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:

I thank the authors for responding to the reviewers' comments.
I particularly appreciate the new experiments on using the dataset to evaluate vision-language models. However, there are still a few elements missing:

R: Thank you for your thoughtful feedback on our response letter and for appreciating the additional experiments we conducted. Below, we provide detailed responses to the additional concerns raised.

1. The new experiments (Table 2), seem to show captioning scores for several models, including GeoChat. However, there are no details as to how the setting of this evaluation. Were the same prompts used as for creating the dataset, but with the satellite image instead of the LC map?
I also miss some discussion as to why GeoChat, which has been trained with RS images, doesn't perform better than the other models. After all, one interpretation is that the proposed dataset may not be so useful for evaluating RS VLMs after all, or maybe the specific evaluation setting is not appropriate. A discussion on this, along with examples of generated captions, would be important.

R: We appreciate the reviewer's comments and respond as follows.

1) **To clarify, the prompt used during dataset creation and the instruction prompt for model evaluation are entirely different. When creating the dataset, we utilize land cover (LC) maps alongside carefully designed prompts to generate rich descriptions of the corresponding satellite images.**

   However, for model training and evaluation, only satellite images are used as visual input, without any LC maps. During evaluation, all models are provided with the same instruction prompt to ensure fairness. Specifically, for the captioning task, the evaluation prompt is a variant of the instruction: "Provide a detailed description of the given image." These evaluation prompts are unrelated to the prompts designed for dataset creation.

To clarify this part further, we have added a description to Section 3.6 as follows:

Note that the prompt used during dataset creation and the instruction prompt for model evaluation are entirely different. Dataset creation involves leveraging land cover maps and designing prompts to generate rich descriptions of satellite images. In contrast, during both training and evaluation of models, only satellite images are used as visual input. To ensure consistency and fairness, all models are evaluated using the same instruction prompt: "Provide a detailed description of the given image" or its variants.

2) **Thank you for highlighting this valuable question: why GeoChat [1] doesn't beat others in the zero-shot evaluation? In our zero-shot evaluation setting, GeoChat, fine-tuned on limited datasets, struggles to perform well on ChatEarthNet due to differences in spatial resolution and content coverage. However, other models are not fine-tuned on specific remote sensing datasets, which sometimes results in better generalizability. To further explain this, we add a comparison of the training datasets used by GeoChat with our ChatEarthNet dataset. GeoChat's performance on our ChatEarthNet dataset is influenced by a significant domain gap between its original training datasets and our proposed dataset. Specifically, GeoChat is trained on six datasets [2]-[7] designed for tasks like object detection, visual question answering, and scene classification on high-resolution remote sensing images, as outlined below:**

- **DOTA [2]: A dataset specifically designed for object detection in remote sensing images, with a focus on high-resolution spatial data and object categories such as ships, tennis courts, and small vehicles.**
- **DIOR [3]: Another object detection dataset with categories such as vehicle, stadium, and wind mill.**
- **FAIR1M [4]: Also an object detection dataset featuring high-resolution remote sensing imagery, providing object categories such as ship, road, and court.**
- **LRBEN [5]: A visual question answering (VQA) dataset in remote sensing, primarily addressing urban-rural classification, presence of elements (e.g., roads and buildings), and simple quantitative or comparative questions. It lacks comprehensive land use land cover (LULC) analysis.**
- **FloodNet [6]: A VQA dataset focusing on flood-related categories like flooded and non-flooded buildings or roads, with a significant domain gap from our dataset.**
- **NWPU-RESISC45 [7]: A classification dataset covering diverse scene types with varying spatial resolutions, such as bridge, church, and intersection.**

Key differences between GeoChat's training datasets and ChatEarthNet include:

- **Objective mismatch: GeoChat's training datasets mainly target object-centric tasks and specific queries, whereas ChatEarthNet emphasizes LULC-related semantics.**
- **Spatial resolution: Most of GeoChat's training data comprises high-resolution images focusing on objects, different from ChatEarthNet's broader geographical context.**
- **Content gap: GeoChat's training datasets lack comprehensive LULC-related descriptions required for detailed LULC analysis.**

We add a more detailed explanation of experimental results in the Section 3.6 as follows:

Specifically, the original GeoChat model exhibits unsatisfactory zero-shot performance on the ChatEarthNet dataset due to the substantial domain differences between its training datasets and our proposed dataset. GeoChat is trained primarily on high-resolution datasets designed for tasks such as object detection, visual question answering, and scene classification, which lack the global-scale land use and land cover-related semantics and descriptions. The differences in spatial resolution, coupled with the lack of comprehensive land cover content, significantly limit GeoChat's performance on ChatEarthNet. These gaps also motivate the need for ChatEarthNet to complement existing datasets.

3) **We would like to explain why we believe ChatEarthNet is well-suited as a benchmark for evaluating RS VLMs. RS VLMs are designed to integrate knowledge across different remote sensing data modalities and domain-specific expertise. ChatEarthNet, with its focus on global-scale LULC analysis, provides a comprehensive and challenging benchmark to test the versatility and generalization ability of these models.**

   When fine-tuned on ChatEarthNet, models show significant performance improvements, demonstrating the dataset's ability to complement existing datasets. This makes ChatEarthNet an essential tool for assessing the breadth and depth of RS VLMs' knowledge in LULC applications on a global scale.

4) **To provide a clearer understanding of the evaluation results, we include examples to compare the original predictions of GeoChat with those obtained after fine-tuning with ChatEarthNet. These examples demonstrate how fine-tuning the model on ChatEarthNet improves the model's ability to generate detailed and accurate descriptions. For consistency, we use the same instruction prompt: "Provide a detailed description of the given image." This comparison highlights the value of ChatEarthNet in enhancing RS VLMs' capabilities for global-scale LULC analysis.**

Fig. 1. Sentinel-2 satellite image and its associated land cover map. Note that during the training and evaluation process, only the sentinel-2 satellite image is used as visual input. The land cover map is only used for creating the dataset.

GeoChat (Original): *In the image, there are two groups of buildings. The first group is located at the left side of the image, while the second group is located at the right side of the image. The buildings in the first group are close to each other, while the buildings in the second group are also close to each other. There are also some trees present in the image, with one tree located at the bottom right corner of the image.*

GeoChat (w/ ChatEarthNet): *In this image, the developed areas are scattered throughout, with a notable concentration in the central part of the image, forming irregular shapes that suggest a mix of urban and suburban layouts. The crop areas dominate the landscape, with expansive patches occupying most parts of the image, especially in the top left and bottom right sections. Small patches of grass are interspersed among the other land cover types, while the water bodies are sparse, appearing as thin, snaking lines that suggest the presence of rivers or streams. The main theme of this image is a predominantly agricultural landscape with significant urban development.*

Label: *In this image, the grass areas are predominant, covering a significant portion of the landscape. The developed areas have a notable presence, with an intricate network of linear and polygonal shapes indicative of urbanization. Crop areas are considerable as well, with a patchwork of rectangular and irregular shapes scattered throughout. Water bodies are visible, with sinuous and branching forms that suggest rivers or streams. The trees occupy the least area, with small, scattered clusters. Overall, the image presents a landscape dominated by grass and agricultural activities, with developed areas and water bodies interspersed, giving a sense of a mixed-use region where urban and rural elements coexist.*

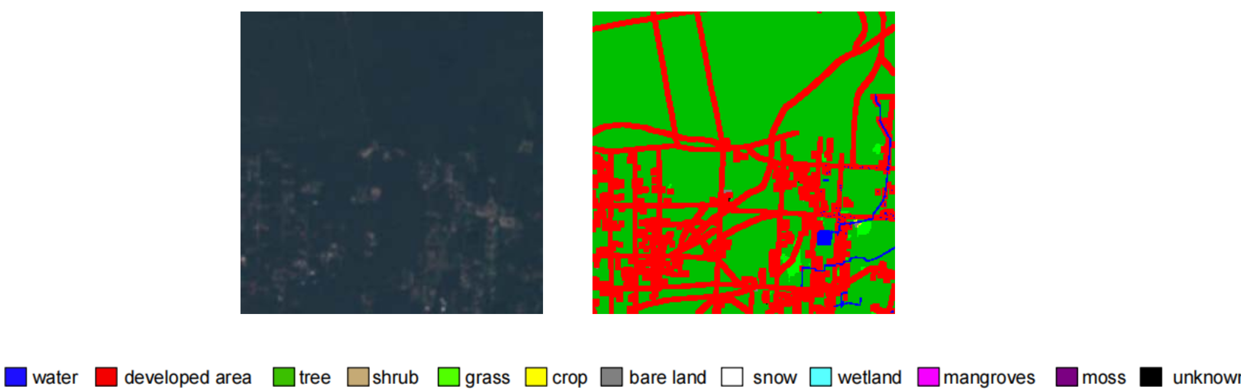

Fig. 2. Sentinel-2 satellite image and its associated land cover map. Note that during the training and evaluation process, only the sentinel-2 satellite image is used as visual input. The land cover map is only used for creating the dataset.

GeoChat (Original): *The image depicts a cityscape with a mix of buildings and trees. In the center of the image, there are two buildings close to each other. At the bottom left, there are two more buildings, and at the bottom right, there is a single building. The trees are scattered throughout the image, with some located at the bottom left, bottom right, and top right. The buildings and trees create a visually interesting and diverse scene.*

GeoChat (w/ ChatEarthNet): *In this image, the developed areas are predominant, with a dense network occupying a significant portion of the landscape, interspersed with patches of trees that cover varying sizes of the terrain. Small, scattered patches of grass are visible throughout, while the water bodies are minimal and irregularly shaped, mostly located near the developed areas. The main theme of this image is a landscape heavily influenced by development, with trees and grass providing some natural contrast.*

Label: *In this image, there is a notable dominance of tree cover in the top left, top right, and particularly in the middle part, suggesting a strong presence of forests or wooded areas. Developed regions are found throughout the image but are more concentrated in the bottom left and bottom right parts, forming patchy and fragmented patterns. A small amount of water is present, mostly noticeable in the bottom right as possibly a river or stream. Sparse patches of grass are almost imperceptible, and the bare ground is present in very small quantities. The image's landscape is primarily characterized by vast treed areas with urban development interspersed, and minimal water bodies.*

2. At the same time, the new results only aim at answering my comment 4 in the first round of review (Is the dataset useful for evaluating RS VLMs?). However, it does not deal with comment 3 (is the dataset useful for learing good RS VLM representations?). Although the authors aim at answering this question in the last two lines of Table 2, it is impossible to judge if the model is overfitting to the specific task when evaluating on the same tasks (even if on different splits). I suggest performing zero- or few-shot experiments using established RS benchmarks (like EuroSAT, BigEarthNet, etc.) in order to compare the performance of each model before and after fine-tuning with the proposed dataset.

R: We thank the reviewer for the valuable suggestion and provide our responses below.

1) **Following the reviewer's suggestion, we conduct additional experiments on scene classification datasets in a zero-shot setting. Specifically, we evaluate models on the widely used UCMerced [8] and AID [9] scene classification datasets. The reason we chose these two datasets is to maintain consistency with GeoChat's [1] experimental setup. Since GeoChat also uses these datasets for scene classification testing, we can directly compare our results with GeoChat's original results on the same datasets. The performance results are as follows:**

Models	**UCMerced**	**AID**
Qwen-VL	62.90	52.60
MiniGPTv2	4.76	12.90
LLaVA-1.5	68.00	51.00
GeoChat	84.43	72.03
GeoChat (w/ ChatEarthNet)	**89.29**	**77.57**

The results demonstrate that after training on the proposed ChatEarthNet dataset, the model retains its ability to perform zero-shot scene classification tasks with strong instruction-following capabilities. This highlights two key points:

- **Model capacity: RS VLMs can effectively learn from diverse datasets and demonstrate their robustness and adaptability.**
- **No overfitting: Fine-tuning with ChatEarthNet does not lead to overfitting but instead enhances the model's representation learning, making it more capable in both detailed captioning and zero-shot scene classification tasks.**

2) **Regarding the concern about overfitting, we acknowledge that fine-tuning the model on a specific dataset can sometimes lead to catastrophic forgetting, a common challenge in machine learning. However, this issue is more close to model design and optimization strategies rather than the dataset itself. Our work focuses on introducing ChatEarthNet as a global-scale image-text dataset, not on developing novel VLM training methodologies. Nevertheless, our experiments show that incorporating ChatEarthNet into GeoChat's training improves performance across different tasks.**

   In summary, we would like to emphasize again that ChatEarthNet contributes to the remote sensing community by:

   - **Providing a framework for creating the large-scale and high-quality image-text dataset on remote sensing data.**
   - **Enhancing VLMs by improving their performance when fine-tuned, without causing overfitting.**

We would also like to mention that the primary aim of our manuscript aligns with the objectives of Journal *Earth System Science Data (ESSD)*: to demonstrate the process of constructing a global-scale image-text dataset using tools like ChatGPT. And the manuscript we submitted is categorized as a "Data Description" paper. While optimal training methods for RS VLMs are indeed an important topic, they fall outside the scope of this work.

We appreciate the reviewer for the deep-thought comments, and we hope these additional experiments and discussions provide a more comprehensive response to your concerns. Thank you again for your valuable suggestions.

[1] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2024.
[2] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. BeLongie, J. Luo, M. Datcu, M. Pelillo, and L. P. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018.
[3] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[4] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu, M. Weinmann, S. Hinz, C. Wang, and K. Fu, "Fair1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022.
[5] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
[6] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post-flood scene understanding," *IEEE Access*, 2021.
[7] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, 2017.
[8] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010.
[9] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

Reviewer #3:

The authors answered the reviewer's questions, supplemented the experiments and made improvements to the manuscript. The reviewer recommends acceptance, but there are still two minor issues:

R: Thank you for recommending our manuscript for acceptance. We are grateful for your acknowledgment of the efforts we made to address your comments. We have carefully considered the two minor issues you raised and provide detailed responses below.

1. In Table A1, the authors are encouraged to unify the digital format in the "#Image-text pairs" field.

R: We appreciate the reviewer's suggestions on this point. We have updated Table A1 to ensure a unified digital format. The revised Table A1 is presented below.

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~~ 5,000,000	Coarse-grained	Model-generated & multiple datasets	RGB, multiple datasets	Global
SkyScript (Wang et al., 2024)	~~2.6 M~~ 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

2. In Table 2, what are the specific differences between the model's original training datasets and the proposed test dataset that lead to a large gap in the metrics?

R: Thank you for raising this insightful question. The performance gap between GeoChat's [1] original training datasets and our proposed ChatEarthNet test dataset arises due to significant domain differences. Below, we provide a detailed comparison of these domain gaps and explain why they lead to the observed performance discrepancies. Additionally, we include examples of predicted results before and after training on ChatEarthNet to better illustrate these differences. Please kindly refer to Comment#1 of Reviewer#1 for more details.

There is indeed a great gap between the datasets for training GeoChat and ours. GeoChat's training datasets include six datasets, summarized below:

- **DOTA [2]: A dataset specifically designed for object detection in remote sensing images, with a focus on high-resolution spatial data and object categories such as ships, tennis courts, and small vehicles.**

- **DIOR [3]: Another object detection dataset with categories such as vehicle, stadium, and wind mill.**
- **FAIR1M [4]: Also an object detection dataset featuring high-resolution remote sensing imagery, providing object categories such as ship, road, and court.**
- **LRBEN [5]: A visual question answering (VQA) dataset in remote sensing, primarily addressing urban-rural classification, presence of elements (e.g., roads and buildings), and simple quantitative or comparative questions. It lacks comprehensive land use land cover (LULC) analysis.**
- **FloodNet [6]: A VQA dataset focusing on flood-related categories like flooded and non-flooded buildings or roads, with a significant domain gap from our dataset.**
- **NWPU-RESISC45 [7]: A classification dataset covering diverse scene types with varying spatial resolutions, such as bridge, church, and intersection.**

We summarize the following three major differences that explain the large performance gap:

- **Objective differences: The GeoChat's training datasets are designed for tasks such as object detection, scene classification, or simple VQA, whereas ChatEarthNet emphasizes detailed LULC-related analysis and descriptions.**
- **Spatial resolution: The above-mentioned datasets predominantly feature high-resolution imagery focusing on individual objects, while ChatEarthNet provides imagery with global-scale, medium-resolution data suited for holistic LULC tasks.**
- **Content and scope: None of GeoChat's training datasets contain the comprehensive, detailed LULC-related image-text pairs that ChatEarthNet offers.**

The gaps identified above demonstrate the need for ChatEarthNet, which is specifically designed to address these limitations, providing global-scale LULC-related image-text data that complements existing datasets. This enables more effective training and evaluation for applications requiring detailed LULC analysis.

We appreciate the reviewer for the deep-thought comments for improving our manuscript. We hope these additional discussions provide a comprehensive response to address your concerns. Thank you again for your valuable suggestions.

[1] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," ***Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*****, 2024.**
[2] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. BeLongie, J. Luo, M. Datcu, M. Pelillo, and L. P. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," ***Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*****, 2018.**
[3] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," ***IEEE Transactions on Geoscience and Remote Sensing*****, 2022.**
[4] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu, M. Weinmann, S. Hinz, C. Wang, and K. Fu, "Fair1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," ***ISPRS Journal of Photogrammetry and Remote Sensing*****, 2022.**

[5] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," ***IEEE Transactions on Geoscience and Remote Sensing*****, 2020.**
[6] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post-flood scene understanding," ***IEEE Access*****, 2021.**
[7] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," ***Proceedings of the IEEE*****, 2017.**