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

Abstract

Pre-trained vision-language models (VLMs), such as CLIP, demonstrate impressive zero-shot classification capabilities with free-form prompts and even show some generalization in specialized domains. However, their performance on satellite imagery is limited due to the underrepresentation of such data in their training sets, which predominantly consist of ground-level images. Existing prompting techniques for satellite imagery are often restricted to generic phrases like "a satellite image of...", limiting their effectiveness for zero-shot land-use/land-cover (LULC) mapping. To address these challenges, we introduce SenCLIP, which transfers CLIP's representation to Sentinel-2 imagery by leveraging a large dataset of Sentinel-2 images paired with geotagged ground-level photos from across Europe. We evaluate SenCLIP alongside other state-of-the-art remote sensing VLMs on zero-shot LULC mapping tasks using the EuroSAT and BigEarthNet datasets with both aerial and ground-level prompting styles. Our approach, which aligns ground-level representations with satellite imagery, demonstrates significant improvements in classification accuracy across both prompt styles, opening new possibilities for applying free-form textual descriptions in zero-shot LULC mapping. Code, dataset and pretrained models are available at https://github.com/pallavijain-pj/SenCLIP

1. Introduction

Monitoring land-use and land-cover (LULC) is essential for understanding human impacts on the environment and assessing related risks [40]. Spaceborne sensors have long been used to map LULC, providing key data for policy-relevant indicators, especially in rural areas [2]. These insights are crucial for sustainable development, land-use planning, habitat preservation, and effective natural resource management [6, 8]. During the last decade, the evolution of deep learning has enhanced the reliability of LULC prediction based on remote sensing [49]. However, these approaches generally depend on large datasets for training the models via supervised learning, requiring a large initial effort and restricting them to a closed set of initial LULC classes. In general, the class labels influence the learning process by guiding the optimisation of the network's parameters towards minimising the classification error for the known classes [10, 13]. Consequently, the network may prioritise learning features that are highly discriminative for these specific set of classes. This limits the usefulness of the representation in recognising unseen classes without additional training or fine-tuning steps.

Recent advancements in vision-language models (VLMs) have revolutionised this paradigm. VLMs leverage Web-scale image/caption datasets to learn a representation that allows for zero-shot predictions across various tasks. These models operate by learning a joint semantic space for image/text pairs through contrastive learning strategies. Prominent examples of VLMs include CLIP [28], ALIGN [12], and BLIP [16, 17], which aim to map both textual descriptions and visual representations into the same latent space, enabling direct semantic comparisons between the two modalities. Despite the significant strides made by VLMs, there are persistent challenges in the practical implementation of zero-shot approaches based on them. One of the obstacles lies in the manual prompting process intrinsic to these approaches, where the choice of vocabulary and specific textual cues for each class play a pivotal role. Even slight variations in prompt formulation, such as the inclusion or exclusion of articles like *"a"* can wield considerable influence over the model's accuracy [46]. Moreover, achieving optimal performance often requires incorporating task-relevant information into the prompts. In

specialised domains like remote sensing, specifying details such as *"centered satellite photo"* or describing specific attributes like *"photo of {class}: a type of broadleaf forest"* alongside the class name becomes imperative for accurate classification [1, 28, 46]. This nuanced prompt construction ensures that the model captures essential contextual cues, leading to improved classification accuracy [43], even if the added context is meaningless [30]. The sensitivity of prompt engineering and prompt learning [45] to a specific context underscores the critical role of prompting for harnessing the full potential of VLMs.

Another significant challenge arises when attempting to move into specialised domains that are underrepresented in the training sets of VLMs, as is the case in remote sensing tasks. This results in challenges both for the image representation, which suffers due to the differences to the more common image modalities in terms of resolution, perspective and radiometry, and the text representation. Indeed, the CLIP representation of satellite images tends to be aligned with coarse concepts (e.g. *this is a satellite image*) that are of little help for understanding LULC [4]. One potential solution lies in the development of large-scale VLMs specifically tailored for remote sensing applications. However, a significant obstacle arises due to the scarcity of textual descriptions associated with remote sensing data. Unlike other domains with abundant image/text pairs, remote sensing lacks a substantial corpus of annotated satellite imagery with corresponding textual descriptions.

Recent advancements like RemoteCLIP [21], SkyCLIP [41], and GeoRSCLIP [44] have successfully integrated remote sensing data with captions and class labels to enhance VLMs. These methods use curated satellite imagery, improving the models' ability to understand spatial contexts and semantic relationships through supervised pretraining, leading to better performance. In contrast, label-free approaches such as Sat2Cap [4] focus on cross-view learning by integrating geotagged ground-level images with high-resolution satellite imagery. This enables the model to associate local scene details with satellite observations, allowing it to interpret ground-level prompts more effectively. By transferring detailed ground-level knowledge to broader satellite contexts, these approaches improve land-use and land-cover classification accuracy.

In this work, we extend the Sat2Cap approach by leveraging the LUCAS dataset, which contains nearly one million geotagged images across the European Union, alongside Sentinel-2 medium-resolution imagery. Our key contributions are:

1) Alignment of Sentinel-2 representations with ground-level images. We advance zero-shot LULC classification by directly aligning Sentinel-2 imagery with co-located ground-level images from LUCAS, which allows the model to understand the satellite imagery via textual descriptions associated with ground-level concepts, as done in Sat2Cap [4]. This is done without relying on labels or captions, in contrast to competing approaches [21, 41, 44]. Unlike in Sat2Cap [4], which focuses on higher-resolution satellite images, our approach bridges the visual representation gap between ground-level perspectives and Sentinel-2's medium-resolution (10 m) multi-spectral data. This is particularly valuable given Sentinel-2's widespread use in environmental and agricultural applications, where capturing fine-grained details remains a challenge.

2) New state-of-the-art in zero-shot LULC classification. By coupling our SenCLIP model with a rich, LLM-generated, set of class-specific prompts, we obtain substantial improvements in terms of zero-shot LULC classification performance on both EuroSAT and BigEarthNet, validating the usefulness of the approach for real-world zero-shot LULC mapping. Additionally, we introduce a **simple prompt selection method**, aiming to optimise performance by identifying the most representative prompts for each class to enhance classification performance.

2. Related work

Zero-shot land use and land cover mapping: The large-scale study of LULC with spaceborne sensors [29] has taken large strides in terms of performance thanks to the application of deep learning-based methods [49]. Although determining some aspects of land cover, such as detecting water, evergreen forests, or built-up areas, is often done at an operational level at large scales [3, 7], many aspects of land use are hard to solve employing a satellite perspective only and require the use of ground-level information [37]. This is even more so the case in a zero-shot setting, when no class labels are available during training [42]. Indeed, the majority of previous work on zero-shot classification in remote sensing focuses on aerial or very high resolution ($<$1 m) satellite imagery, where minute details can be interpreted by humans or pretrained computer vision models [15, 18, 19, 38], with most methods for medium resolution satellite imagery ($\approx$ 10 m) focusing on the more relaxed few-shot setting, where a few training samples are available [31, 32]. In this paper, we leverage the fine-grained, easily detectable, details in geotagged ground-level images to enable zero-shot LULC classification using medium resolution satellite imagery.

Prompt tuning: Manual prompt engineering for zero-shot tasks with VLMs often prioritises linguistic nuances over visual cues, potentially limiting accuracy. To address this, large language models (LLMs) are increasingly used to improve prompting methods for better accuracy and robustness [23, 27]. Context Optimisation (CoOp) methods [45, 46] take a different route by learning non-textual prompts from training data, although they struggle with fine-grained tasks [43]. Recent approaches, like incorpo-

rating visual cues into prompts [43], or adding noise to enhance robustness [30], have also improved VLM performance. In our work, we adapt standard, LLM-based, and ground-level perspective prompting for remote sensing and close the visual representation gap through cross-view learning with geotagged images.

Cross-view learning: Geo-localised ground-level photos offer a promising avenue to leverage the descriptive capabilities of ground-level vision models, often surpassing those using remote sensing data. Cross-view methodologies provide insights into image similarity, localization, and orientation [20, 34, 35]. VIGOR [48] employs contrastive learning to compare features between aerial- and street-view images, enhancing scene analysis from diverse viewpoints. Similarly, TransGeo [47] uses attention-guided non-uniform cropping to enrich aerial-view features with ground-level details. However, bridging the gap between ground-level and satellite features remains challenging, particularly with low-resolution data. To address this, Sat2Cap [4] introduces a cross-view modeling framework that predicts CLIP embeddings for ground-level scenes using overhead imagery. Sat2Cap focuses on retrieval tasks and does not address the challenges of LULC classification or the handling of lower-resolution satellite images. Our work extends the Sat2Cap approach by targeting medium-resolution Sentinel-2 imagery and rural LULC tasks, achieving higher precision. While Sat2Cap captures satellite image features effectively, we enhance this by integrating four directional ground-level images per location from the LUCAS dataset. This cross-view integration enriches semantic context, leading to more accurate Sentinel-2 image representations and a better understanding of both ground and overhead modalities.

3. Method

Our method makes use of the rich and text-aligned representation provided by CLIP [28] on ground-level images and transfers them to a satellite image representation via geotagged photos, similarly to [4]. As shown in Fig. 1, we use two separate CLIP image encoders: one is kept frozen and provides frozen ground-level embeddings, while the other is fine-tuned for satellite data. The ground-level embeddings serve as targets, guiding the satellite image embedding to learn the semantic space that enables the alignment of satellite-derived data with the manifold of original CLIP embeddings from the ground-level perspective.

3.1. Self-supervised training dataset

The ground-level images used in this study were obtained from the LUCAS dataset [5], a comprehensive rural survey dataset providing Land Use and Land Cover information across Europe. The 2018 LUCAS survey consists of approximately 235,000 geotagged locations. Each location is associated with four directional images (taken from the north, east, west, and south), resulting in a total of around 900,000 images. This large-scale dataset offers a rich source of information for analyzing land use and cover patterns.

Using the LUCAS geolocations, we accessed Sentinel-2 data via the Planetary Computer API [36]. Our data retrieval focused on acquiring imagery from specific months and years corresponding to the LUCAS dataset. To ensure data quality, we applied a cloud coverage filter, selecting images with less than 10-20% cloud cover. The obtained Sentinel-2 data included RGB bands with a resolution of 10 meters per pixel and scene dimensions of 100×100 pixels. More details on the dataset and its distribution across Europe are provided in supplementary.

3.2. Ground-level representation of satellite images

For the two CLIP image encoders (frozen and trainable), pairs of ground-level images ($\mathbf{Y}$) and geolocated satellite images ($\mathbf{x}$) are utilised, denoted as $\{(\mathbf{Y}_1, \mathbf{x}_1), (\mathbf{Y}_2, \mathbf{x}_2), \ldots, (\mathbf{Y}_N, \mathbf{x}_N)\}$. Here, $\mathbf{Y}_i = \{\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \ldots, \mathbf{y}_{i,K}\}$ represents the set 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, denoted as f_G, such that ground-level image embeddings are computed as: $\mathbf{g}_{i,k} = f_G(\mathbf{y}_{i,k})$. Simultaneously, the satellite image encoder f_S is initialised with the original CLIP image encoder and undergoes fine-tuning with Sentinel data. This results in our fine-tuned models, which we refer to as SenCLIP when combined with the projection head defined below.

Pooling For each location, the frozen embeddings correspond to a set of ground-level images. To consolidate these into a single embedding $\mathbf{G}_i$ per location, represented by a set of quadruplet embeddings $\{\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \ldots, \mathbf{g}_{i,K}\}$, we explore two different pooling methods: average and attention pooling. Average pooling (AvgPool) is a simple and efficient approach that assigns equal importance to all four directional images. The embedding $\mathbf{G}_i$ is defined as follows:

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

Attention pooling (AttPool) allows the model to focus on the most informative features from each location. The embedding $\mathbf{G}_i$ is obtained as follows:

$$\mathbf{G}_i = \sum_{k=1}^{K} (w_{i,k} \cdot \mathbf{g}_{i,k}), \tag{2}$$

where $w_{i,k}(\mathbf{g}_{i,k})$ represents the trainable attention weight for the i-th location and the k-th image, parameterised using a fully connected neural network.

Projection head The projection head H, adapted from the implementation in [33], transforms the embeddings into a

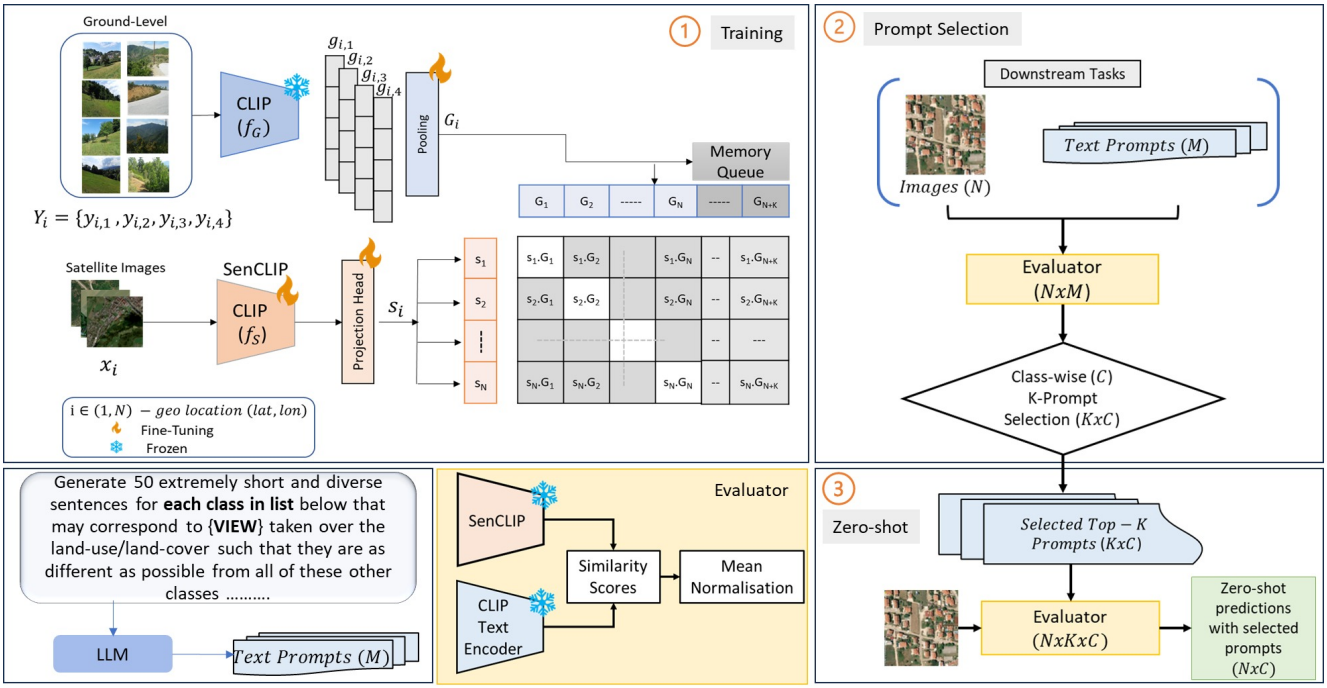


Figure 1. Architecture: The figure illustrates the three-step pipeline consisting of Pre-Training, Prompt Selection, and Zero-shot Predictions. It also demonstrates the prompt generation process from LLMs, which is utilised for prompt selection and then selected prompts for zero-shot prediction.

new space to capture richer relationships. It consists of a two-layer linear neural network, with GELU activation, dropout regularisation, and layer normalisation.

A residual connection is created by adding the output of the first linear layer to the output of the dropout layer. The projection head is trained alongside the rest of the model. When combined with the trainable CLIP image encoder, the final model produces an embedding $\mathbf{s}$ for the satellite image $\mathbf{x}$, as shown below:

$$\mathbf{s_i} = \text{SenCLIP}(\mathbf{x_i}) = H(f_S(\mathbf{x_i})). \tag{3}$$

Training During training, two components are trained: the image encoder and a projection head for satellite images. Additionally, a pooling head is incorporated into the frozen encoder to consolidate ground-level quadruplets. The training process uses a dictionary implemented as a queue, inspired by the MoCo framework [9], to manage pooled frozen ground-level embeddings. This mechanism allows the reuse of encoded keys from previous iterations, with samples gradually replaced and the oldest batch removed to maintain consistency with newer samples. The training objective is to optimize the InfoNCE (Information Noise Contrastive Estimation) [25] contrastive loss function between the pooled embeddings $\mathbf{G}_i$ from ground-level quadruplets and the fine-tuned satellite embeddings $\mathbf{s}_i$. This loss function evaluates both the similarity and dissimilarity between these two sets of embeddings, offering valuable insights into the effectiveness of the model's training process.

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\mathbf{G}_i \cdot \mathbf{s}_i/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{G}_i \cdot \mathbf{s}_j/\tau))}\right) \tag{4}$$

where τ is the temperature and N, the number of samples.

3.3. Prompting and zero-shot inference

Class-specific view-dependent prompts. We used an LLM to generate view-specific prompts encompassing aerial/overhead and ground-level views, as depicted in Fig.1. Examples of generated prompts shown in supplementary. Based on the training described above, we expect SenCLIP to perform well on prompts describing LULC classes from a ground-level perspective. We generated a fixed number of prompts, T, for each class $c \in [1, \ldots, C]$.

Prompt-based zero-shot classification. The generated prompts, represented as vectors $\mathbf{a}_{c,t}$, each prompt can be considered to be an attribute of its corresponding class c. The similarity between a satellite image embedding $\mathbf{s}_i$ and the full set of text embeddings is determined through a dot product. In order to obtain the class scores, we employ Direct Attribute Prediction (DAP) [14]. The similarity between $\mathbf{s_i}$ and each attribute vector $\mathbf{a}_{c,t}$, calculated as $\mathbf{s}_i \cdot \mathbf{a}_{c,t}$, serves as a proxy for $p(\mathbf{a}_{c,t}|\mathbf{s}_i)$ i.e. probability of class at-

tribute given the satellite image embedding. The final classification assignment is then computed as

$$c_i = \arg\max_c \prod_{t=1}^{T} \frac{p(\mathbf{a}_{c,t}|\mathbf{s}_i)}{p(\mathbf{a}_{c,t})}, \tag{5}$$

where $p(\mathbf{a}_{c,t})$ is empirically estimated as the mean similarity of $\mathbf{a}_{c,t}$ with the full image set.

Prompt selection method. This process involves computing a goodness score for each prompt $\mathbf{a}_{c,t}$, based on its similarity to the rest of the prompts. These scores are then used to identify the top prompts for each class.

Specifically, we use a weighted mean score to measure how well each prompt represents its corresponding class. Prompts that are more representative of their class characteristics are assigned higher weighted scores. This is achieved by calculating the mean within class similarity

$$\alpha_{c,t} = \frac{\sum_{q=1}^{T} \mathbf{a}_{c,t} \cdot \mathbf{a}_{c,q}}{T} \tag{6}$$

for each class, and comparing it to the overall mean similarity scores

$$\beta_{c,t} = \frac{\sum_{d=1}^{C} \sum_{q=1}^{T} \mathbf{a}_{c,t} \cdot \mathbf{a}_{d,q}}{C \cdot T}. \tag{7}$$

The ratio $w_{c,t} = \frac{\alpha_{c,t}}{\beta_{c,t}}$ effectively evaluates the representativeness of each prompt, emphasizing those with higher relevance to specific class characteristics. Higher weighted scores indicate prompts that more effectively encapsulate the essence of their class, serving as strong indicators of class-specific attributes. This approach aligns with the objective of prioritizing prompts that are more indicative and representative of their classes, while diminishing the influence of less relevant prompts.

4. Experiments and results

This section outlines the training implementation and hyperparameters, followed by the quantitative and qualitative results for SenCLIP. Quantitative results cover zero-shot inference, the impact of prompt selection, and performance improvements with LaFTer [23]. Qualitative analysis using image captioning and cross-view image retrieval highlights the quality of the learned representations.

4.1. Implementation details

We fine-tuned SenCLIP using two backbone CLIP encoders: ResNet50 (RN50) and ViT-B/32. The fine-tuning strategy varied based on the model architecture: for RN50, all layers were fine-tuned, whereas for ViT-B/32, we fine-tuned only the last transformer block, linear layer, and the projection head. The AdamW optimizer, proposed by [22], was used with initial learning rates (LR) of 10^{-5} for RN50 and 10^{-4} for ViT-B/32. Training was conducted over 20 epochs with a batch size of 32, incorporating a step scheduler with a step size of 5 and a decay multiplier of 0.95. The temperature parameter τ was set to 0.07 to scale the similarity scores. Data augmentation techniques included resizing, center cropping to 32×32, random flipping, and rotation. For generating the prompts, we used GPT-3.5 [26] as LLM. The models were trained on a single NVIDIA Titan X GPU.

Model Arch	Prompt Templates/Models	Generic	Aerial	Ground
EuroSAT				
RN50	CLIP	40.55	47.64	32.28
	RemoteCLIP	25.20	29.95	22.13
	SenCLIP-AvgPool	53.89	57.54	56.71
	SenCLIP-AttPool	**56.53**	**57.78**	**57.95**
ViT-B/32	CLIP	47.26	54.87	51.66
	RemoteCLIP	44.74	48.95	43.21
	SkyCLIP50	55.66	66.04	59.98
	GeoRSCLIP	**63.40**	68.02	65.82
	SenCLIP-AvgPool	61.18	**71.22**	65.54
	SenCLIP-AttPool	62.24	70.78	**66.91**
BigEarthNet				
RN50	CLIP	27.71	29.77	24.09
	RemoteCLIP	23.04	33.00	20.23
	SenCLIP-AvgPool	32.74	32.41	34.39
	SenCLIP-AttPool	**34.61**	**34.88**	**34.80**
ViT-B/32	GeoRSCLIP*	41.95	37.36	32.10
	CLIP	29.80	29.50	28.37
	RemoteCLIP	27.17	26.87	27.76
	SkyCLIP50	20.16	29.87	20.21
	SenCLIP-AvgPool	**34.72**	**36.78**	**37.40**
	SenCLIP-AttPool	33.78	35.29	37.07

Table 1. Zero-shot performance comparison on EuroSAT and BigEarthNet using RN50 and ViT-B/32 backbones, highlighting the effect of unified and class-specific prompt strategies. Prompts include a generic format, '*centered satellite photo of* {*class*}', alongside various GPT-3.5-generated aerial and ground view descriptions. *Note: GeoRSCLIP, trained on BigEarthNet with paired text, is considered supervised rather than zero-shot.

4.2. Quantitative results

To evaluate the effectiveness of the learned representations, we utilised two well-established Sentinel-2 benchmark datasets: EuroSAT [11] and BigEarthNet [39]. EuroSAT contains images with 10 distinct, single-class land use/cover categories. On the other hand, BigEarthNet offers a more extensive set of annotations, with 19 multi-label classes. Consequently, Top-1 accuracy was used as the evaluation metric for EuroSAT, while mean average precision (mAP) was employed for BigEarthNet. All evaluations were performed in a zero-shot setting, where the models were not exposed to the training sets of the benchmarks. Classification was carried out by comparing image features with class-associated text features.

Zero-shot inference. Table 1 summarises the zero-shot classification performance of various models using differ-

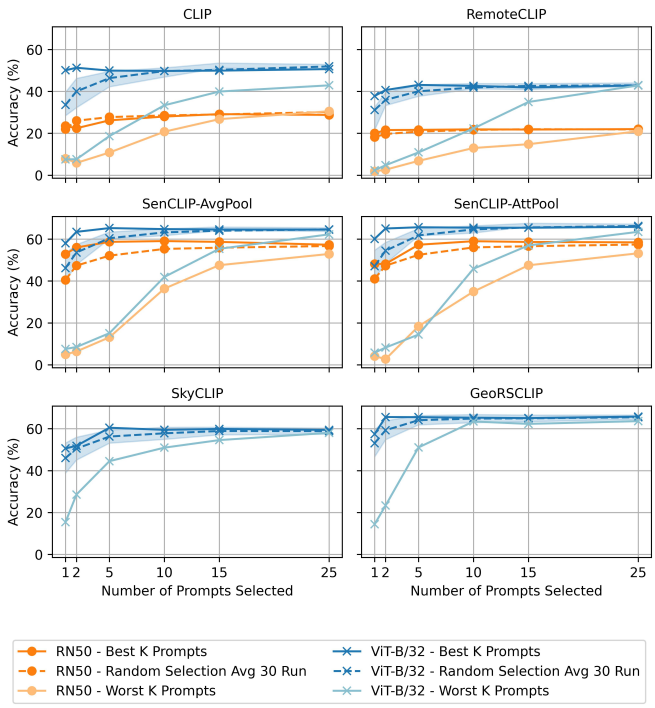


Figure 2. Effect of prompt selection strategies on model (RN50 and ViT-B/32 Backbone) performance on the EuroSAT dataset, varying the number of used prompts. It compares our prompt selection method, which ranks prompts from the most to least descriptive prompts (labelled as Best K and Worst K) for each class, against a random prompt selection baseline. For the random selection baseline we also show the standard deviation over 30 trials.

ent types of prompts. Models were tested with both generic prompts (e.g., "a centered satellite photo of class") and class-specific prompts tailored to aerial and ground-level perspectives. Each class was represented by 50 prompts for each perspective. The comparison includes baseline models like CLIP, RemoteCLIP, SkyCLIP [41], GeoRSCLIP [44], alongside SenCLIP employing two pooling techniques: AvgPool and AttPool. SenCLIP consistently outperforms all other models, demonstrating its superior ability to leverage both aerial and ground-view prompts for enhanced classification accuracy on both dataset with both the aerial and ground-level prompts. Specifically, for the ViT-B/32 architecture, SenCLIP improves classification performance by over 10% for aerial prompts and ground-view prompts compared to CLIP and RemoteCLIP. While GeoRSCLIP performs competitively with generic satellite prompts, similar to those it has been trained on, its performance declines with more descriptive prompts. It achieves a notable accuracy on the BigEarthNet dataset. However, GeoRSCLIP is trained on BigEarthNet with captions, making it a supervised model rather than zero-shot, limiting its comparability to other models in this context. This is also reflected in

Model/Prompts	EuroSAT		BigEarthNet	
	Aerial	Ground	Aerial	Ground
CLIP	59.88 ± 6.08	48.33 ± 5.79	29.69 ± 5.47	29.17 ± 0.93
SenCLIP AvgPool	**77.90 ± 1.22**	72.87 ± 1.12	**34.67 ± 0.71**	**44.55 ± 1.52**
SenCLIP AttPool	73.58 ± 0.92	**75.92 ± 2.13**	34.20 ± 1.81	42.79 ± 4.14

Table 2. EuroSAT evaluation with LaFTer on top of CLIP and SenCLIP with ViT-B/32 backbone. LaFTer text classifier is trained for 400 epochs and fine-tuned for 20 and 5 epochs for CLIP and SenCLIP, respectively. We used aerial and ground-level prompts. Results are averaged over different seeds.

the fact that GeoRSCLIP underperforms SenCLIP by 5% in BigEarthNet when using ground-level prompts, which are further away from the generic ones used to train it. The class-specific ground-view prompts, along with the unified aerial prompts, showcases the SenCLIP's ability to capture relevant visual and semantic information, leading to superior performance compared to CLIP and remote sensing VLMs.

Prompt selection and zero-shot. Fig. 2 illustrates the impact of prompt selection strategies on model performance on the EuroSAT dataset with RN50 and ViT backbones. Particularly for ViT, the best results are obtained with our approach by selecting a few (2 or 5) prompts. For the RN50 backbone, prompt selection with CLIP and RemoteCLIP exhibit similar results to random prompt selection due to their lower effectiveness with ground-level prompts, resulting in varying performance between good and random prompts. The worst K-prompt selection, with worse-than-random performance when few prompts are selected, highlights the efficacy of the proposed prompt selection strategy.

Zero-shot classifier tuning (LaFTer [23]) on top of SenCLIP. Table 2 presents the integration of LaFTer with CLIP and SenCLIP models, revealing the influence of text classifier training on performance. The text classifier trained for the LaFTer default setting of 400 epochs. SenCLIP demonstrates superior performance when employing ground-view and aerial-view prompts (50 per class), for both EuroSAT and BigEarthNet. We found that SenCLIP achieves high level performance with fine-tuning for up to 5 epochs, while CLIP requires 20 epochs to achieve the best results without over-fitting. This could be attributed to SenCLIP being optimised specifically for remote sensing tasks, whereas CLIP has more generalised embeddings that require additional training to optimise.

4.3. Qualitative results on cross-modal retrieval

Qualitative analysis of SenCLIP embeddings using ClipCap [24] To further analyse the embeddings learned by SenCLIP, we conducted a qualitative analysis using the ClipCap image caption generator [24]. As shown in Fig. 3, the captions generated by SenCLIP provide detailed descriptions of the ground-level view. In contrast, captions

		SenCLIP AvgPool	SenCLIP AttPool	CLIP	RemoteCLIP	SkyCLIP	GeoRSCLIP
Annual Crop Land		the new crop of peppers.	the new crop of rice.	the colors of the flag.	the camera moves along the edge of the field.	blue and gold stars on a blue background.	the green field of the sunflower.
Forest		the man who posted this photo of himself in the woods.	the full text of the letter.	a close up of the sky.	the video shows the ship's silhouette as it passes through the water.	bottoms of a young woman.	a boat is floating on the water.
Herbaceous Vegetation Land		the rocks are being protected by a new plan to build a new road.	a map of the area.	the video shows the area around the site where the blast took place.	a close up of a rock.	turquoise and pink painted on a blue background.	aerial view of the forest.
Highway or Road		a sign is seen in front.	a new sign is seen on the side of a bus.	a car drives through a tunnel.	a close up of a car wash.	it's a cute idea to make a flower pot out of a hat.	the video shows the city's roads and buildings.
Industrial Buildings		a city is a city.	person, the new logo for the store.	the aerial view of the city.	a city is a small town.	tourist attraction is a popular place to get a photo op.	the city of the future.
Pasture		the farm is now in a state of disrepair.	the farmhouse was abandoned by the family when it was built.	the aircraft was flying over the field.	a photo of a water tank.	stars on a branch.	the greenhouses are located in the middle of a large field.
Permanent Crop Land		the garden of the future.	the fruit is still in the process of being harvested.	the video shows the area around the site where the explosion took place.	a photo of a house.	tis the season to be a star!.	the rice fields in the countryside.
Residential Buildings		this is a house i built.	the house of the artist.	a satellite image taken shows the village.	a close up of a dirty room.	it is a beautiful day with a lot of stars and the moon.	aerial view of the city.
River		the bridge is one of the most famous landmarks.	the bridge is one of the most iconic bridges in the world.	the bridge is seen from above.	the man's body was found in the river.	it's a boy's world, we all have our reasons.	the river flows through the city.
Sea or Lake		the beach is so beautiful.	a photo of the ship.	a cloud of water droplets.	a cloud in the water.	to the man in the garden.	a large black jellyfish floats in the ocean.

Figure 3. Image captioning on EuroSAT images using ClipCap [24]

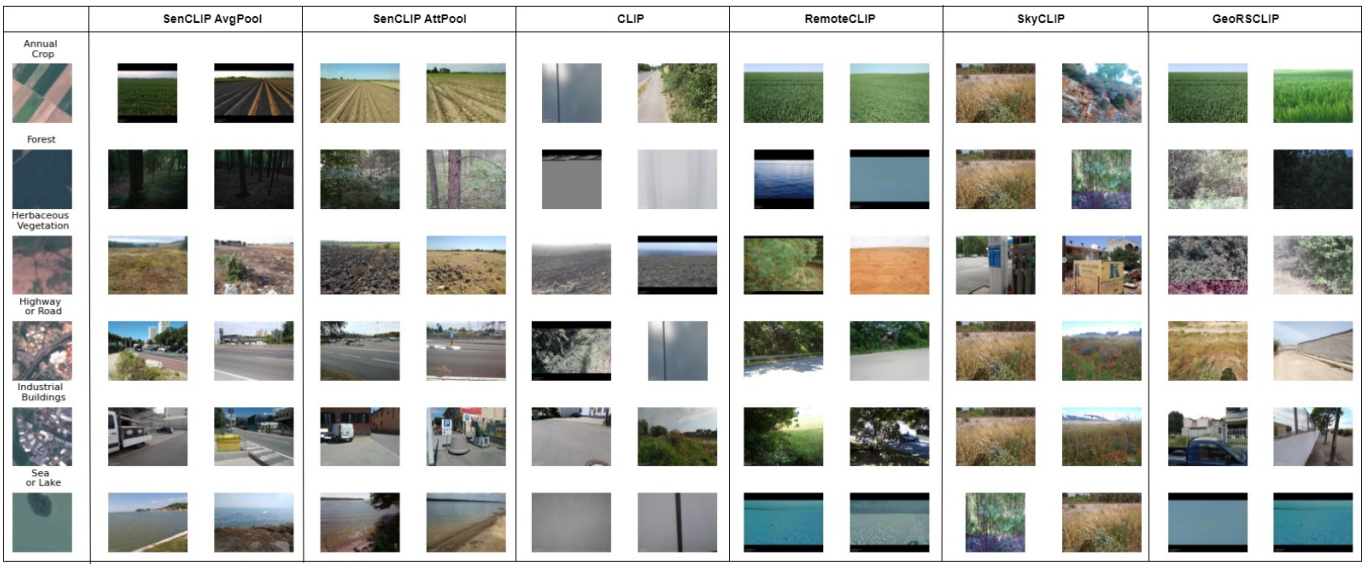
produced by CLIP and RemoteCLIP predominantly reflect an aerial perspective, often including terms such as *"aerial view"*, *"photo of"* and referencing perspectives from aircraft, video, and camera angles. This distinction underscores SenCLIP's superior ability to capture ground-level details, offering a more nuanced understanding of the scene compared to its counterparts.

Image-to-image retrieval - satellite to ground-level. We conducted an image-to-image retrieval experiment to identify ground-level images corresponding to a given satellite image. As depicted in Fig. 4a, we leveraged EuroSAT images from different classes to identify the two nearest neighbor ground-level LUCAS images. The results demonstrate SenCLIP's strong ability to accurately retrieve ground-level images that align with the given class, significantly outperforming CLIP and RemoteCLIP. The latter models struggled across several classes and frequently selected outlier LUCAS images, highlighting their limitations in establishing robust mappings between satellite and ground-level views. Notably, SenCLIP's fine-grained feature alignment is especially apparent in the *annual crops* class, where the retrieved ground-level images showcase detailed views of plantations.

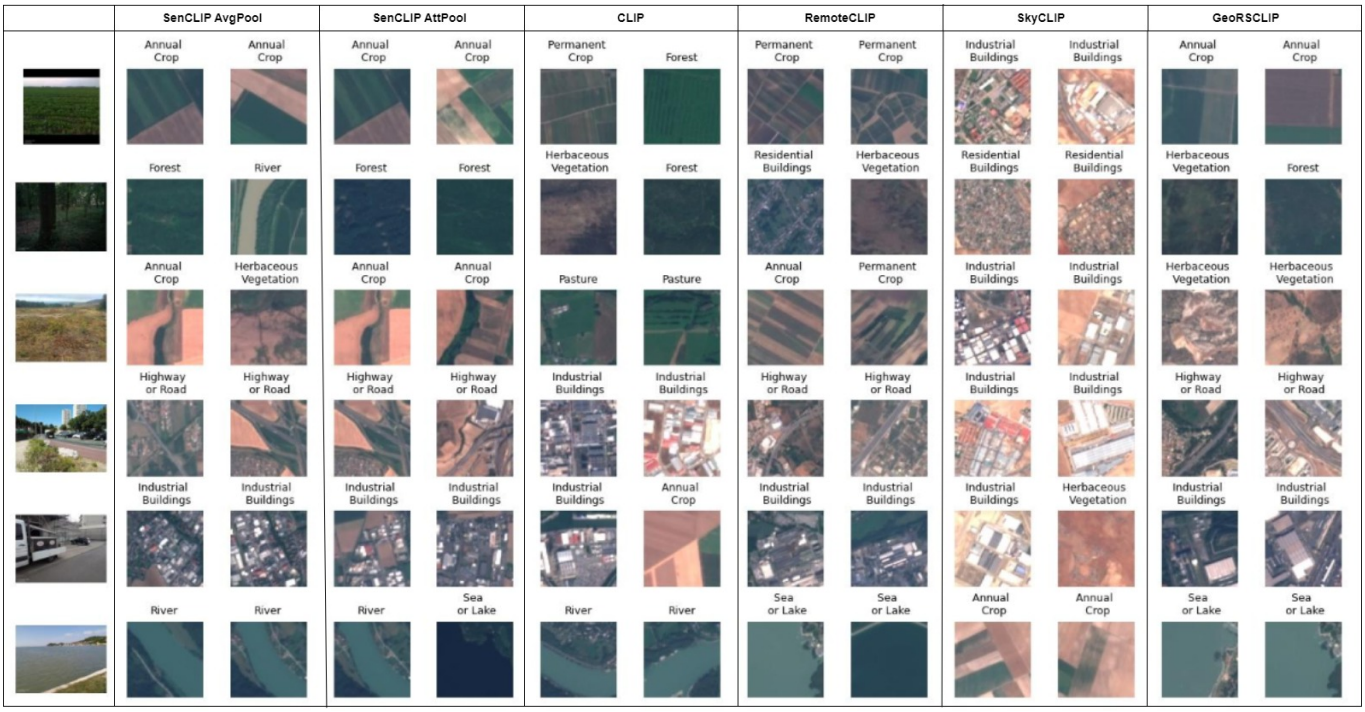
Image-to-image retrieval - ground-level to satellite. In our exploration of image-to-image retrieval, we investigated the process of mapping LUCAS ground-level images to EuroSAT satellite images, as shown in Fig. 4b. For this analysis, we used the top-2 LUCAS images given an example from EuroSAT, as depicted in Fig. 4a. The results demonstrate that SenCLIP excels in accurately identifying the correct classes in the majority of cases. However, it is worth noting that CLIP and RemoteCLIP perform better in this scenario than in EuroSAT to LUCAS.

5. Conclusion

This study highlights the benefits of cross-view fine-tuning of CLIP, enabling a remote sensing model to effectively capture ground-level semantic details using medium-resolution Sentinel-2 images. Unlike conventional models, which often struggle with domain-specific terms and predefined class names, our approach demonstrates remarkable flexibility in accommodating diverse prompting styles for zero-shot LULC classification. This flexibility paves the way for the creation of custom LULC maps without requiring any additional training data. The model's success can be attributed to its comprehensive self-supervised training, which aligns Sentinel-2 representations with CLIP representations of co-located, geotagged, ground-level images from the European Union-wide LUCAS dataset. Furthermore, we introduce an efficient prompt selection method, highlighting the importance of prompt curation. Overall, this work combines cross-view



(a) **EuroSAT to LUCAS**



(b) **LUCAS To EuroSAT**

Figure 4. **Qualitative image-to-image retrieval**: This analysis demonstrates the qualitative effectiveness of SenCLIP embeddings in both directions. By identifying the top-2 nearest LUCAS embeddings from EuroSAT images, the results indicate that the model successfully learns the fine-grained relationships between ground-level and satellite imagery. Conversely, using ground-level LUCAS images to find the top-2 nearest neighbor satellite images from EuroSAT, the analysis further demonstrates the model's capability to map embeddings bidirectionally, effectively capturing and relating fine-grained details between the two image domains.

training and prompt selection to empower models like SenCLIP, enabling them to surpass the limitations of traditional remote sensing methods. By incorporating ground-level landscape descriptions, SenCLIP sets a new benchmark for zero-shot LULC mapping.

Acknowledgements

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.

References

[1] James Urquhart Allingham, Jie Ren, Michael W Dusenberry, Xiuye Gu, Yin Cui, Dustin Tran, Jeremiah Zhe Liu, and Balaji Lakshminarayanan. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *International Conference on Machine Learning*, pages 547–568. PMLR, 2023. 2

[2] James Richard Anderson. *A land use and land cover classification system for use with remote sensor data*, volume 964. US Government Printing Office, 1976. 1

[3] György Büttner. Corine land cover and land cover change products. In *Land use and land cover mapping in Europe: practices & trends*, pages 55–74. Springer, 2014. 2

[4] 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. 2, 3

[5] 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. 3

[6] Gregory Giuliani, Paolo Mazzetti, Mattia Santoro, Stefano Nativi, Joost Van Bemmelen, Guido Colangeli, and Anthony Lehmann. Knowledge generation using satellite earth observations to support sustainable development goals (sdg): A use case on land degradation. *International Journal of Applied Earth Observation and Geoinformation*, 88:102068, 2020. 1

[7] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853, 2013. 2

[8] Peter K. Hargreaves and Gary R. Watmough. Satellite earth observation to support sustainable rural development. *International Journal of Applied Earth Observation and Geoinformation*, 103:102466, 2021. 1

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[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*, 12(7):2217–2226, 2019. 5

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[14] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009. 4

[15] Aoxue Li, Zhiwu Lu, Liwei Wang, Tao Xiang, and Ji-Rong Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):4157–4167, 2017. 2

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[18] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023. 2

[19] Yansheng Li, Deyu Kong, Yongjun Zhang, Yihua Tan, and Ling Chen. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179:145–158, 2021. 2

[20] 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*, pages 5007–5015, 2015. 3

[21] 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. 2

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[23] M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Mateusz Kozinski, Horst Possegger, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5, 6

[24] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 6, 7

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[26] 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. 5

[27] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 2

[28] 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. 1, 2, 3

[29] John Rogan and DongMei Chen. Remote sensing technology for mapping and monitoring land-cover and land-use change. *Progress in planning*, 61(4):301–325, 2004. 2

[30] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. *arXiv preprint arXiv:2306.07282*, 2023. 2, 3

[31] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 200–201, 2020. 2

[32] Marc Rußwurm, Sherrie Wang, and Devis Tuia. Humans are poor few-shot classifiers for sentinel-2 land cover. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 4859–4862. IEEE, 2022. 2

[33] M. Moein Shariatnia. Simple CLIP, 4 2021. 3

[34] 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*, pages 17010–17020, 2022. 3

[35] 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*, pages 4064–4072, 2020. 3

[36] Microsoft Open Source, Matt McFarland, Rob Emanuele, Dan Morris, and Tom Augspurger. microsoft/planetarycomputer: October 2022, oct 2022. 3

[37] Shivangi Srivastava, John E Vargas Munoz, Sylvain Lobry, and Devis Tuia. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science*, 34(6):1117–1136, 2020. 2

[38] Gencer Sumbul, Ramazan Gokberk Cinbis, and Selim Aksoy. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):770–779, 2017. 2

[39] Gencer Sumbul, Jian Kang, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, and Begüm Demir. Bigearthnet dataset with a new class-nomenclature for remote sensing image understanding. *arXiv preprint arXiv:2001.06372*, 2020. 5

[40] Billie L Turner, Eric F Lambin, and Anette Reenberg. The emergence of land change science for global environmental change and sustainability. *Proceedings of the National Academy of Sciences*, 104(52):20666–20671, 2007. 1

[41] 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. 2, 6

[42] Meiliu Wu, Qunying Huang, Song Gao, and Zhou Zhang. Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. *International Journal of Applied Earth Observation and Geoinformation*, 125:103591, 2023. 2

[43] Yi Zhang, Ce Zhang, Ke Yu, Yushun Tang, and Zhihai He. Concept-guided prompt learning for generalization in vision-language models. *arXiv preprint arXiv:2401.07457*, 2024. 2, 3

[44] 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. 2, 6

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2

[47] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 3

[48] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 3

[49] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 1, 2