SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing

Zhecheng Wang, Rajanie Prabha*, Tianyuan Huang*, Jiajun Wu, Ram Rajagopal

Stanford University
{zhecheng, rajanie, tianyuah, ramr}@stanford.edu, jiajunwu@cs.stanford.edu

Abstract

Remote sensing imagery, despite its broad applications in helping achieve Sustainable Development Goals and tackle climate change, has not yet benefited from the recent advancements of versatile, task-agnostic vision language models (VLMs). A key reason is that the large-scale, semantically diverse image-text dataset required for developing VLMs is still absent for remote sensing images. Unlike natural images, remote sensing images and their associated text descriptions cannot be efficiently collected from the public Internet at scale. In this work, we bridge this gap by using geo-coordinates to automatically connect open, unlabeled remote sensing images with rich semantics covered in OpenStreetMap, and thus construct SkyScript, a comprehensive vision-language dataset for remote sensing images, comprising 2.6 million image-text pairs covering 29K distinct semantic tags. With continual pre-training on this dataset, we obtain a VLM that surpasses baseline models with a 6.2% average accuracy gain in zero-shot scene classification across seven benchmark datasets. It also demonstrates the ability of zero-shot transfer for fine-grained object attribute classification and cross-modal retrieval. We hope this dataset can support the advancement of VLMs for various multi-modal tasks in remote sensing, such as open-vocabulary classification, retrieval, captioning, and text-to-image synthesis.

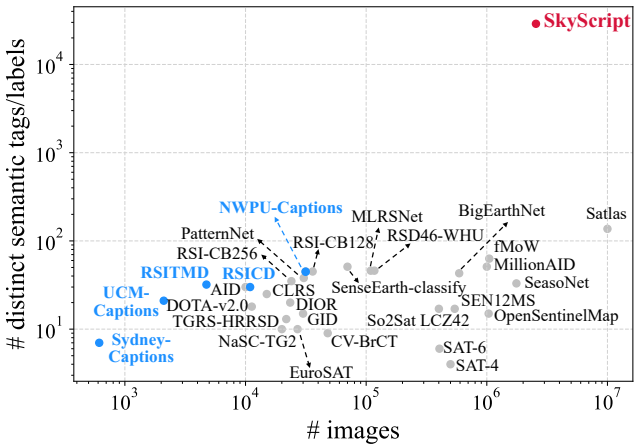


Figure 1: Comparison of general-purpose remote sensing datasets. Grey circles denote datasets for classification, object detection, or semantic segmentation (only datasets with $\geq$10,000 images are shown). Blue circles denote image-text datasets. Our dataset, SkyScript, is over two orders of magnitude more semantically diverse than existing remote sensing image-text datasets.

Introduction

Remote sensing imagery plays an important role for achieving Sustainable Development Goals (SDGs). Applying computer vision on remote sensing images can automate a broad range of applications, such as poverty estimation (Jean et al. 2016), crop yield prediction (You et al. 2017), deforestation detection (Torres et al. 2021), and renewable energy mapping (Yu et al. 2018; Kruitwagen et al. 2021). Facing the increasing risk of climate change, remote sensing imagery and the vision models built for it can further contribute to both mitigation and adaptation by enabling the observation of the Earth surface (Helber et al. 2019), detection of high-pollution industry (Lee et al. 2021), evaluation of carbon stock (Reiersen et al. 2022), and identification of vulnerable infrastructure and populations (Huang et al. 2021).

Despite numerous task-specific supervised learning models developed for remote sensing images, this domain has not yet fully benefited from the recent advancements of task-agnostic, versatile vision language models (VLMs) such as CLIP (Radford et al. 2021; Jia et al. 2021; Alayrac et al. 2022). This is because a key ingredient in the VLM's recipe for gaining versatility and generalizability is the large and semantically diverse collection of image-text pairs, which is still not readily available for remote sensing images. In the recent development of VLMs, such image-text pairs, ranging from million to billion scale, are usually collected from the public Internet through web crawling (Schuhmann et al. 2022). By contrast, remote sensing images are collected and owned exclusively by Earth observation companies, government agencies, or intergovernmental organizations (e.g., European Space Agency). Although these imagery data may be accessed through specialized, often non-free data pipelines (e.g., paid API service), they cannot be crawled from the web at scale. Even when these images can be obtained, they

*These authors contributed equally.

are usually standalone—not like web images which are often surrounded by semantically relevant text.

Due to the domain familiarity required for annotating remote sensing images, obtaining large and semantically diverse data through human annotations is also challenging. As Figure 1 shows, the sizes of existing remote sensing datasets can rarely reach the million-level, and the semantic classes are no more than 150. This significantly constrains the development of VLMs in the remote sensing domain.

In this work, we bridge this gap by constructing SkyScript[1], a large and semantically diverse image-text dataset for remote sensing. We achieve this by using geo-coordinates to connect open, unlabeled remote sensing images on Google Earth Engine (GEE) with rich semantic information covered in the OpenStreetMap (OSM) database. This yields a dataset with global coverage containing 2.6 million image-text pairs and covering 29K distinct semantic tags—two orders of magnitude richer than the existing remote sensing image-text datasets (Figure 1). The semantic information spans across object categories, subcategories, and detailed attributes (e.g., road surface materials). We demonstrate the value of this dataset by using continual pre-training to obtain a CLIP model that substantially outperforms the original CLIP and other baselines on three downstream remote sensing tasks: zero-shot scene classification, fined-grained attribute classification, and cross-modal retrieval. Our major contributions are summarized as follows:

- We create SkyScript, a large-scale, semantically diverse vision-language dataset for remote sensing images.
- SkyScript enables the development of a CLIP model for remote sensing that outperforms the original CLIP and other baselines in zero-shot scene classification.
- We further demonstrate the capabilities of SkyScript and its derived models on remote sensing cross-modal retrieval and zero-shot fine-grained attribute classification.

Related Work

Vision-language model (VLM). Connecting images with their corresponding text descriptions has been shown to be an effective approach for learning visual representations. (Joulin et al. 2016; Li et al. 2017a; Sariyildiz, Perez, and Larlus 2020; Desai and Johnson 2021). The development of CLIP and subsequent research further show that training on large-scale image-text datasets can yield VLMs that are generalizable to a wide variety of domains and robust to distribution shifts (Radford et al. 2021; Jia et al. 2021; Li et al. 2022; Yu et al. 2022; Alayrac et al. 2022; Chen et al. 2022). In particular, CLIP aligns image and text representations with contrastive learning, enabling zero-shot transfer of the learned representations to various computer vision tasks in the open world (Radford et al. 2021). However, VLMs for remote sensing images are underexplored, constrained by the availability of large-scale remote sensing image-text dataset.

[1]The dataset and associated models are publicly available at https://github.com/wangzhecheng/SkyScript

Foundation models for remote sensing. Following the recent advancement of self-supervised pre-training in computer vision, research aiming at building remote sensing foundation models has explored two major directions. The first line of research aims to learn the representations of remote sensing images through establishing the similarity metrics between multiple images captured at different geo-locations or multiple views of the same object/location (Jean et al. 2019; Manas et al. 2021; Ayush et al. 2021). Another line of research attempts to train a vision model with the goal to reconstruct masked patches of the input image (Cong et al. 2022; Fuller, Millard, and Green 2022a,b; Reed et al. 2022; Mendieta et al. 2023; Cha, Seo, and Lee 2023). Both these two types of models involve imagery only, requiring further fine-tuning to adapt the model to different downstream tasks. The development of remote sensing VLMs integrating both imagery and text for zero-shot transfer is constrained by the availability of remote sensing image-text data. A contemporaneous work on remote-sensing-specialized CLIP (Liu et al. 2023) relied on image-text pairs transformed from existing remote sensing datasets, hence limited by the scale and semantic diversity of the existing data.

Remote sensing dataset. Owing to the domain knowledge needed for annotating remote sensing images, the construction of large and semantically diverse remote sensing datasets is particularly challenging. Only a few remote sensing datasets contain million-scale images (Figure 1), but they cover only fixed sets of semantic classes no more than 150 (Yang and Newsam 2010; Cheng, Han, and Lu 2017; Xia et al. 2017; Christie et al. 2018; Zhou et al. 2018; Helber et al. 2019; Sumbul et al. 2019; Xiong et al. 2022; Long et al. 2021; Wang et al. 2021; Johnson, Treible, and Crispell 2022; Bastani et al. 2023; Jonathan Roberts and Albanie 2023). Existing remote sensing image-text datasets are particularly small (Qu et al. 2016; Lu et al. 2017; Yuan et al. 2022a; Cheng et al. 2022), with a size ranging from several hundred to tens of thousand images. Liu et al. (2023) constructed a remote sensing image-text dataset containing 166K images by aggregating multiple existing remote sensing datasets and automatically assembling captions based on annotated bounding boxes or segmentation masks. However, the number of unique classes involved is bounded by the semantic diversity of existing remote sensing datasets.

Dataset

Data Collection Approach

We construct the SkyScript dataset from the wild by linking large-scale yet unlabeled remote sensing image data with geo-tagged semantic information from OSM (Figure 2a). Here we introduce the data sources, the data selection approach, and how we derive image-text pairs.

Source of images. The data included in an open remote sensing image-text dataset ideally should have no licensing restrictions for research use. To this end, we acquire satellite and aerial images using the Google Earth Engine (GEE) platform which provides open access to large-scale remote sensing image collections from various sources allowing

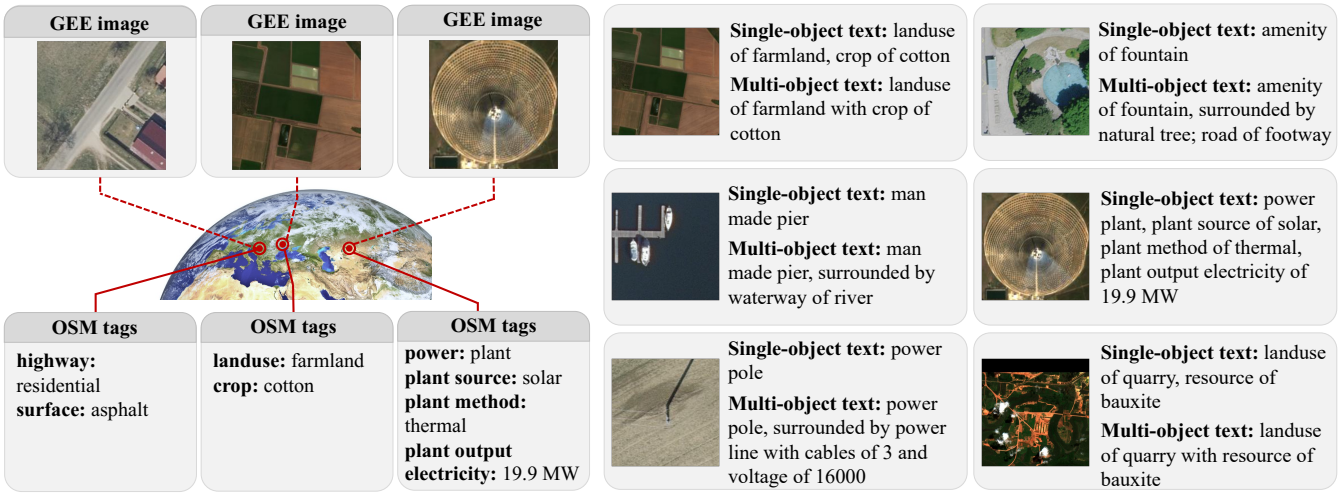


Figure 2: (a) The overview of the dataset construction approach. Remote sensing images of different sources are obtained from Google Earth Engine (GEE) platform, while semantic tags are obtained from OpenStreetMap (OSM). Images and tags are paired based on geo-coordinates. (b) Examples of image-caption pairs in the SkyScript dataset. Each image corresponds to a caption describing a single object and a caption describing multiple objects.

public sharing and redistribution. Specifically, Table 1 lists the image collections used in SkyScript, forming a multi-source, multi-resolution image pool with ground sampling distance (GSD) ranging from 0.1 m/pixel to 30 m/pixel. For each image collection, we only consider RGB bands even if multispectral images are present. The inclusion of additional bands is left for future research.

Image collection	GSD	Country
SWISSIMAGE 10 cm RGB imagery	0.1	Switzerland
Spain RGB orthophotos	0.1	Spain
Brandenburg RGBN orthophotos	0.2	Germany
Finland RGB NLS orthophotos	0.5	Finland
National Agriculture Imagery Program	0.6-1	U.S.
Planet SkySat Public Ortho, RGB	0.8	global
Planet SkySat Public Ortho, MS	2	global
Harmonized Sentinel-2 MSI, Level-2A	10	global
Landsat 8 C2 T1 TOA Reflectance	30	global
Landsat 9 C2 T1 TOA Reflectance	30	global

Table 1: Google Earth Engine (GEE) image collections used in SkyScript and their ground sampling distance (GSD) in meters and country information. MS: multispectral, MSI: multispectral instrument, TOA: top of atmosphere.

Source of semantics. To enable the generalizability of VLM, the semantics covered in the dataset should ideally include not just a large variety of object categories, but also fine-grained subcategories and attributes. To bridge this gap for remote sensing images, we leverage the rich semantic information contained in OpenStreetMap just (OSM), an open, crowdsourced geographic database. In OSM, each object on the map is described by one or more *tags*. Each *tag* consists of two free-form text fields, *key* and *value*. A *key*, by definition, is used to describe a topic, a category, or a type of feature (e.g., "surface"), while a *value* describes the specific feature, attribute, or subcategory given the *key* (e.g., "asphalt"). Figure 2a shows more examples of tags.

Previously, the rich semantics in OSM have not been fully exploited in constructing remote sensing datasets for supervised learning, primarily due to the concern of its uncurated nature. However, the capability of noisy but semantically diverse image-text datasets based on web crawling have been demonstrated in contrastive image-text pre-training (Radford et al. 2021; Jia et al. 2021). Our exploitation of uncurated yet rich semantics in OSM is based on this intuition.

Connect images with appropriate semantics. In OSM, some tags can be visually grounded in remote sensing images (e.g., "waterway": "stream"), while others cannot (e.g., "house number": "3"). Also, given an image of a certain GSD (e.g., 10 m), sufficiently large objects (e.g., "natural": "coastline") can be visually grounded at this image resolution while others cannot (e.g., "power": "pole").

To determine which tags should be included to describe an image, we develop a two-stage tag classification approach with CLIP embeddings of tags (key + value) as inputs (Figure 3). We use CLIP embeddings as they have already encoded visual information of tags after the image-text pre-training. In the first stage, a binary logistic regression model is used to predict whether the tag can be visually grounded at all. If a tag is predicted to be visually groundable, a second logistic regression model is further used to predict the maximum GSD (i.e., lowest resolution) at which the tag can be visually grounded. The prediction is one of the options:

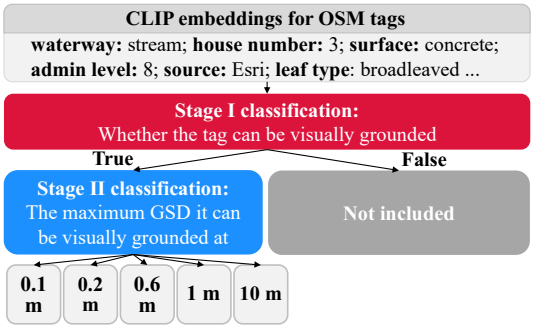


Figure 3: Tag classification. Given the CLIP embedding of a tag, first determine whether it can be visually grounded in remote sensing images, and if yes, then determine the maximum ground sampling distance (GSD) at which it can be visually grounded. Manually-curated labels are provided to train classification models at both stages.

0.1 m, 0.2 m, 0.6 m, 1 m, and 10 m. This is used to determine whether the tag should be included for describing an image given its GSD. See more details in Appendix A.1.

Data selection. Data selection consists of two steps: object selection and image selection. To ensure global representativeness and semantic diversity, we perform a two-stage object selection based on OSM. The first stage is random object selection. Specifically, we randomly select 400K grids across the globe, each with $0.01° \times 0.01°$ latitude/longitude intervals. Then we query objects located in these grids from the OSM database. This enables the inclusion of a random, globally representative subset of objects with common tags (e.g., "highway": "residential", "waterway": "river"). The second stage is targeted object selection. For rare tags that are not covered in the random object selection (e.g., "archaeological site": "tumulus", "castle type": "palace"), we directly query all objects containing these tags from the OSM database. This enables the inclusion of rare semantics into our dataset as well. For both stages, OverPass API is used for querying the OSM database.

For image selection, we use a object-centered scheme to determine the image collection and tile boundary for each image tile. Specifically, for objects represented as points or polylines, we use the maximum allowable GSD predicted by our second-stage tag classification model to determine the suitable image collections of which the GSD is smaller than the maximum allowable GSD. Tile boundaries are initially determined by making the object point or a randomly selected node on the object polyline located at the image center. For objects represented as polygons, we combine the bounding box of the polygon, a range of desired image tile size, and GSD information of each image collection to determine the suitable image collections to use. The bounding box of the polygon is used as the image tile boundary. An object will be skipped if no suitable image collection can be found for it. To add variations of the object location in an image, we further alter the image tile boundary randomly so that the object deviates moderately from the image center. See Appendix A.2 for more details about image selection.

Assembling caption. Each object can have multiple tags, and each tag consists of a key and a value. We convert tags into a caption by first connecting the key and the value with a connecting word (e.g., "of", "is") and then connecting multiple tags with comma or "and". Appendix A.3 provides details about the rules used for assembling caption.

For each image tile, we generate two captions. The first caption describes only the object that is used for determining the image tile boundary. We also obtain other OSM objects contained in the image tile by using geospatial overlay and then assemble a second caption describing multiple objects in the image tile. For example, if a object with a tag {"power": "pole"} is surrounded by another object with tags {"power": "minor line", "cables": "3", "voltage": "16000"}. Then the two captions are:

Single-object caption: *power pole*
Multi-object caption: *power pole, surrounded by power minor line with cables of 3 and voltage of 16000*

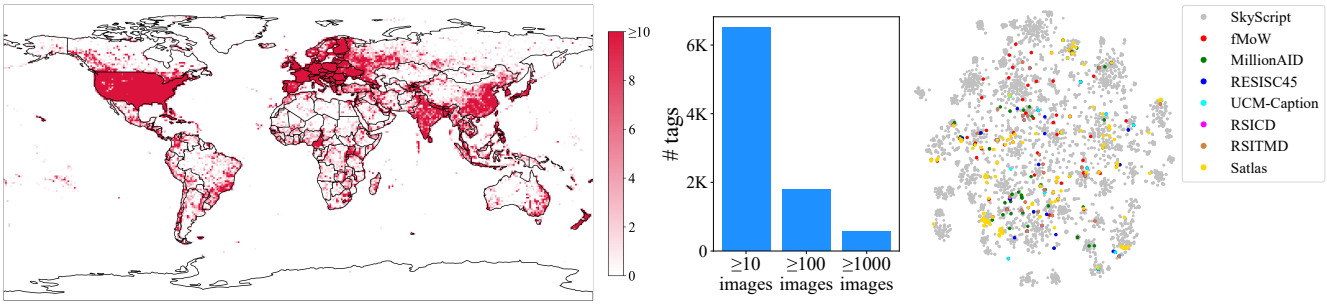
Filtering out uncorrelated image-text pairs. To reduce noisy information in this wild, uncurated dataset, we filter out the image-text pairs where the image and its corresponding caption are not sufficiently relevant. Specifically, after a caption is assembled from tags, we apply OpenAI's ViT-L/14 CLIP model to obtain the image and caption embeddings for each image-caption pair and calculate their cosine similarity. Low-similarity pairs indicate noisy samples (e.g., the object described by the caption is obscured by trees). We only keep image-caption pairs with a similarity value among the top. We experiment with top 20%, 30%, and 50% and the results are discussed in the next section.

Dataset Analysis

Dataset overview. Using the data collection approach described above, we obtain 5.2 million unfiltered remote sensing image-text pairs covering 44K distinct tags. Figure 2b show examples of image-text pairs. By keeping only image-text pairs with a similarity value among the top 50%, we obtain 2.6 million image-text pairs covering 29K distinct tags. These tags form 100K distinct single-object captions and 1.2 million distinct multi-object captions. GSD of images range from 0.1 m to 30 m. We randomly sample 1,000 image-tag pairs and manually check the tag accuracy, which is 96.1%.

30,000 image-text pairs are set aside for testing cross-modal retrieval performance. This test set is denoted as "SkyScript-retrieval". We also collect additional samples to form a classification dataset containing 70 classes for validating classification performance. Objects in this set are not covered by the main dataset. This auxiliary dataset, denoted as "SkyScript-classification", is detailed in Appendix A.4.

Geographic coverage. As Figure 4a shows, SkyScript has a good geographic coverage for all continents except Antarctica. The U.S. and Europe have a particularly high volume of objects covered, as high-resolution images ($<$1 m) required for visually grounding small objects are concentrated in these two regions. In the rest of the world, regions with



(a) (b) (c)

Figure 4: (a) Geographic coverage of the SkyScript dataset, represented as the number of covered objects in each of $1° \times 1°$ latitude/longitude grids. (b) Number of distinct tags with $\geq$10, $\geq$100, and $\geq$1,000 images included in SkyScript, respectively. (c) t-SNE visualization of semantic tag/label embeddings of different remote sensing datasets.

denser population tend to have more objects covered (e.g., east China, India). This reflects higher object density (e.g., building) and probably more complete OSM annotations in these densely-populated areas.

Semantic diversity. SkyScript is semantically diverse. As Figure 4b shows, 580 tags have $\geq$1,000 images, and more than 1,800 tags have $\geq$100 images included in the dataset. By using t-SNE to project the CLIP embeddings of tags covered in SkyScript as well as the semantics classes of other datasets into the 2D space (Figure 4c), we find that the semantics covered in the SkyScript dataset can be viewed as a superset of those covered in previous datasets. As demonstrated in Figure 2b, SkyScript includes not only broad category information, but also fine-grained information about object attributes (e.g., crop type of farmland, road surface).

Comparison with the remote sensing subset in LAION. LAION-2B is a huge English text-image dataset containing 2.3B images and their English captions obtained by web crawling (Schuhmann et al. 2022). We obtain a remote sensing subset of it by applying a binary classification model to determine whether an image in LAION-2B is a remote sensing image (see details in Appendix A.5). This subset, denoted as LAION-RS, contains 726K remote sensing image-text pairs—only 0.03% of all samples. This shows that web crawling cannot efficiently collect remote sensing image-text pairs at scale. We compare SkyScript with LAION-RS by evaluating them in downstream tasks.

Applications and Limitations

SkyScript can be used for developing models for a variety of tasks in remote sensing, such as open-vocabulary classification, cross-modal retrieval, image captioning, and text-to-image generation. It has potential values in a broad range of applications for sustainable development, such as monitoring the conditions of infrastructures (e.g., road, bridge), identifying illegal mining, tracking land use, and mapping distributed renewable energy resources. SkyScript alone can be used to develop domain-specialized VLMs for remote sensing by using continual pre-training. It can also be combined with other image-text datasets together to pre-train general-purpose VLMs from scratch.

SkyScript may have inherent bias. First, as currently we only consider remote sensing imagery without licensing restriction, high-resolution images are mainly limited to the U.S. and Europe, making the rest of the world underrepresented in the dataset. Second, OSM annotations are less complete in developing countries, which, again, makes the samples in these regions less abundant. Potential mitigation approaches include partnership with Earth observation companies or government agencies to obtain high-resolution images with wider coverage, as well as taking both models and human annotations in the loop to annotate more objects in underrepresented countries and regions. Moreover, we only use a simple rule-based approach to automatically assemble captions from tags. Using large language models (LLMs) to generate more natural and meaningful captions from tags warrants future exploration.

Experiments

We demonstrate the value of the SkyScript dataset by using it to develop a CLIP model for remote sensing images through continual pre-training. This model is further evaluated in zero-shot classification and cross-modal retrieval, showing substantive performance gains compared with the VLMs pre-trained on web-based image-text datasets.

CLIP Continual Pre-training

We perform continual pre-training on SkyScript to obtain a remote-sensing-specialized CLIP model, denoted as SkyCLIP. Specifically, we initialize the CLIP model with the weights that were pre-trained on web-based image-text data, and then further train it on the SkyScript dataset using image-text contrastive learning (Radford et al. 2021). We consider two ViT versions for CLIP: CLIP/ViT-B-32 and CLIP/ViT-L-14. CLIP/ViT-B-32 tokenizes images by patches of 32×32 pixels and has 12 transformer layers. By contrast, CLIP/ViT-L-14 uses a larger ViT that tokenizes images by patches of 14×14 pixels and has 24 transformer layers. For CLIP/ViT-B-32, we use the LAION-2B model weights to initialize the model, while for CLIP/ViT-L-14, we use the OpenAI model weights to initialize the model,

Model		Scene classification								Fine-grained classification		
	SkyScript	AID	Euro	fMoW	M-AID	P-Net	RESISC	RSI-CB	Avg.	Roof	Smooth	Surface
CLIP/ViT-B-32												
CLIP-original	40.16	69.55	32.11	17.62	57.27	64.09	65.71	41.26	49.66	31.50	26.80	61.36
Curated captions	40.03	71.05	33.85	18.02	57.48	66.56	66.04	42.73	50.82	28.50	27.80	60.91
RemoteCLIP	27.06	**87.05**	30.74	11.13	46.26	56.05	67.88	44.55	49.09	30.50	21.00	43.86
CLIP-laion-RS	40.77	69.55	37.63	19.16	56.59	64.79	64.63	41.79	50.59	28.83	27.60	62.27
SkyCLIP-50	52.98	70.90	33.30	19.24	62.69	72.18	66.67	46.20	53.02	26.00	**38.00**	**67.73**
CLIP/ViT-L-14												
CLIP-original	55.06	69.25	41.89	26.19	57.88	71.39	66.70	43.02	53.76	37.50	25.40	42.73
Curated captions	56.09	72.95	41.96	26.33	58.47	74.86	68.70	44.60	55.41	37.00	26.60	40.00
RemoteCLIP	34.40	70.85	27.81	16.77	47.20	61.91	**74.31**	**50.79**	49.95	34.33	34.20	55.45
CLIP-laion-RS	58.81	71.70	**54.30**	27.21	60.77	72.68	71.21	48.21	57.87	40.50	37.60	53.41
SkyCLIP-20	67.94	71.95	53.63	**28.04**	65.68	78.62	70.70	50.03	59.81	44.83	26.80	61.36
SkyCLIP-30	69.08	72.15	52.44	27.77	66.40	79.67	70.77	50.19	59.91	46.17	30.80	64.32
SkyCLIP-50	**70.89**	71.70	51.33	27.12	**67.45**	**80.88**	70.94	50.09	**59.93**	**46.83**	35.80	67.50

Table 2: Top-1 accuracy (%) for zero-shot classification. Euro: EuroSAT. M-AID: Million-AID. P-Net: PatternNet.

both based on the performances of initial model weights. The continual pre-training is conducted on 4 NVIDIA A100 GPUs with a batch size of 512 and total epochs of 20.

Zero-Shot Scene Classification

Benchmark datasets. We evaluate the zero-shot scene classification performance of SkyCLIP and other baseline models on seven common and comparatively large benchmark datasets: AID (Xia et al. 2017), EuroSAT (Helber et al. 2019), fMoW (Christie et al. 2018), Million-AID (Long et al. 2021), PatternNet (Zhou et al. 2018), NWPU-RESISC45 (Cheng, Han, and Lu 2017), and RSI-CB256 (Li et al. 2017b). Appendix A.6 provides more details about benchmark datasets. As a reference, we also evaluate the performance on the SkyScript classification dataset.

Definition of zero-shot transfer. Following the use of the term "zero-shot" in (Radford et al. 2021), rather than measuring the model generalizability on unseen object categories, here we evaluate the "zero-shot" transfer capability on unseen datasets. Although we cannot eliminate the possibility of SkyScript overlapping with the benchmark datasets in terms of geo-locations, they are unlikely to overlap in terms of images. This is because images in SkyScript are obtained from open image collections from GEE, while those in benchmark datasets are usually from commercial satellite images captured by different cameras, hence rendering different images even at the same geo-location.

Results. We use the original CLIP models as baselines, denoted as "CLIP-original". We use continual pre-training to train a CLIP model on the LAION-RS dataset, denoted as "CLIP-laion-RS". We also compare RemoteCLIP (Liu et al. 2023), as well as a dataset aggregating existing human-curated captioning datasets (NWPU-Captions, UCM-Captions, Sydney-Captions, RSICD) denoted as "Curated captions". Models trained on SkyScript are differentiated by the fraction of samples included (top 20%, 30%, or 50%) based on the descending order of CLIP similarity of image-text pairs, denoted as "SkyCLIP-20", "SkyCLIP-30", and "SkyCLIP-50", separately.

Table 2 shows the top-1 accuracies on different scene classification datasets. We find that SkyCLIP not only significantly improves the performance on the in-domain SkyScript classification dataset, but also consistently generalizes better to unseen benchmark datasets than the baseline models. For SkyCLIP-50 with the ViT-L-14 backbone, the average top-1 accuracy is 59.93%, with a 6.17% gain compared with the original CLIP/ViT-L-14, and a 2.06% gain compared with the CLIP model trained on the remote sensing subset of LAION-2B (CLIP-laion-RS). This suggests that SkyScript can yield a VLM with better zero-shot transfer capability on unseen remote sensing datasets than the general VLM. Also, simply taking the remote sensing subset of web-scale data for continual pre-training cannot reach the same performance as training on SkyScript. Moreover, SkyCLIP-50 outperforms RemoteCLIP (+9.98% average accuracy) and the model developed with human-curated captions (+4.52% average accuracy), indicating the importance of semantic diversity of image-text datasets in developing VLMs for remote sensing images.

As Table 2 shows, SkyCLIP-50 outperforms SkyCLIP-20 and SkyCLIP-30 on some benchmark datasets (Million-AID, PatternNet, RESISC45) but not on others. This reflects the trade-off between increasing the number of image-text pairs and ensuring their high quality as measured by their predicted pairwise similarity.

Zero-Shot Fine-Grained Classification

Compared with previous remote sensing datasets, the rich semantics covered in SkyScript can further facilitate the learning of fine-grained subcategories or attributes. Here we also demonstrate the fine-grained classification ability of the CLIP trained with SkyScript by testing its zero-shot performance on classifying three different attributes: roof shape, road smoothness, and road surface materials.

Model	SkyScript-retrieval		RSICD		RSITMD		UCM-Captions	
	img2txt	txt2img	img2txt	txt2img	img2txt	txt2img	img2txt	txt2img
Supervised cross-modal retrieval models								
AMFMN	-	-	(14.62)	(18.21)	(25.74)	(33.69)	(43.65)	(48.51)
LW-MCR	-	-	(12.68)	(18.50)	(24.70)	(32.45)	(43.02)	(47.68)
GaLR	-	-	(19.16)	(18.77)	(29.65)	(33.17)	-	-
Vision-language models								
CLIP-original	2.97	1.95	19.67	13.84	27.51	24.10	68.41	56.76
Curated captions	3.28	2.18	(20.56)	(16.37)	27.88	28.47	(70.95)	(59.59)
CLIP-laion-RS	3.85	2.81	22.66	18.52	30.24	29.67	69.68	57.56
RemoteCLIP	5.08	2.81	(36.32)	(33.20)	(43.95)	(44.94)	(79.05)	(74.98)
SkyCLIP-30	(8.53)	(7.73)	23.70	19.97	30.75	30.58	72.22	59.33

Table 3: Mean recall (%) for cross-modal retrieval. SkyCLIP-30 is trained on SkyScript (top 30% samples in terms of pairwise similarity) using multi-object captions. If a dataset is involved in training ("in-domain"), then it is bracketed with "()".

Test data. We construct a test set for evaluating the classification of each of the three attributes. Roof shapes include 6 classes: flat, hipped, gabled, dome, pyramidal, and round. Road smoothness includes 5 classes: excellent, good, intermediate, bad, and very bad. Road surface includes 5 classes: asphalt, concrete, grass, gravel, and sand. Each class contains 60 to 100 images. To abide by the zero-shot principle, we use Google and Bing Maps API instead of GEE to collect these images and do not consider the objects that have already been included in SkyScript. This ensure that images and objects in these fine-grained classification test sets are not seen during training. See more details in Appendix A.7.

Results. The rightmost panel of Table 2 shows the top-1 accuracies of SkyCLIP and other models on the fine-grained classification test sets. For roof shape classification, SkyCLIP with the ViT-L-14 image encoder achieves the highest top-1 accuracy (46.83%), while for road smoothness and surface material classification, SkyCLIP with the ViT-B-32 image encoder, though smaller, achieves the highest top-1 accuracies (38.00% and 67.73%). For both ViT-B-32 and ViT-L-14, SkyCLIP outperforms its CLIP counterpart by a significant margin (+6.37% to +24.77%), except a performance decrease of -5.5% for ViT-B-32 on roof shape classification. It is noteworthy that the performance gain of SkyCLIP in zero-shot fine-grained classification is more substantial than that on scene classification. This suggests the rich semantics covered in SkyScript can benefit those remote sensing applications without abundant labeled data, such as monitoring the conditions of civil infrastructure.

Cross-Modal Retrieval

We further demonstrate the value of SkyScript on cross-modal retrieval. Given the image/text query, the model needs to find the best match of text/image. This is done by calculating the cosine similarity between the query image/text embedding and the embedding of each text/image candidate.

Benchmark datasets. We evaluate the cross-retrieval performance on the test set of SkyScript ("in-domain") containing 30,000 image-text pairs, together with three benchmark image-text datasets for remote sensing ("zero-shot"): UCM-Captions (Qu et al. 2016), RISCD (Lu et al. 2017), and RSITMD (Yuan et al. 2022a).

Results. In addition to the comparison with CLIP-based models, We also compare SkyCLIP with three recently-developed remote sensing cross-modal retrieval models: AMFMN (Yuan et al. 2022a), LW-MCR (Yuan et al. 2022b), and GaLR (Yuan et al. 2022c). Benchmark datasets were seen during the training of these three models.

The mean recall, defined as the average of recall@1, recall@5, and recall@10, is used to measure the overall performance for both image-to-text and text-to-image retrieval. As Table 3 shows, SkyCLIP can steadily achieve better retrieval performance than CLIP and CLIP-laion-RS on three benchmark datasets that have never been seen during training (+2.57% to +6.48%). Its mean recalls on the benchmark datasets are comparable or even higher than the three supervised models (AMFMN, LW-MCR, GaLR) that have already seen the benchmark datasets during training. Notably, on UCM-Captions dataset, SkyCLIP outperforms the supervised models by $>$10% for text-to-image retrieval and $>$25% for image-to-text retrieval. This demonstrates the versatility of the visual and text representations learned from SkyScript, which can be transferred to unseen cross-modal remote sensing tasks in a zero-shot setting.

Conclusion

We present SkyScript, a large and semantically diverse image-text dataset for remote sensing images. We demonstrate its value by using it to derive a remote-sensing-specialized CLIP model outperforming baseline models across three downstream tasks in the zero-shot setting: scene classification, fine-grained classification, and cross-modal retrieval. The limitation in its geographic representativeness has also been discussed. Future work can explore its usage in other remote sensing tasks such as image captioning and text-to-image generation.

Acknowledgements

This work is in part supported by the Stanford Human-Centered AI (HAI) Fellowship, Stanford Data Science Scholarship, U.S. Department of Energy DE-EE0009359, and Stanford Precourt Institute for Energy.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10181–10190.

Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; and Kembhavi, A. 2023. Satlas: A Large-Scale Dataset for Remote Sensing Image Understanding. arXiv:2211.15660.

Cha, K.; Seo, J.; and Lee, T. 2023. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865–1883.

Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; and Wang, Z. 2022. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6180.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

Desai, K.; and Johnson, J. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11162–11173.

Fuller, A.; Millard, K.; and Green, J. R. 2022a. Satvit: Pretraining transformers for earth observation. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Fuller, A.; Millard, K.; and Green, J. R. 2022b. Transfer Learning with Pretrained Remote Sensing Transformers. *arXiv preprint arXiv:2209.14969*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. 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.

Huang, B.; Yang, J.; Streltsov, A.; Bradbury, K.; Collins, L. M.; and Malof, J. M. 2021. GridTracer: Automatic Mapping of Power Grids Using Deep Learning and Overhead Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 4956–4970.

Jean, N.; Burke, M.; Xie, S. M.; Davis, W. M.; Lobell, D.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353: 790 – 794.

Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; and Ermon, S. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3967–3974.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Johnson, N.; Treible, W.; and Crispell, D. 2022. Opensentinelmap: A large-scale land use dataset using openstreetmap and sentinel-2 imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1333–1341.

Jonathan Roberts, K. H.; and Albanie, S. 2023. SATIN: A Multi-Task Metadataset for Classifying Satellite Imagery using Vision-Language Models. *arXiv preprint arXiv:2304.11619*.

Joulin, A.; Van Der Maaten, L.; Jabri, A.; and Vasilache, N. 2016. Learning visual features from large weakly supervised data. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, 67–84. Springer.

Kruitwagen, L.; Story, K. T.; Friedrich, J.; Byers, L.; Skillman, S.; and Hepburn, C. 2021. A global inventory of photovoltaic solar energy generating units. *Nature*, 598: 604 – 610.

Lee, J.; Brooks, N. R.; Tajwar, F.; Burke, M.; Ermon, S.; Lobell, D.; Biswas, D.; and Luby, S. P. 2021. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

Li, A.; Jabri, A.; Joulin, A.; and Van Der Maaten, L. 2017a. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, 4183–4192.

Li, H.; Tao, C.; Wu, Z.; Chen, J.; Gong, J.; and Deng, M. 2017b. RSI-CB: A Large-Scale Remote Sensing Image Classification Benchmark Using Crowdsourced Data. *Sensors (Basel, Switzerland)*, 20.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; and Zhou, J. 2023. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *arXiv preprint arXiv:2306.11029*.

Long, Y.; Xia, G.-S.; Li, S.; Yang, W.; Yang, M. Y.; Zhu, X. X.; Zhang, L.; and Li, D. 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14: 4205–4230.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2017. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195.

Manas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; and Rodriguez, P. 2021. Seasonal contrast: Unsupervised pretraining from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9414–9423.

Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; Chen, C.; and Li, M. 2023. Gfm: Building geospatial foundation models via continual pretraining. *arXiv preprint arXiv:2302.04476*.

Qu, B.; Li, X.; Tao, D.; and Lu, X. 2016. Deep semantic understanding of high resolution remote sensing image. *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–5.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2022. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*.

Reiersen, G.; Dao, D.; Lütjens, B.; Klemmer, K.; Amara, K.; Steinegger, A.; Zhang, C.; and Zhu, X. 2022. ReforesTree: A Dataset for Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery. In *AAAI Conference on Artificial Intelligence*.

Sariyildiz, M. B.; Perez, J.; and Larlus, D. 2020. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 153–170. Springer.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sumbul, G.; Charfuelan, M.; Demir, B.; and Markl, V. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–5904. IEEE.

Torres, D. L.; Turnes, J. N.; Vega, P. J. S.; Feitosa, R. Q.; Silva, D. E.; Junior, J. M.; and de Almeida, C. A. 2021. Deforestation Detection with Fully Convolutional Networks in the Amazon Forest from Landsat-8 and Sentinel-2 Images. *Remote. Sens.*, 13: 5084.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*.

Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; and Lu, X. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3965–3981.

Xiong, Z.; Zhang, F.; Wang, Y.; Shi, Y.; and Zhu, X. X. 2022. EarthNets: Empowering AI in Earth Observation. *arXiv:2210.04936*.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.

You, J.; Li, X.; Low, M.; Lobell, D.; and Ermon, S. 2017. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In *AAAI Conference on Artificial Intelligence*.

Yu, J.; Wang, Z.; Majumdar, A.; and Rajagopal, R. 2018. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; and Sun, X. 2022a. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Yuan, Z.; Zhang, W.; Rong, X.; Li, X.; Chen, J.; Wang, H.; Fu, K.; and Sun, X. 2022b. A Lightweight Multi-Scale Crossmodal Text-Image Retrieval Method in Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Yuan, Z.; Zhang, W.; Tian, C.; Rong, X.; Zhang, Z.; Wang, H.; Fu, K.; and Sun, X. 2022c. Remote Sensing Cross-Modal Text-Image Retrieval Based on Global and Local Information. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Zhou, W.; Newsam, S.; Li, C.; and Shao, Z. 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145: 197–209.