WildSAT: Learning Satellite Image Representations from Wildlife Observations

Rangel Daroya[1] Elijah Cole[2] Oisin Mac Aodha[3] Grant Van Horn[1] Subhransu Maji[1]
[1]University of Massachusetts, Amherst [2]GenBio AI [3]University of Edinburgh

Abstract

Species distributions encode valuable ecological and environmental information, yet their potential for guiding representation learning in remote sensing remains underexplored. We introduce WildSAT, which pairs satellite images with millions of geo-tagged wildlife observations readily-available on citizen science platforms. WildSAT employs a contrastive learning approach that jointly leverages satellite images, species occurrence maps, and textual habitat descriptions to train or fine-tune models. This approach significantly improves performance on diverse satellite image recognition tasks, outperforming both ImageNet-pretrained models and satellite-specific baselines. Additionally, by aligning visual and textual information, WildSAT enables zero-shot retrieval, allowing users to search geographic locations based on textual descriptions. WildSAT surpasses recent cross-modal learning methods, including approaches that align satellite images with ground imagery or wildlife photos, demonstrating the advantages of our approach. Finally, we analyze the impact of key design choices and highlight the broad applicability of WildSAT to remote sensing and biodiversity monitoring.

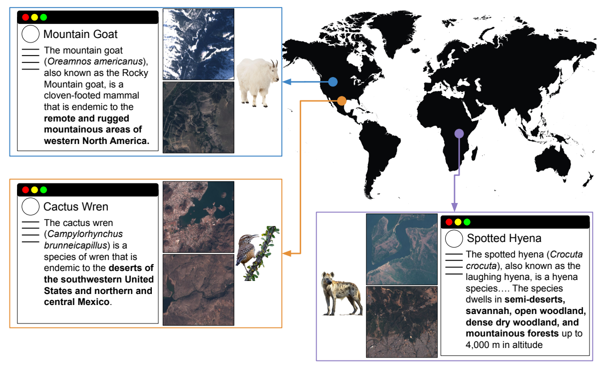


Figure 1. **Wildlife observations can provide valuable supervision for learning satellite image representations.** Known wildlife locations derived from human observations, coupled with descriptive information on species range, habitat, and other ecological attributes on Wikipedia, serve as a rich source of contextual information for satellite imagery. Our **WildSAT** approach leverages these additional data sources to (i) learn robust satellite image representations for downstream tasks, and (ii) complement and further improve existing models using continual pre-training.

1. Introduction

The growth in the number of satellites with imaging capabilities deployed over the past 50 years has provided an unprecedented ability to monitor the surface of the earth [33, 72, 74]. The image data derived from these remote sensors has been shown to be highly effective for diverse tasks such as estimating global tree canopy height [35, 63], detecting illegal fishing activity [20, 32, 52], crop monitoring [17, 29, 68], disaster management [53, 61, 67], among others. Central to building computer vision models for these tasks is the need for mechanisms for learning effective representations from image data. As a result of the distribution shift between remote sensing imagery and web-sourced images, a large body of work has emerged exploring the merits and trade-offs between different sources of supervision.

Direct supervision in the form of paired images and labels (*e.g*. image tiles with labels denoting land cover type) can be prohibitively expensive to obtain at a global scale [24]. To address this, there is growing interest to develop methods that learn remote sensing representations from self-supervision [28, 42, 44], multiple paired modalities [13, 43, 58], or other auxiliary sources [14, 66]. A useful supervision source needs to be globally distributed, correlated with the local landscape as viewed from an image, and able to discriminate regions at a fine spatial scale.

A promising auxiliary supervision source is provided by locations where different species of plants and animals can be found around the world. For example, *Mountain Goats (Oreamnos americanus)* are found in rugged mountainous areas, while habitat specialists like the *Cactus Wren (Campylorhynchus brunneicapillus)* are typically found in deserts nesting in spiny cacti (Fig. 1). Species location data offer a rich source of supervision, reflecting the local natural environment around each observation. It is also readily available from citizen science platforms such as iNaturalist [1] and eBird [59] which host hundreds of millions of wildlife observations. While species location data has improved fine-grained species classification [5, 9, 40], its potential for learning remote sensing representations remains unclear. Prior works have largely relied on anthropogenic labels (*e.g.* human-made features like roads, buildings, industrial areas) to learn satellite image representations [37, 43, 64], whereas we explore using wildlife observations as a complementary and potentially valuable signal.

We introduce a new approach that uses signals derived from species location observations. We take inspiration from recent work that attempt to fuse multi-modal ecological data and remote sensing imagery into a shared common embedding space [25, 57, 58]. WildSAT uses a contrastive learning objective to align satellite image, text, and location based on species observation data, bringing embeddings from the same area closer together and pushing those from different areas further apart. Through this method, we utilize information about the preferred habitats of species to improve satellite image representations.

We make the following contributions: (i) We introduce **WildSAT**, a new approach to learning remote sensing representations using species observation locations as a supervisory signal. (ii) We show WildSAT-derived representations are competitive with state-of-the-art satellite representations, while enabling zero-shot satellite image retrieval. (iii) We present a thorough evaluation that highlights WildSAT-derived representations not only outperform but also complement existing methods focused on anthropogenic labels by incorporating wildlife information. (iv) We perform ablation studies to show the impact of each component of our approach, and show WildSAT outperforms recent cross-modal methods like GRAFT [43] and TaxaBind [58]. The code and dataset are available at https://github.com/cvl-umass/wildsat.

2. Related Work

Previous works learn satellite image representations by training on large-scale remote sensing datasets from programs like Landsat [46], Sentinel [16, 26], or NAIP [47]. These methods range from using self-supervised [11, 28, 44], supervised [4, 50, 60], and cross-modal [13, 21, 43, 49, 54, 58] learning to learn rich image representations for downstream satellite-based tasks.

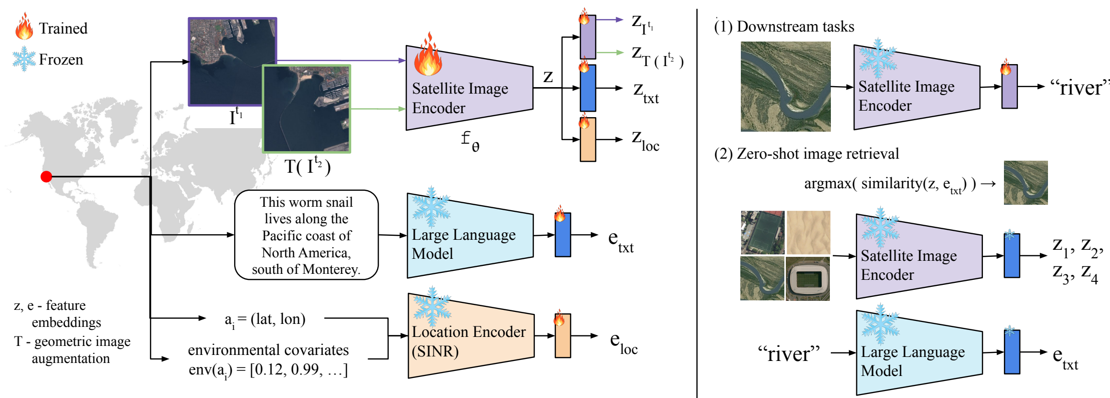
Several works have explored adding other modalities while training on satellite images [13, 21, 25, 30, 31, 43, 57, 58, 66]. A common approach uses geo-tagged images and pre-trained image-text encoders like CLIP [55], aligning new modalities to their embedding space using contrastive learning [13, 27, 31, 37, 43, 66]. This strategy has been used for various tasks: satellite image localization in GeoCLIP [66], bird species classification and mapping in BirdSAT [57], and improving plant species image representations in CRISP [25]. Models like GRAFT [43], TaxaBind [58], RemoteCLIP [37], and GeoBind [13] align multiple modalities at the same time for cross-modal retrieval and zero-shot tasks. Zermatten et al. [73] have also demonstrated the benefits of aligning satellite imagery with species observation data for zero-shot classification. TaxaBind was the first to use species geographic locations and satellite imagery, but it focuses on ecological tasks rather than satellite image tasks. We also expand on their approach by leveraging open-source Wikipedia text instead of taxonomic hierarchy data, offering a more diverse supervision for satellite image representations. Beyond contrastive learning, other methods use supervised learning to fuse embeddings of different modalities for predicting species range maps and encounter rates [14, 22, 62]. Most similar to our work is WikiSatNet [64] which uses location-aligned Wikipedia articles and satellite images to improve satellite image representations. However, while WikiSatNet and previous works [31, 37, 43, 64] primarily focus on anthropogenic data, we explore the impact of wildlife observations. Specifically, we investigate how species distributions—capturing habitat preferences, climate, and environmental factors—can serve as powerful signals.

While previous work has focused on improving species distribution modeling [14, 39, 57, 62] or fine-grained image classification [13, 41, 58] using satellite images, our work improves satellite image representations using wildlife observations. Our experiments show that both randomly initialized models and strong baselines, such as Prithvi [28], SatlasNet [4], and SeCo [44], benefit from this supervision on a wide range of satellite image tasks (Tab. 1, Tab. 2).

3. Method

We define the problem as follows: given an image encoder $f_\theta : \mathbf{I} \rightarrow \mathbf{z}$ with parameters θ, we want to find an optimal set of parameters θ^* that improves the performance of f on various remote sensing tasks through a robust satellite image feature representation $\mathbf{z}$. It takes an image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$ as input and outputs an embedding $\mathbf{z} \in \mathbb{R}^d$. We propose to optimize θ using our WildSAT framework using data consisting of satellite images, locations, environmental covariates, and text. We hypothesize that leveraging known environmental context around each species observation (Fig. 1) allows for more effective optimization of model parameters.

WildSAT. To supplement satellite images, we take advantage of additional modalities that naturally align based on the distribution of species throughout the globe. Information on species habitat can provide a rich source of supervision for improving satellite image representations, and we describe how to leverage this through our proposed WildSAT framework. Fig. 2 shows the architecture used to train a satellite image encoder f_θ. The encoder f can be any architecture (*e.g.* a ResNet50 [23], ViT-B/16 [15], *etc.*). The initial parameters θ can be randomly initialized, pre-trained on a different domain (*e.g.* ImageNet [12]), or pre-trained on a related dataset (*e.g.* SatlasPretrain [4]). The output em-



(a) Training Pipeline **(b) Inference Pipeline**

Figure 2. **Architecture for training and evaluating the satellite image encoder**. (a) The training pipeline uses the location of a species, the satellite images at those locations, the environmental covariates, and the Wikipedia text associated with the species. In addition to the alignment of image, text, and location modalities, the encoder is encouraged to learn additional image features by using temporal and geometric image transformations on the input satellite image. (b) Downstream tasks use the frozen satellite image encoder with an additional trainable layer (or layers). Alternatively, the predicted image embeddings can be used for zero-shot retrieval via text queries.

bedding $\mathbf{z}$ can be used for downstream remote sensing tasks such as classification and zero-shot image retrieval.

WildSAT aims to improve f by training on additional modalities related to species observation data. To incorporate other modalities into the satellite images, we use pre-trained models, *e.g.* SINR [10] for location and GritLM [48] for text. Given an initial image encoder f, we add three sets of linear layers to predict embeddings for images ($\mathbf{z}_{I^t}$), text ($\mathbf{z}_{\text{txt}}$), and locations ($\mathbf{z}_{\text{loc}}$). Similarly, both the pre-trained LLM (GritLM) and location encoder (SINR) have an added trainable linear layer to project their respective feature embeddings to $\mathbb{R}^d$ as $\mathbf{e}_{\text{txt}}$ and $\mathbf{e}_{\text{loc}}$, respectively. In addition to text and location, we also fine-tune the model on image-specific features by forcing embeddings of similar satellite images to be close to each another. For two satellite images $\mathbf{I}^{t_1}$ and $\mathbf{I}^{t_2}$ taken at the same location but at different times, their corresponding feature representations should be similar. We also apply geometric augmentations T such as flipping and random cropping on the latter image such that $f(\mathbf{I}^{t_1}) \approx f(T(\mathbf{I}^{t_2}))$.

Training. Our framework uses a contrastive learning objective to improve satellite image encoder embeddings. We jointly optimize the parameters of the model f_θ and the additional linear layers through the training objective in Eqn. 1. These loss terms correspond to a contrastive objective over image embeddings ($\mathcal{L}_{\text{img}}$), text embeddings ($\mathcal{L}_{\text{txt}}$), and location embeddings ($\mathcal{L}_{\text{loc}}$) of f. The embeddings $\mathbf{z}_{\mathbf{I}^t}, \mathbf{z}_{\text{txt}}, \mathbf{z}_{\text{loc}}$ are linear projections of the image embedding for each of the modalities (see Fig. 2):

$$\min_\theta \left[\underbrace{\mathcal{L}(\mathbf{Z}_{\mathbf{I}^{t_1}}, \mathbf{Z}_{T(\mathbf{I}^{t_2})})}_{\mathcal{L}_{\text{img}}} + \underbrace{\mathcal{L}(\mathbf{Z}_{\text{txt}}, \mathbf{E}_{\text{txt}})}_{\mathcal{L}_{\text{txt}}} + \underbrace{\mathcal{L}(\mathbf{Z}_{\text{loc}}, \mathbf{E}_{\text{loc}})}_{\mathcal{L}_{\text{loc}}} \right] \quad (1)$$

We compute distance between two sets of embeddings $\mathbf{Z}$ and $\mathbf{E}$ using a minibatch of n samples with the i-th embedding in $\mathbf{Z}$ aligned with the i-th embedding of $\mathbf{E}$ [51, 55].

Implementation details. During training, we fine-tune all satellite image encoders and added linear layers on the species observation dataset using Eqn. 1. For models pre-trained on out-of-domain datasets (*e.g.* ImageNet1K [12]), we apply parameter-efficient fine-tuning (PEFT) tailored to each architecture: ResNet50 uses scale and shift fine-tuning [19, 36], tuning only BatchNorm parameters, while ViT and Swin [38] use DoRa [45] on the attention layers. These techniques enable gradual parameter updates, allowing models to learn new satellite image features without forgetting those from their original domain. For a randomly initialized model or a model pre-trained in the same domain (*i.e.* satellite images), we fine-tune all parameters.

4. Dataset

To train the model, we combine images, text, location, and environmental covariates from publicly available datasets [3, 10, 16, 18, 22]. For a given species, we obtain its corresponding observation data (*e.g.* location) through iNaturalist [65], and a text description of its preferred habitat from its corresponding Wikipedia [3] page (Fig. 1). Environmental covariates are obtained from WorldClim2 [18] for a given location. Satellite images from Sentinel-2 are then retrieved based on the species observation locations. A total of 980,376 training samples were collected with location, satellite image, and text.

5. Experiments

We evaluate the representations learned by **WildSAT** via linear probing experiments. Starting with different mod-

	Average (w/ random)		Average (no random)	
Dataset	Base	+WS	Base	+WS
AID [69]	61.2	**77.0**	72.7	**79.4**
BEN20k [34, 60]	38.5	**53.1**	45.7	**53.4**
EuroSAT [24]	80.2	**93.8**	88.9	**94.3**
FMoW [8]	33.4	**41.1**	39.0	**43.3**
RESISC45 [7]	65.3	**81.0**	77.8	**83.5**
So2Sat20k [34, 75]	32.6	**47.6**	37.9	**48.2**
UCM [70]	68.8	**86.1**	81.8	**87.9**

Table 1. **Linear probing performance improvement on seven downstream datasets without (Base) and with WildSAT (+WS) fine-tuning**. Accuracy is visualized for all dataset plots except BEN20k that visualizes micro F1 score. The tables show average performance across all architectures: the left columns include models with random weights, and the right columns exclude them.

	Cashew1k [71]		SAcrop3k [2]	
	Base	+WS	Base	+WS
ImageNet [12]	70.3%	70.6%	24.3%	25.0%
MoCov3 [6]	71.4%	73.3%	22.9%	24.9%
SeCo [44]	62.6%	73.3%	22.3%	22.8%
SatlasNet [4]	55.2%	71.0%	19.4%	20.5%
Random	40.1%	72.6%	18.0%	20.3%
Average	59.9%	**72.2%**	21.4%	**22.7%**

Table 2. **Downstream satellite image segmentation results, reported using IoU, show WildSAT (+WS) improving on existing models**.

els and different parameter initializations (either random or pre-trained), we evaluate the performance before and after fine-tuning. When probing for each downstream dataset, the trained satellite image encoder is frozen and a randomly initialized decoder is added (Fig. 2b.1). For all tasks except segmentation, a linear layer is used for the decoder. Segmentation tasks use a convolutional-based decoder with a U-Net architecture [56]. Only the decoder is trained for each downstream task to assess the impact of the image embedding $\mathbf{z}$ without diluting its representation.

Base models in subsequent experiments refer to the different pre-trained encoders before we fine-tune with WildSAT. We experiment on 12 pre-training methods spanning random initialization, in-domain pre-training, and out-of-domain pre-training. These cover different architectures ResNet50, Swin-T, Swin-B, ViT-B/16, and ViT-L/16 for a total of 20 base models.

6. Results and Discussion

Classification and segmentation performance. Tab. 1 and Tab. 2 display the results on the 7 downstream classification datasets and 2 segmentation datasets, respectively, across 15 different architectures and pre-training methods. The addition of WildSAT improves 108 of the 115 settings

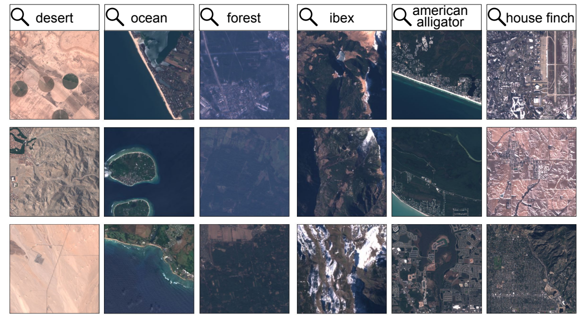


Figure 3. **Zero-shot results for text-based satellite image retrieval**. The columns show the top 3 images returned given the text query on top. A model can be queried using general landscape descriptions (*e.g.* 'desert', 'ocean', 'forest'). In addition, specific wildlife text such as 'ibex' and 'house finch' can be used as queries to view the types of environment they inhabit.

with an overall average improvement ranging from 7.7% to 17.4% in the different datasets (4.3% to 10.4% without the randomly initialized models).

The results in Tab. 1 and Tab. 2 highlight the performance improvements WildSAT contributes. These improvements may be attributed to our use of diverse supervision—integrating images, text embeddings, and species data at scale. This strategy ultimately helps in downstream tasks, particularly for both increasing true positive rates on classes related to habitats (*e.g.* forests, deserts), while reducing false positives on the same types of classes.

Zero-shot image retrieval. When trained using our WildSAT framework, we observe that models learn wildlife-specific attributes. By using the frozen satellite image encoder and a large language model, a user can input text to query satellite images. The top k images with the most similar embeddings to the text embeddings (computed using cosine similarity) can be retrieved (Fig. 2b.2). Fig. 3 displays examples of satellite images retrieved given different text queries. General descriptions of landscapes or locations can be used for querying such as 'desert', 'ocean', or 'forest'. At the same time, specific wildlife text can also be used as queries such as 'ibex' or 'house finch'. Zero-shot retrieval returns images of the habitat of the wildlife.

7. Conclusion

While satellite images are often used to interpolate sparse wildlife observations to create species range maps, our work demonstrates that these observations also provide a rich source of supervision for learning satellite image representations. WildSAT can not only learn high-quality representations from scratch but also improve performance of strong pre-trained models, such as those trained on ImageNet and satellite imagery datasets, across a range of satellite imagery tasks.

Acknowledgements. We thank the iNaturalist community for providing the data used for training. Experiments were performed on the University of Massachusetts GPU cluster funded by the Mass. Technology Collaborative. RD and SM were supported in part by NASA grant 80NSSC22K1487 and NSF grant 2329927.

References

[1] iNaturalist. https://www.inaturalist.org. Accessed on 2025-03-03. 1

[2] Planet, Radiant Earth Foundation, Western Cape Department of Agriculture, German Aerospace Center (DLR): A fusion dataset for crop type classification in Western Cape, South Africa. https://source.coop/esa/fusion-competition. Accessed on 2025-03-03. 4

[3] Wikipedia. https://www.wikipedia.org. Accessed on 2025-03-03. 3

[4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 2, 4

[5] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014. 1

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 4

[7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 4

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 4

[9] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 1

[10] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International Conference on Machine Learning*, pages 6320–6342. PMLR, 2023. 3

[11] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 4

[13] Aayush Dhakal, Subash Khanal, Srikumar Sastry, Adeel Ahmad, and Nathan Jacobs. Geobind: Binding text, image, and audio through satellite images. In *International Geoscience and Remote Sensing Symposium*, 2024. 1, 2

[14] Johannes Dollinger, Philipp Brun, Vivien Sainte Fare Garnot, and Jan Dirk Wegner. Sat-sinr: High-resolution species distribution models through satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2024. 1, 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[16] ESA. Sentinel-1-missions-sentinel online-sentinel online. *Eur. Sp. Agency*, 2022. 2, 3

[17] Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P Gomes. A gnn-rnn approach for harnessing geospatial and temporal information: application to crop yield prediction. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11873–11881, 2022. 1

[18] Stephen E Fick and Robert J Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315, 2017. 3

[19] Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. In *International Conference on Learning Representations*, 2021. 3

[20] Rollan C Geronimo, Erik C Franklin, Russell E Brainard, Christopher D Elvidge, Mudjekeewis D Santos, Roberto Venegas, and Camilo Mora. Mapping fishing activities and suitable fishing grounds using nighttime satellite images and maximum entropy modelling. *Remote Sensing*, 10(10):1604, 2018. 1

[21] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27672–27683, 2024. 2

[22] Max Hamilton, Christian Lange, Elijah Cole, Alexander Shepard, Samuel Heinrich, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. Combining observational data and language for species range estimation. In *Advances in Neural Information Processing Systems*, 2024. 2, 3

[23] 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. 2

[24] 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. 1, 4

[25] Andy V Huynh, Lauren E Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *European Conference on Computer Vision*, 2024. 2

[26] Markus Immitzer, Francesco Vuolo, and Clement Atzberger. First experience with sentinel-2 data for crop and tree species classifications in central europe. *Remote sensing*, 8(3):166, 2016. 2

[27] Pallavi Jain, Diego Marcos, Dino Ienco, Roberto Interdonato, Aayush Dhakal, Nathan Jacobs, and Tristan Berchoux. Aligning geo-tagged clip representations and satellite imagery for few-shot land use classification. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 319–323. IEEE, 2024. 2

[28] Johannes Jakubik, S Roy, CE Phillips, P Fraccaro, D Godwin, B Zadrozny, D Szwarcman, C Gomes, G Nyirjesy, B Edwards, et al. Foundation models for generalist geospatial artificial intelligence, 2023. *URL https://arxiv.org/abs/2310.18660*, 2023. 1, 2

[29] Priyabrata Karmakar, Shyh Wei Teng, Manzur Murshed, Shaoning Pang, Yanyu Li, and Hao Lin. Crop monitoring by multimodal remote sensing: A review. *Remote Sensing Applications: Society and Environment*, 33:101093, 2024. 1

[30] Subash Khanal, Srikumar Sastry, Aayush Dhakal, and Nathan Jacobs. Learning tri-modal embeddings for zero-shot soundscape mapping. In *BMVC*, 2023. 2

[31] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *AAAI Conference on Artificial Intelligence*, 2024. 2

[32] Andrey A Kurekin, Benjamin R Loveday, Oliver Clements, Graham D Quartly, Peter I Miller, George Wiafe, and Kwame Adu Agyekum. Operational monitoring of illegal fishing in ghana through exploitation of satellite earth observation and ais data. *Remote Sensing*, 11(3):293, 2019. 1

[33] Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017. 1

[34] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geobench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36, 2023. 4

[35] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *Nature Ecology & Evolution*, 2023. 1

[36] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 3

[37] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 3

[39] Titouan Lorieul, Elijah Cole, Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, and Alexis Joly. Overview of geolifeclef 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data. In *CLEF (Working Notes)*, pages 1940–1956, 2022. 2

[40] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1

[41] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, pages 23498–23515. PMLR, 2023. 2

[42] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 1

[43] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[44] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 1, 2, 4

[45] Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. *arXiv preprint arXiv:2405.17357*, 2024. 3

[46] Jeffrey G Masek, Michael A Wulder, Brian Markham, Joel McCorkel, Christopher J Crawford, James Storey, and Del T Jenstrom. Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment*, 248:111968, 2020. 2

[47] Aaron E Maxwell, Timothy A Warner, Brian C Vanderbilt, and Christopher A Ramezan. Land cover classification and feature extraction from national agriculture imagery program (naip) orthoimagery: A review. *Photogrammetric Engineering & Remote Sensing*, 83(11):737–747, 2017. 2

[48] Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024. 3

[49] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, 2024. 2

[50] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 2

[51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[52] Fernando Paolo, Tsu-ting Tim Lin, Ritwik Gupta, Bryce Goodman, Nirav Patel, Daniel Kuster, David Kroodsma, and Jared Dunnmon. xview3-sar: Detecting dark fishing activity using synthetic aperture radar imagery. *Advances in Neural Information Processing Systems*, 2022. 1

[53] Gustavo Perez, Subhransu Maji, and Daniel Sheldon. Discount: counting in large image collections with detector-based importance sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22294–22302, 2024. 1

[54] Lukas Picek, Christophe Botella, Maximilien Servajean, César Leblanc, Rémi Palard, Théo Larcher, Benjamin Deneu, Diego Marcos, Pierre Bonnet, and Alexis Joly. Geoplant: Spatial plant species prediction dataset. *Advances in Neural Information Processing Systems - Datasets and Benchmark*, 2024. 2

[55] 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. 2, 3

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI*, pages 234–241. Springer, 2015. 4

[57] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7136–7145, 2024. 2

[58] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 1, 2

[59] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009. 1

[60] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 2, 4

[61] Wenjuan Sun, Paolo Bocchini, and Brian D Davison. Applications of artificial intelligence for disaster management. *Natural Hazards*, 103(3):2631–2689, 2020. 1

[62] Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. *Advances in Neural Information Processing Systems - Datasets and Benchmarks*, 36, 2023. 2

[63] Catherine Torres de Almeida, Jéssica Gerente, Jamerson Rodrigo dos Prazeres Campos, Francisco Caruso Gomes Junior, Lucas Antonio Providelo, Guilherme Marchiori, and Xinjian Chen. Canopy height mapping by sentinel 1 and 2 satellite images, airborne lidar data, and machine learning. *Remote Sensing*, 14(16):4112, 2022. 1

[64] Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. Learning to interpret satellite images using wikipedia. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019. 2

[65] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 3

[66] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geolocalization. *Advances in Neural Information Processing Systems*, 2024. 1, 2

[67] Stefan Voigt, Thomas Kemper, Torsten Riedlinger, Ralph Kiefl, Klaas Scholte, and Harald Mehl. Satellite image analysis for disaster and crisis-management support. *IEEE transactions on geoscience and remote sensing*, 45(6):1520–1528, 2007. 1

[68] Bingfang Wu, René Gommes, Miao Zhang, Hongwei Zeng, Nana Yan, Wentao Zou, Yang Zheng, Ning Zhang, Sheng Chang, Qiang Xing, et al. Global crop monitoring: a satellite-based hierarchical approach. *Remote Sensing*, 7(4): 3907–3933, 2015. 1

[69] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 4

[70] Yi Yang and Shawn Newsam. 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*, pages 270–279, 2010. 4

[71] Leikun Yin, Rahul Ghosh, Chenxi Lin, David Hale, Christoph Weigl, James Obarowski, Junxiong Zhou, Jessica Till, Xiaowei Jia, Nanshan You, et al. Mapping smallholder cashew plantations to inform sustainable tree crop expansion in benin. *Remote Sensing of Environment*, 295:113695, 2023. 4

[72] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment*, 241:111716, 2020. 1

[73] Valerie Zermatten, Javiera Castillo-Navarro, Pallavi Jain, Devis Tuia, and Diego Marcos. Ecowikirs: Learning ecolog-

ical representation of satellite images from weak supervision with species observations and wikipedia. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2275–2285, 2025. 2

[74] 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. *Geoscience and remote sensing magazine*, 5 (4):8–36, 2017. 1

[75] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, et al. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. 4