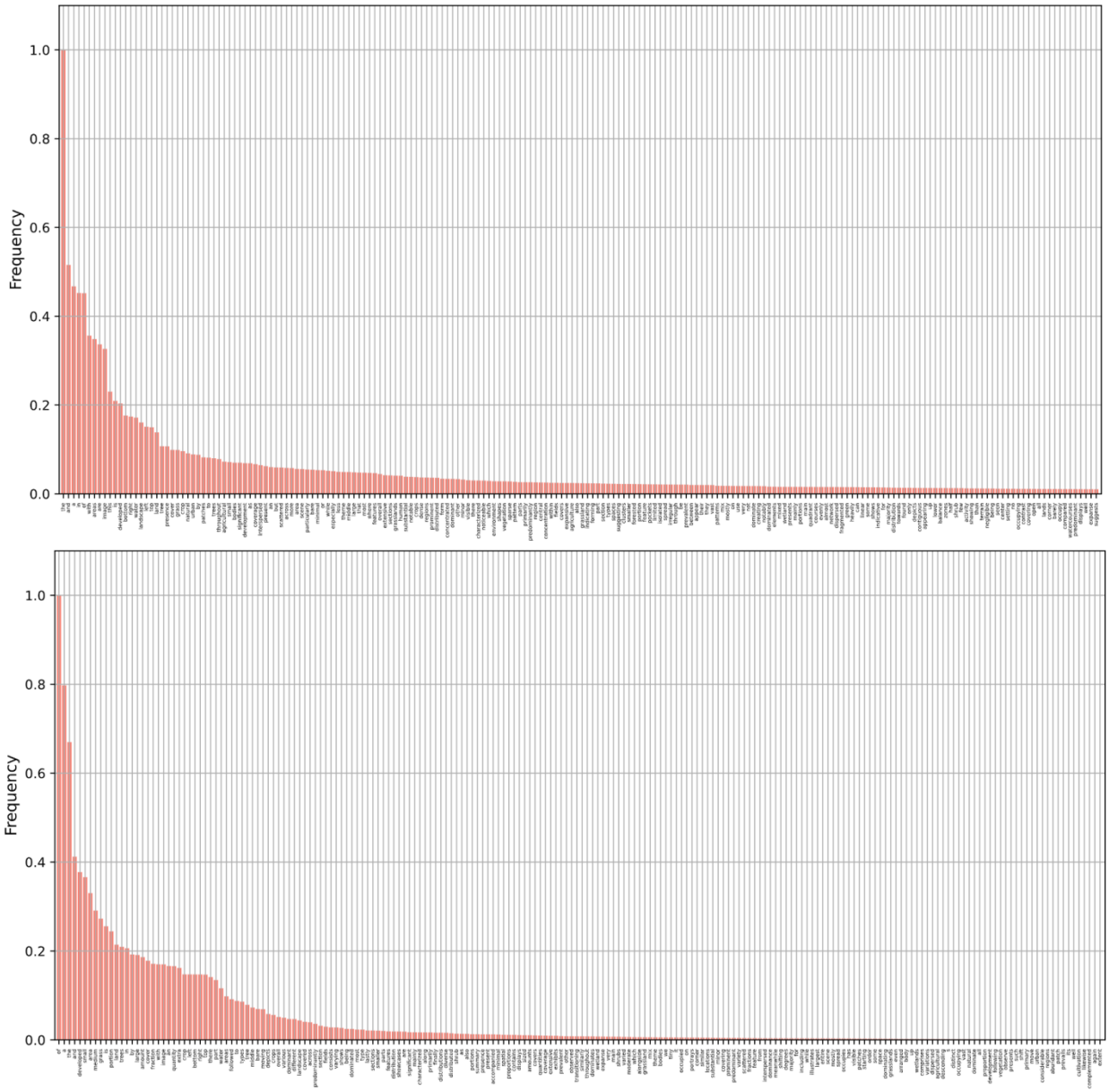
In this example, "The fourth image" changes to "This image," "with a larger extent than in the other images," is removed, and "Similarly," is removed. This shows how the manual verification process is done by removing comparative elements. We hope this example helps clarify the manual verification process.

3. The y-axis of Figs. 9-10 are suggested to be revised to the same to make contrasts clearer.

R: Thank you for your valuable suggestion. We agree that using the same y-axis scale for both figures could make comparisons clearer. However, there are some considerations to keep in mind.

If we adjust the y-axis to the maximum frequency value of 2,000,000 for both figures, the plot for ChatGPT-4V will become difficult to interpret due to its significantly lower frequency numbers. Alternatively, normalizing the frequency to a range of 0 to 1, as shown below, results in plots that are visually identical to Figs. 9-10 in the manuscript. However, Figs. 9-10 provide more insight into the actual number of samples in the two subsets by retaining the true value range of the frequencies.

While we appreciate the reviewer's suggestion, we believe that preserving the original y-axis scale provides more meaningful information and would prefer to retain the current version.



4. The authors claim that “it stands out as the first dataset offering high-quality detailed land cover descriptions on a global scale” on line 230 of page 14. Please replace this expression with a more accurate description.

R: Thank you for pointing out this. We have revised this sentence as follows:

“In terms of the number of image-text pairs, the ChatEarthNet dataset is not the largest dataset available, but it offers high-quality detailed land cover descriptions on a global scale. ”

5. Page 10 has gaps, and the authors are encouraged to reformat the article.

R: Thank you for the insightful suggestion. We have addressed the gaps in the layout. Please kindly check out the revised version.