ALIGNING GEO-TAGGED CLIP REPRESENTATIONS AND SATELLITE IMAGERY FOR FEW-SHOT LAND USE CLASSIFICATION

Pallavi Jain[a,b], *Diego Marcos*[b], *Dino Ienco*[b,c], *Roberto Interdonato*[b,d], *Aayush Dhakal*[e], *Nathan Jacobs*[e] and *Tristan Berchoux*[a]

[a] Mediterranean Agronomic Institute of Montpellier - CIHEAM-IAMM, Montpellier, France
[b] Inria, Univ. of Montpellier, Montpellier, France
[c] INRAE, UMR TETIS, Univ. of Montpellier, Montpellier France
[d] Cirad, UMR TETIS, Univ. of Montpellier, Montpellier France
[e] Washington University in St. Louis, USA

ABSTRACT

A major difference between ground-level and satellite imagery of landscapes lies in their semantic granularity: ground-level images tend to offer details on objects and human activities, while satellite images provide broader geographic context but, typically, with coarser semantics. This study aims to leverage this complementary information by integrating fine-grained insights from a ground-level view into the analysis of satellite image data. To achieve this integration, we propose to align a satellite image representation with co-located geo-tagged ground-level image CLIP representations. This method focuses on enriching satellite image visual features by leveraging the inherent visual characteristics found in ground-level images as a reference in a contrastive manner, without relying on additional textual information to guide the learning process. We evaluate the quality of the learned representations on the EuroSAT benchmark in various few-shot settings.

Index Terms— computer vision, satellite images, land use, contrastive learning

1. INTRODUCTION

Characterising land use patterns remains a critical task, enabling comprehensive assessments of geographical landscapes and their evolving features. These analyses provide invaluable insights into the utilisation, transformation, and management of land, particularly in rural areas. Exploiting remote sensing data, like satellite and aerial imagery, has the potential of furnishing crucial information for analysing and interpreting land use patterns at a continental scale.

Approaches based on deep learning have been extensively used to extract useful information from the vasts amounts of remote sensing data that are being produced daily. However, conventional methodologies often rely on extensive labelled datasets, limiting adaptability to new and diverse scenarios. Recent advancements in self-supervised learning, particularly contrastive approaches, offer opportunities to learn useful representations without extensive labelled data. This is done by leveraging natural co-occurrence patterns within massive, unlabelled, datasets. For instance, DINO [1] uses patches stemming from the same image, while CLIP (Contrastive Language-Image Pre-training) [2] leverages textual captions associated to online images, thus exploiting a multimodal co-occurrence.

In practice, the CLIP framework comprises of text and image encoders. The text encoder processes the textual descriptions to generate embeddings that represent the semantic meaning of the text. Simultaneously, the image encoder processes images to produce embeddings that encapsulate visual features present in the images. These encoders work in tandem to establish a joint embedding space where text and image embeddings share similarities. This allows CLIP framework to learn the similarity between any text-image pair, resulting in high similarity only when the text adequately describes the image.

This enable CLIP as a tool to explore possible interpretations of a given image via natural language. However, its focus on natural images found online presents a challenge when dealing with the specificities of remote sensing data, limiting its effectiveness in remote sensing data analysis and contextual understanding of land use patterns. This is not only due to differences with respect to the sensors used, but also with the kind of views captured by space- and air-borne cameras; remote sensing imagery often provides coarse details, with much lower spatial resolution than typical ground-level photos, about objects that may help infer useful characteristics of a landscape, such as land use, that are not adequately captured within the CLIP embedding space [3].

One way of improving the CLIP representation of remote

This research was supported by the ‘Giving Rural Actors Novel Data and Re-Usable Tools to Lead Public Action in Rural Areas’ (GRANULAR) project, which has received funding from the European Union’s Horizon Europe Research and Innovation Programme under Grant Agreement No. 101061068.

sensing images, as proposed in RemoteCLIP [4] is to fine-tune the CLIP image encoder on a set of captioned remote sensing images, in line with CLIP's text-image similarities. This framework targets specialised feature learning designed for remote sensing. However, its dependence on labelled data persist, since it requires a curated set of text-image pairs.

In order to eliminate the need for curated labels, another option, as proposed in [3], is to add a new modality of freely available pair data, in the form of pairs of remote sensing images and co-located geo-tagged photos. The latter already provides rich semantic information via the pre-trained CLIP image encoder, meaning than only one additional model needs to be trained in order to encode the remote sensing images. In our work, we thus leverage cross-view geo-localised images to capitalise on the potential of ground-level images as superior feature descriptors to remote sensing images. Cross-view approaches have been extensively employed in various studies to comprehend image similarity, localisation, and orientation [5, 6, 7].

Our methodology involves leveraging CLIP's knowledge acquired from natural images to extract frozen ground-level feature embeddings. We then fine-tune another CLIP image encoder using Bing/Sentinel images, adapting CLIP's embedding space for remote sensing data by incorporating ground images. Hence, this work harnesses the advantage of CLIP's comprehension of detailed ground-level features to improve its understanding of remote sensing data, creating a mutually beneficial relationship between the two domains.

We train the remote sensing image encoder using the geo-tagged ground level images of the LUCAS project [8] paired with two different satellite modalities: (i) very high resolution images provided by Bing Maps and (ii) Sentinel-2 images. We evaluate the learned representations on EuroSAT, a land use / land cover classification benchmark.

2. METHOD

The proposed framework links ground-level and satellite image data through their spatial coordinates, as illustrated in Figure 1.This allows to align satellite-derived data with the manifold of the original CLIP embeddings, thus allowing to interact with satellite data using textual descriptions from a ground-level perspective.

We initiate the process by extracting ground-level image embeddings using the frozen CLIP encoder. We then fine-tune a separate encoder using satellite imagery, aiming to align its representation with the co-located ground-level CLIP representation.

2.1. Dataset

Ground-level image data utilised in this study were sourced from the LUCAS dataset [8], a comprehensive rural survey dataset encompassing Land Use and Land Cover information across Europe. This dataset spans multiple years, including data from 2006, 2009, 2012, 2015, and 2018. Specifically, our work focuses on leveraging the LUCAS 2018 dataset, which comprises approximately 235,000 geo-tagged locations. Each location is associated with four directional images (north, east, west, south), resulting in an aggregate of about 900,000 images.

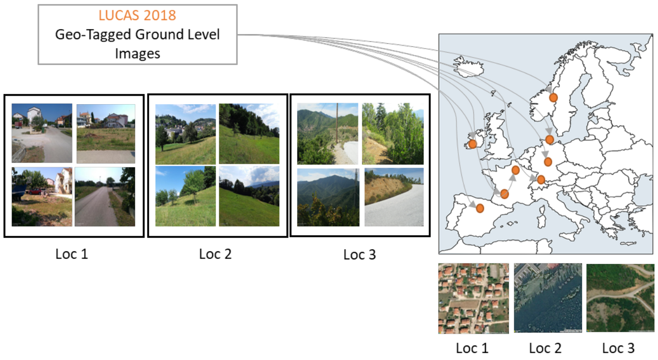

Fig. 1. Bing and Sentinel images collected across Europe from the same locations as geo-tagged LUCAS photos

.

Bing and Sentinel-2 images were collected for their unique strengths: Bing for high-resolution details in specific regions, and Sentinel-2 for broader coverage despite lower resolution, ensuring a fair comparison in data utilisation for training. The images were obtained to cover approximately 1 sqkm around geographical location from the LUCAS 2018 dataset. For Bing aerial data, we utilised Bing Maps API with specific parameters: a zoom level of 18, and 500 x 500 pixels image size.

In a similar fashion, Sentinel-2 data is accessed using the Planetary Computer API [9]. The data retrieval strategy focused on obtaining imagery corresponding to specific months and years from the LUCAS dataset. Cloud coverage filtering was implemented to select images featuring less than 10-20% cloud cover. The acquired Sentinel-2 data consisted of RGB bands, offering imagery at a resolution of 10 meters per pixel and dimensions of 100 x 100 pixels per scene.

2.2. Approach

We use pairs of ground-level image quadruplets and satellite images, denoted as $\{(\mathbf{Y}_1, \mathbf{x}_1), (\mathbf{Y}_2, \mathbf{x}_2), \ldots, (\mathbf{Y}_N, \mathbf{x}_N)\}$. With $\mathbf{Y}_i = \{\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \mathbf{y}_{i,3}, \mathbf{y}_{i,4}\}$ the quadruplet of ground level images corresponding to the i^{th} location. The frozen embeddings are obtained from ground-level images using the pre-trained CLIP encoder, $\mathbf{g}_{i,k} = f_G(\mathbf{y}_{i,k})$. Simultaneously, the satellite image encoder $\mathbf{s}_i = f_s(\mathbf{x}_i)$ is initialised with the original CLIP image encoder and undergoes fine-tuning with Bing and Sentinel data, resulting in modified models referred to as BingCLIP and SenCLIP respectively.

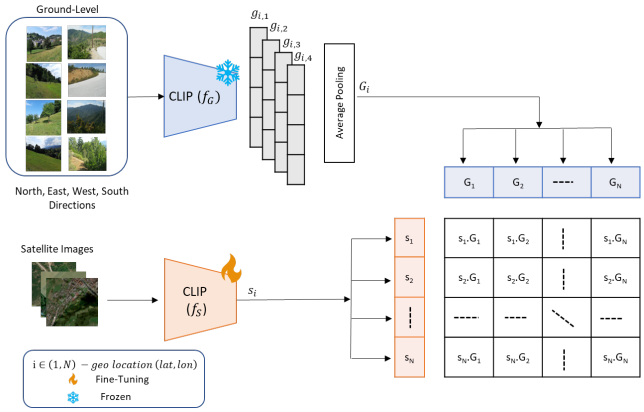

Fig. 2. Architecture: Frozen CLIP encoder (f_G) extracts ground-level embeddings, while a separate CLIP encoder (f_s) is fine-tuned using Sentinel/Bing images. Comparing these embeddings aligns CLIP features with satellite images while maintaining proximity to ground-level context.

For each location, the frozen embeddings correspond to four LUCAS directional images. To consolidate these into a single embedding $\mathbf{G}_i$ per location, represented by a set of quadruplet embeddings $\{\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \mathbf{g}_{i,3}, \mathbf{g}_{i,4}\}$, an average pooling operation is performed as follows:

$$\mathbf{G}_i = \frac{1}{4}\sum_{k=1}^{4}(\mathbf{g}_{i,k}). \tag{1}$$

To assess the alignment between the frozen ground-level embeddings and the fine-tuned CLIP embeddings for satellite images, we employed the InfoNCE loss. This loss function, derived from Noise Contrastive Estimation (NCE) principles, facilitated the evaluation of both the similarity and dissimilarity between these two sets of embeddings.

The InfoNCE loss can be expressed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\text{sim}(\mathbf{G}_i, \mathbf{s}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{G}_i, \mathbf{s}_j)/\tau))}\right) \tag{2}$$

Here N is the number of data samples, where $\mathbf{G}_i$ and $\mathbf{s}_i$ represent the frozen ground-level embeddings and the learnable satellite image embeddings, respectively, pertaining to the i^{th} geographic location. In addition, $\text{sim}(\mathbf{G}, \mathbf{s})$ denotes the cosine similarity function between these two embeddings, and τ denotes a temperature parameter that scales the similarity scores.

The objective of this loss function is twofold: it aims to increase the cross-modal similarity between embeddings from the same geographic location while diminishing the similarity between embeddings from different locations. The ultimate goal is to ensure alignment between the fine-tuned CLIP embeddings and the distinctive characteristics inherent in the ground-level embeddings. This alignment enhances the model's capacity to discern and extract features specific to satellite imagery, thereby enhancing the accuracy of land use classification.

In addition to the InfoNCE loss, we also explored using a cosine similarity loss, that utilises only positive pair of ground-level and satellite image embeddings. The cosine similarity loss function between individual ground-level images and their corresponding positive pairs can be represented as:

$$\mathcal{L}_{cossim} = \sum_{i=1}^{N}\frac{\mathbf{G}_i \cdot \mathbf{s}_i}{\|\mathbf{G}_i\|\|\mathbf{s}_i\|} \tag{3}$$

3. EXPERIMENTS AND RESULTS

3.1. Pre-Training

This study involved the fine-tuning of a ResNet50 architecture used in the CLIP framework for BingCLIP and SenCLIP. To optimise the model's parameters, we employed the AdamW optimiser, as proposed by Loshchilov and Hutter [10], setting the initial learning rate (LR) to $5 \cdot 10^{-6}$. The training procedure employed a batch size of 32 and extended across 100 epochs, incorporating a cosine annealing warm-start scheduler to enhance the training process. τ is set to 0.007 to scale the similarity scores. To transform the images, a series of techniques including resizing, centre cropping, flipping, and rotation were applied. These transformations were implemented to ensure diverse and comprehensive augmentation of the image data for improved model training.

In addition to fine-tuning the CLIP model, we conducted separate experiments where we trained models from scratch using Bing and Sentinel-2 data. This approach allowed for a fair and comparative analysis to discern the benefits derived from fine-tuning the CLIP model. By training models from scratch, we aimed to understand the differences in performance, feature learning capabilities, and overall effectiveness between fine-tuned CLIP and models developed from the ground up.The hyperparameter settings remain same in both fine-tune and scratch settings. All models were trained on single NVIDIA Titan X GPU.

3.2. Evaluation

For the evaluation process, we opted to utilise the EuroSAT benchmark dataset [11] due to its incorporation of Sentinel-2 images and Land Use/Cover classes, aligning well with our primary objectives. The evaluation is done on 5000 instances of EuroSAT that comprises two approaches: first, splitting the data into 80:10:10 for train, validation, and test datasets, and second, utilising few-shot learning, wherein 1, 2, 5, 10, 15, and 20 instances per class were provided to assess the model's performance with limited instances. We present results through two distinct methodologies: Linear probing, which involves assessing the models' frozen weights and utilising a single linear layer as the classification layer, and fine-tuning the whole last residual block and last linear layer of ResNet50. The utilisation of linear probing offers a more robust evaluation methodology, allowing for a comprehensive understanding of the efficacy of the learned representations. This approach aids in gauging the efficiency of the learned features and their applicability to the classification task.

Models	Linear Probing	Fine-Tune
Supervised Scratch ResNet50	96.42 *	NA
CLIP	87.60	95.73
RemoteCLIP	92.53	96.67
Bing Scratch	94.53	95.33
BingCLIP-InfoNCE	97.33	98.27
BingCLIP-CosSim	97.60	**98.53**
Sentinel Scratch	95.47	96.40
SenCLIP-InfoNCE	**98.13**	97.73
SenCLIP-CosSim	97.47	97.60

Table 1. Top-1 accuracy on EuroSAT dataset. (* represent scratch training on EuroSAT instead of Linear Probing result).

For the fair evaluation comparing the advantages of transferring the ground-level CLIP representation to satellite data, we trained a ResNet50 from scratch on EuroSAT. As illustrated in Table 1, both Bing-based and Sentinel-based models showcased superior performance in the classification tasks via linear probing and fine-tuning, outperforming the supervised, original CLIP, and Remote CLIP methods. However, the precise impact of the loss function remains a point to investigate. In the case of the Sentinel-based InfoNCE model, it surpassed others in linear probing, yet experienced a slight decline of 0.04% when fine-tuning the last block. Conversely, fine-tuning BingCLIP with cosine similarity loss demonstrated superior performance across all models. Both SenCLIP and BingCLIP consistently delivered improved overall performance in comparison to text or label-based training. In fact, models trained from scratch using Bing and Sentinel data exhibited better performance in comparison to CLIP and RemoteCLIP. Nonetheless, the performance of these models remains slightly lower in contrast to the supervised method. This comparison underscores the potential and efficacy of transferring ground-level representations to satellite in enhancing feature learning for image classification tasks, offering competitive performance conversely to traditional supervised approaches.

In the evaluation conducted to assess our models' capabilities using few-shot learning, Table 2 presents the linear probing performance across varying shot values, namely 1, 2, 5, 10, 15, and 20. Notably, the results underscore the remarkable performance of the SenCLIP models in handling few-shot learning scenarios. Across the spectrum of shots, SenCLIP models consistently showcased exceptional accuracy, demonstrating their adeptness in learning from limited instances.

Models/K-Shot	1	2	5	10	15	20
CLIP	44.50	54.00	66.50	71.50	68.00	75.00
RemoteCLIP	51.00	65.50	72.00	74.00	79.00	81.50
Bing Scratch	59.00	61.00	79.00	81.00	80.50	85.00
BingCLIP-InfoNCE	56.00	68.50	84.00	87.50	87.00	89.50
BingCLIP-CosSim	64.50	69.00	83.00	86.00	89.00	90.50
Sentinel Scratch	63.50	71.00	83.00	88.50	87.00	90.50
SenCLIP-InfoNCE	60.50	**72.00**	82.50	88.50	**91.50**	91.50
SenCLIP-CosSim	**66.00**	71.00	**84.50**	**91.00**	90.50	**93.50**

Table 2. K-Shot Linear Probing Top-1 Accuracy on EuroSAT

One prominent observation from the evaluation is the significant performance disparity between BingCLIP/SenCLIP models and CLIP or RemoteCLIP. This gap highlights the superior feature learning capacity of the former models, emphasising their proficiency in comprehending labels even with minimal instances provided per class.

Moreover, the outcomes obtained from the models trained from scratch also surpass the performance of CLIP and RemoteCLIP. This finding further reinforces the notion that ground-level images serve as good alternative descriptors compared to textual information, that contributes to the learning of effective representations.

In the comparison between Bing and Sentinel models, the superior performance of the Sentinel-based models aligns with expectations, considering that the EuroSAT dataset primarily comprises Sentinel images. This outcome highlights the advantages of using images from a similar source during model training, leading to improved recognition and classification of data within that specific dataset.

4. CONCLUSION

In this research, we investigated the potential of using geotagged ground-level imagery to improve satellite image feature learning. By fine-tuning CLIP-based models on Sentinel-2 and Bing imagery, using ground-level photos from the LUCAS European land use project, our study aimed to bridge the gap between detailed ground-level context and broader, less-detailed satellite data. Our results demonstrated that models trained with ground-level imagery as descriptors outperformed other pre-training strategies, such as CLIP, in image classification. The comparison revealed significant advantages, particularly with limited labeled and Sentinel-2 data. Furthermore, our few-shot learning evaluation highlighted the adaptability and accuracy of our approach, even with minimal instances per class. These findings suggest the promising use of ground-level imagery to enhance satellite image analysis.

5. REFERENCES

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, “Emerging properties in self-supervised vision

transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[2] 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*. PMLR, 2021, pp. 8748–8763.

[3] Aayush Dhakal, Adeel Ahmad, Subash Khanal, Srikumar Sastry, and Nathan Jacobs, "Sat2cap: Mapping fine-grained textual descriptions from satellite images," *arXiv preprint arXiv:2307.15904*, 2023.

[4] 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.

[5] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5007–5015.

[6] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.

[7] Yujiao Shi and Hongdong Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17010–17020.

[8] 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*, vol. 7, no. 1, pp. 352, 2020.

[9] Microsoft Open Source, Matt McFarland, Rob Emanuele, Dan Morris, and Tom Augspurger, "microsoft/planetarycomputer: October 2022," oct 2022.

[10] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," 2018.

[11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.