SenCLIP: Enhancing zero-shot land-use mapping for Sentinel-2 with ground-level prompting

Pallavi Jain[1, 2, 6], Dino Ienco[2, 3, 5, 6], Roberto Interdonato[2, 4, 5, 6],
Tristan Berchoux[1] and Diego Marcos[2, 6]
[1]Mediterranean Agronomic Institute of Montpellier - CIHEAM-IAMM, [2] Inria, [3] INRAE,
[4] Cirad, [5] UMR TETIS, [6] Univ. of Montpellier, Montpellier France
{pallavi.jain, dino.ienco, roberto.interdonato, diego.marcos}@inria.fr
berchoux@iamm.fr

A. Appendices

A.1. Dataset Overview

The LUCAS dataset [2], encompassing data from 2006, 2009, 2012, 2015, and 2018, includes high-resolution images captured at 1600×1200 pixels. For this work, we focused on the 2018 dataset and downsampled these images to 512×512 pixels using the LANCZOS [1] interpolation method. This technique was selected for its superior resampling quality, ensuring the preservation of image detail and clarity. Figure 1 illustrates the geographical distribution of geo-tags across Europe, while Figure 2 showcases examples of the four directional LUCAS images alongside their corresponding Sentinel-2 images obtained from the Planetary API.

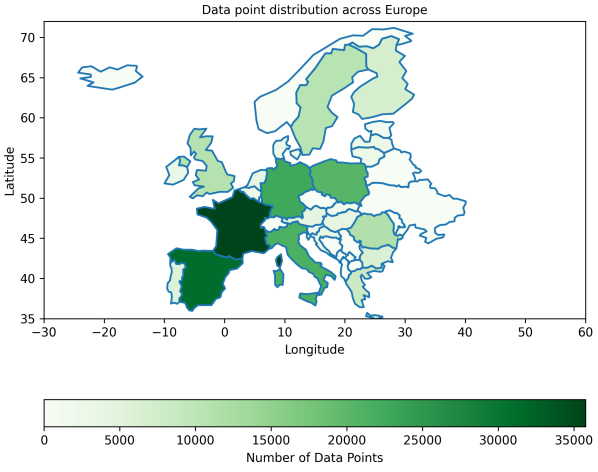

Figure 1. Data Distribution across Europe

A.2. Meta Prompts and Prompt Example

In this work, we generated ground and aerial prompts using the meta-prompt approach described in "*Meta Prompt for Ground View Prompts*", and further refined them based on different views, such as "aerial" shown in "*Meta Prompt for Aerial View Prompts*", as well as prompt length. An example of the generated prompts for the "Forest" class is shown in Fig 3.

Meta Prompt for Ground View Prompts

Generate 50 extremely short and diverse sentences that may correspond to factual visual descriptions of photos taken over the land-use/land-cover class 'Annual Crop' such that they are as different as possible from all of these other classes:
['Industrial', 'Pasture', 'River', 'Forest', 'Herbaceous Vegetation', 'Permanent Crop', 'Highway', 'Residential', 'Sea Lake']
Try to describe visual features or objects that are likely to be visible in such images, even if they are not stereotypical. Make sure they cover as many of all the possible random photos that could be taken over that land-use/land-cover and that they sound as objective as possible, covering different seasons and states of the land-use/land-cover. Make sure to add some examples related to the class for the image visual description. Do not make poetic sentences but more factual.

Figure 2. Sentinel-2 collected images from Geo-Tagged LUCAS data points. LUCAS images include four directional views, which are displayed alongside the Sentinel-2 imagery with 10m resolution.

Aerial	"Satellite image shows a network of narrow trails winding through thick woodland." "Aerial view reveals a mix of coniferous and deciduous trees, creating a varied texture." "Bare patches within the forest might indicate recent logging or natural clearings."
Aerial Long	"From above, the forest appears as a vast expanse of lush greenery, with the tree canopies forming a complex mosaic of different shades, interspersed with narrow, shadowy lines indicating trails or streams.", "The overhead perspective reveals a thick forest, its canopy a rich tapestry of varied hues of green, punctuated by the occasional, brighter patches of wildflowers in bloom."
Aerial Short	"Dense canopy of trees covering a large area.", "Varying shades of green indicating different tree species."
Ground	"The ground is covered with a thick layer of forest undergrowth and ferns." "The forest features a mix of old-growth trees and younger saplings." "Mushrooms and fungi are visible growing on fallen logs and tree stumps."
Ground Long	"A family of deer grazes peacefully amidst the trees, their ears twitching as they listen for signs of danger while they browse on tender shoots of grass and leaves.", "A pack of wolves hunts for prey under the cover of darkness, their keen senses and stealthy movements allowing them to move through the forest with ease."
Ground Short	"A fallen tree trunk provides a home for mosses and fungi.", "A hiker pauses to admire the view from a forest overlook."

Figure 3. Examples of different style prompts generated for *"Forest"* class by GPT3.5 [4]

Meta Prompt for Aerial View Prompts

Generate 50 extremely short and diverse sentences that may correspond to factual visual descriptions of aerial or satellite view over the land-use/land-cover class 'Annual Crop' such that there are as different as possible from all of these other classes:
['Annual Crop', 'Industrial', 'Pasture', 'River', 'Forest', 'Herbaceous Vegetation', 'Permanent Crop', 'Highway', 'Residential', 'Sea Lake']
Try to describe aerial or satellite visual features or objects that are likely to be visible in such images, even if they are not stereotypical. Add aerial view context with patterns and use "aerial", "satellite photo" terms. Make sure they cover as many of all the possible random photos that could be taken over that land-use/land-cover and that they sound as objective as possible, covering different seasons and states of the land-use/land-cover. Make sure to add some examples related to the class for the image aerial or satellite visual attributes. Do not make poetic sentences but more factual.

A.3. Zeroshot Results based on Length of Prompts

In addition to generating prompts from aerial and ground perspectives, we further diversified the prompt styles by incorporating varying lengths: short sentences (10 words) and long sentences (50 words), tailored to the specific class under consideration, as illustrated in Fig 3. Table 1 reveals several key insights regarding the impact of prompt length and view-type on zero-shot classification performance.

Prompt Templates/Models		Aerial Short	Aerial Long	Ground Short	Ground Long
EuroSAT					
RN50	CLIP [5]	31.95	**38.42**	28.56	**31.32**
	RemoteCLIP [3]	24.82	**27.80**	**23.84**	20.90
	SenCLIP-AvgPool	**57.74**	*57.02*	**59.10**	55.46
	SenCLIP-AttPool	***58.68***	56.34	**60.12**	55.92
ViT-B/32	CLIP [5]	45.09	**49.61**	40.97	**41.76**
	RemoteCLIP [3]	37.94	**38.85**	**37.54**	35.30
	SkyCLIP [6]	**61.05**	59.20	53.25	**54.03**
	GeoRSCLIP [7]	62.46	**62.91**	**60.00**	57.88
	SenCLIP-AvgPool	63.78	**66.89**	**64.28**	58.80
	SenCLIP-AttPool	*64.10*	***67.54***	**63.04**	57.82
BigEarthNet					
RN50	CLIP [5]	27.60	**30.02**	**24.41**	23.78
	RemoteCLIP [3]	**32.60**	32.14	**31.74**	30.66
	SenCLIP-AvgPool	33.57	**33.60**	30.02	**32.70**
	SenCLIP-AttPool	*35.18*	***35.89***	32.74	**35.09**
ViT-B/32	CLIP [5]	28.58	**34.12**	27.51	**28.94**
	RemoteCLIP [3]	**32.75**	28.38	**29.97**	25.44
	SkyCLIP [6]	25.77	**28.08**	**23.43**	21.55
	GeoRSCLIP* [7]	*37.24*	***39.05***	30.95	**33.75**
	SenCLIP-AvgPool	33.08	**35.66**	*34.36*	***34.57***
	SenCLIP-AttPool	33.67	**33.76**	33.80	**33.95**

Table 1. Zero-shot classification results with RN50 and ViT-B/32 backbones on EuroSAT and BigEarthNet datasets, highlighting the effectiveness of various prompt lengths and types. The comparison includes specific class prompts with short (10-word) and long (50-word) sentence descriptions for aerial and ground views, all generated using GPT-3.5. *Note: GeoRSCLIP [7], trained on BigEarthNet with paired text, is considered supervised rather than zero-shot. **Bold** indicates the each model's performance on short versus long prompts, while ***italic*** highlights the overall best-performing model across short and long.

Effect of Prompt Length: Across both the EuroSAT and BigEarthNet datasets, longer prompts (50 words) generally outperform shorter ones (10 words), particularly for aerial views. This trend holds across all models, which

show improved accuracy when provided with more detailed prompts. For instance, SenCLIP exhibits notable accuracy improvements with longer prompts on both datasets, especially in aerial views, with the exception of RN50 on EuroSAT.

Ground Views and Prompt Length: In contrast to aerial views, the effect of prompt length is less pronounced for ground-level images. Models like RemoteCLIP [3] and SenCLIP often perform equally well or slightly better with shorter prompts compared to longer ones. This is likely due to the rich visual context inherent in ground-level images, where detailed descriptions in longer prompts may add limited value. For example, in the EuroSAT dataset, SenCLIP-AvgPool and SenCLIP-AttPool show minimal gains from longer prompts in ground views, suggesting that prompt specificity may matter more than length for ground-level imagery.

Moreover, SenCLIP variants consistently outperform other models in both aerial and ground views across both datasets, demonstrating the robustness of its cross-view strategies in leveraging detailed descriptions. While GeoRSCLIP [7], a supervised model, benefits from longer prompts (particularly in BigEarthNet), it is consistently outperformed by SenCLIP in ground-view scenarios. For example, SenCLIP achieves significant performance gains on ground-level imagery within the BigEarthNet dataset.

References

[1] Claude E Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology and Climatology*, 18(8):1016–1022, 1979. 1

[2] Raphaël d'Andrimont, Momchil Yordanov, Laura Martinez-Sanchez, Beatrice Eiselt, Alessandra Palmieri, Paolo Dominici, Javier Gallego, Hannes Isaak Reuter, Christian Joebges, Guido Lemoine, et al. Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union. *Scientific data*, 7(1):352, 2020. 1

[3] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023. 3, 4

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[6] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5805–5813, 2024. 3

[7] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023. 3, 4