WildSAT: Learning Satellite Image Representations from Wildlife Observations

Supplementary Material

A. Datasets

A.1. Training Data Distribution

Fig. A1 shows the spatial distribution of all data we collected across the globe. Most of the data are from United States and Europe, corresponding to the wildlife observation data available from citizen science platforms [1, 2]. Despite this, we show that models trained with WildSAT can still generalize to areas beyond the US and Europe, with segmentation improvements even in areas like Africa (see Fig. A3).

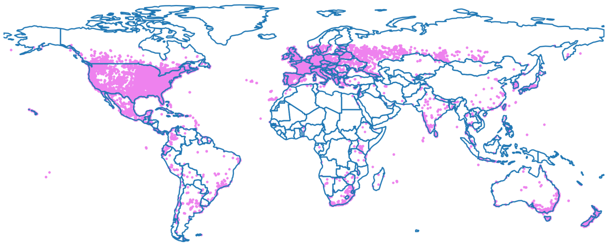

Figure A1. **Distribution of data points with satellite image, environmental covariates, and text**. The alignment of different modalities is guided by the geographic distribution of species.

A.2. Satellite Image Evaluation Datasets

Below we briefly describe the different satellite image classification datasets used for evaluation. For each of the datasets, the splits for training, validation, and testing follow those provided from the respective sources.
UCM [78] is an image classification dataset that contains 21 classes with 100 each covering USA. Each image is 256×256 with a resolution of 1 ft.
AID [76] is an image classification dataset that contains 30 class, each containing from 220 to 400 images from Google Earth. Each image is 600×600 with a resolution ranging from 0.5 to 8 m.
RESISC45 [8] is an image classification dataset that contains 45 classes with 700 images each, sourced from Google Earth. An image is 256×256 with a resolution ranging from 0.2 to 30 m.
FMoW [9] is an image classification dataset that contains 63 classes from over 200 countries with a total of over 400k images from QuickBird, GeoEye, and WorldView satellites. Images vary in size and resolution and each class has a different number of images.
EuroSAT [28] is an image classification dataset that contains 10 classes of land use and land cover from Europe. Each image is 64×64 with 10 m resolution. The dataset has 27k images with each class having a different number of images.
So2Sat20k [84] is an image classification dataset that contains 17 classes across different climate zones with global coverage. The full dataset contains 400k pairs of Sentinel-1 and Sentinel-2 images. We use the GEO-Bench [38] version referred to as "So2Sat20k" which contains 20k training samples.
BigEarthNet20k (BEN20k) [65] is a multi-label classification dataset with 43 classes. The full dataset is from 10 countries in Europe with 590k Sentinel-2 images. We use the GEO-Bench [38] version "BEN20k" which contains 20k training samples.
Cashew1k [79] is a segmentation dataset with 7 categories mapping cashew plantations in Benin, Africa. It contains more than 1k training examples of Sentinel-2 images. We use the available dataset in GEO-bench.
SACrop3k [3] is a segmentation dataset with 10 categories mapping crop types in South Africa. It contains 3k training samples of Sentinel-2 images. We use the dataset provided in GEO-bench.

B. Additional Implementation Details

WildSAT Training Details. We train each base model on the WildSAT framework for 25 epochs using an Adam optimizer with a learning rate of 1×10^{-4}, with an embedding dimension $d = 512$, and a batch size of 64. Random cropping, resizing, jitter, and channel mixing are applied on satellite images as augmentations. Each satellite image is paired with a

	Encoder	UCM [78]		AID [76]		RESISC45 [8]		FMoW [9]		EuroSAT [28]		So2Sat20k [84]		BEN20k [65]	
		Base	+WS	Base	+WS	Base	+WS	Base	+WS	Base	+WS	Base	+WS	Base	+WS
ViT-B/16	ImageNet [14]	93.2	97.5	84.4	88.9	88.2	93.0	43.8	51.4	94.5	97.3	41.8	55.2	52.3	58.2
	MoCov3 [7]	94.2	95.1	86.0	86.9	89.1	90.3	51.1	52.9	95.9	97.1	47.6	56.6	51.6	57.0
	CLIP [59]	94.5	96.3	86.3	88.0	92.1	93.0	51.5	52.8	92.2	97.1	37.6	49.7	47.1	59.1
	Prithvi-100M* [32]	49.7	85.5	35.9	71.2	42.6	73.5	19.2	30.5	67.3	93.5	21.5	45.1	33.6	50.6
	SatCLIP* [35]	38.2	50.3	37.4	46.4	40.4	46.2	19.0	20.1	74.6	79.4	39.0	43.1	27.0	28.7
	Random weights	4.1	75.5	3.8	62.1	1.9	62.4	8.0	26.0	11.1	90.4	5.9	46.8	0.0	51.2
Swin-T	ImageNet [14]	94.0	96.9	87.9	89.0	90.4	91.8	47.6	50.7	96.2	97.3	48.3	51.5	54.1	57.7
	SatlasNet [5]	89.6	91.2	74.3	81.2	80.2	86.5	31.8	44.6	90.8	95.5	36.4	53.1	48.7	56.5
	Random weights	21.0	81.7	19.5	72.0	19.9	74.9	12.1	33.4	59.9	92.7	21.9	45.9	9.8	52.4
ResNet50	ImageNet [14]	94.2	93.6	87.8	86.7	90.5	90.1	47.3	46.0	95.5	96.0	36.1	46.6	55.8	57.5
	MoCov3 [7]	92.0	93.5	83.0	83.3	88.0	87.6	50.2	45.7	93.5	95.1	27.2	42.5	46.6	53.8
	SatlasNet [5]	86.8	90.1	72.5	79.4	81.8	85.4	34.7	42.4	93.5	95.4	33.9	44.8	44.9	56.4
	SeCo [48]	86.1	88.8	74.3	79.6	80.2	86.3	35.9	42.8	89.7	95.5	39.9	46.0	44.3	57.3
	SatCLIP* [35]	69.4	76.2	63.1	71.8	70.2	78.8	36.2	39.9	83.4	92.9	45.4	44.9	42.3	48.2
	Random weights	24.7	79.9	22.3	68.2	24.5	74.7	12.7	36.9	65.2	92.2	5.9	42.3	19.9	51.3
Overall average		68.8	**86.1**	61.2	**77.0**	65.3	**81.0**	33.4	**41.1**	80.2	**93.8**	32.6	**47.6**	38.5	**53.1**
Average w/o random		81.8	**87.9**	72.7	**79.4**	77.8	**83.5**	39.0	**43.3**	88.9	**94.3**	37.9	**48.3**	45.7	**53.4**

Table A1. **Results of linear probing different models on seven downstream datasets without (Base) and with (+WS) WildSAT fine-tuning**. Accuracy is reported for all datasets except BEN20k that reports micro F1 score. Base refers to the original models specified as the encoder and +WS refers to the same models further trained on the species observation data. Fine-tuning models with species observation show significant improvement over the base models. Both Prithvi-100M and SatCLIP are pre-trained with multispectral images, but for consistency across downstream datasets and models, only RGB bands are used. We include results on multispectral images in Tab. A4.

wildlife observation location for positive samples. For each of these pairs, a section of text is randomly sampled from the Wikipedia [4] page of the species. A satellite image of the same location, but from a different time, is also randomly sampled for image augmentation. Negative samples are randomly selected from other species and locations that do not correspond to the observed wildlife locations. For multiple species that are found in the same location, multiple texts are associated with the same location. For the same species present in different locations, the text for each location is randomly sampled from the same Wikipedia page (*i.e.* each location would correspond to a random section on the same Wikipedia page). Downstream tasks follow the train/val/test split provided in each of the datasets. Training takes an average of 8 hours on 2 NVIDIA L40S.

Satellite Image Filtering. All satellite images are from Sentinel-2A and Sentinel-2B. We follow the same data collection procedure from SatlasPretrain [5], where satellite images are downloaded from EU's Sentinel Data [18]. Each image is 512×512 pixels with a 10 m resolution per pixel. Only images that are tagged with significantly less cloud cover from [5] are used. In addition, we only use satellite images that were taken in the same time range as the wildlife observation data (from 2017 to 2021). This is done since we do not use the exact observation date and time as an input to the model; we consider all observations throughout the time range. At the same time, the text descriptions we use also refer to all types of habitats regardless of time of year.

Multi-spectral Baselines. Prithvi-100M [32] and SatCLIP [35] originally use multi-spectral data in their pre-training. However, for general applicability and easy comparisons with other models, we only use RGB bands in Table 1. When WildSAT is applied to these models, we only fine-tune with the three bands, and set other bands to zero. At the same time, when applying both the base models and WildSAT fine-tuned models on downstream satellite image datasets, we also set other bands to zero. We also explore using multiple bands as inputs in Tab. A4, and discuss the results in the next section. In this case, when a band used by a model is not available in the dataset, we set the values of the band to zero. Otherwise, we use all the bands available.

Encoder		UCM [78]		AID [76]		RESISC45 [8]		FMoW [9]		EuroSAT [28]		So2Sat20k [84]		BEN20k [65]	
		Base	+WS	Base	+WS	Base	+WS	Base	+WS	Base	+WS	Base	+WS	Base	+WS
ResNet50	ImageNet1K V2 [60]	94.2	93.6	87.8	86.7	90.5	90.1	47.3	46.0	95.5	96.0	36.1	46.6	55.8	57.5
ResNet50	ImageNet1K V1 [14]	92.5	93.5	90.4	88.8	85.1	84.7	40.7	37.0	88.0	94.9	38.8	48.2	46.7	53.7
ViT-L/16	SatMAE [13]	23.8	86.1	25.1	70.6	26.1	74.6	13.9	33.4	48.3	94.5	15.6	48.6	18.4	51.0

Table A2. **Additional linear probing results on satellite image classification datasets**. Accuracy is reported for all datasets except BEN20k that reports micro F1 score. 'Base' refers to the original models specified as the encoder and '+WS' refers to the same models further trained with WildSAT. Consistent with results thus far, fine-tuning models with species observation generally show significant improvement over the base models.

	UCM [78]	AID [76]	RESISC45 [8]	FMoW [9]	EuroSAT [28]	So2Sat20k [84]	BEN20k [65]
TaxaBind [63]	80.5	67.7	72.6	31.2	85.2	33.9	47.6
GRAFT [47]	81.1	76.1	83.3	39.3	90.9	36.6	48.0
RemoteCLIP [41]	96.1	86.1	90.9	45.7	93.3	35.5	49.4
CLIP [59]	94.5	86.3	92.1	51.5	92.2	37.6	47.1
WildSAT (Ours)	**96.3**	**88.0**	**93.0**	**52.8**	**97.1**	**49.7**	**59.1**

Table A3. **Linear probing results on downstream satellite classification datasets using models with CLIP as the base.** Results are reported as accuracy, except for BEN20k which uses micro F1. TaxaBind, GRAFT, and RemoteCLIP fine-tune a CLIP backbone and use additional modalities such as text, ground images, and satellite images for cross-modal tasks. WildSAT outperforms both the standard CLIP and the previous methods that also fine-tune on CLIP.

Encoder		So2Sat20k [84]		BEN20k [65]	
		Base	+WS	Base	+WS
ViT-B/16	Prithvi-100M [32]	28.7	50.1	34.4	53.9
ViT-B/16	SatCLIP [35]	48.8	48.8	33.4	36.1
ResNet50	SatCLIP [35]	45.8	46.3	46.7	54.3
Average		41.1	**48.4**	38.2	**48.1**

Table A4. **Linear probing results on models using multispectral images**. Accuracy is reported for So2Sat20k and micro F1 score for BEN20k. 'Base' refers to the original models specified as the encoder and '+WS' refers to the same models further trained with WildSAT. We see improved base model performance with using more bands (compared to using only RGB), and show that the addition of WildSAT further improves performance even with multispectral models and datasets.

C. Additional Results

C.1. Satellite Image Classification

Linear Probing results. Tab. A1 shows the raw numbers from which Fig. 3 was generated.

ImageNet V1 results. Tab. A2 shows results on the ImageNet V1 base model that previous works have used [46, 48]. The results in Table 1 in the main paper include a base model using ImageNet V2 (also included in Tab. A2 for reference) which has generally better performance across the downstream satellite image datasets. Tab. A2 additionally shows results on SatMAE [13], a ViT-L/16 model that was pre-trained with the MAE framework on satellite images. Similar to previous results, WildSAT improves performance across the seven satellite image classification datasets evaluated.

Multi-spectral results. Tab. A4 shows additional results when using multispectral images as input to the models that support this capability. Prithvi-100M and SatCLIP accept multiple bands as input, while So2Sat20k and BEN20k are datasets on GEO-bench that contain multiple bands. We find that WildSAT also improves on these multispectral models, demonstrating the broad applicability of our method.

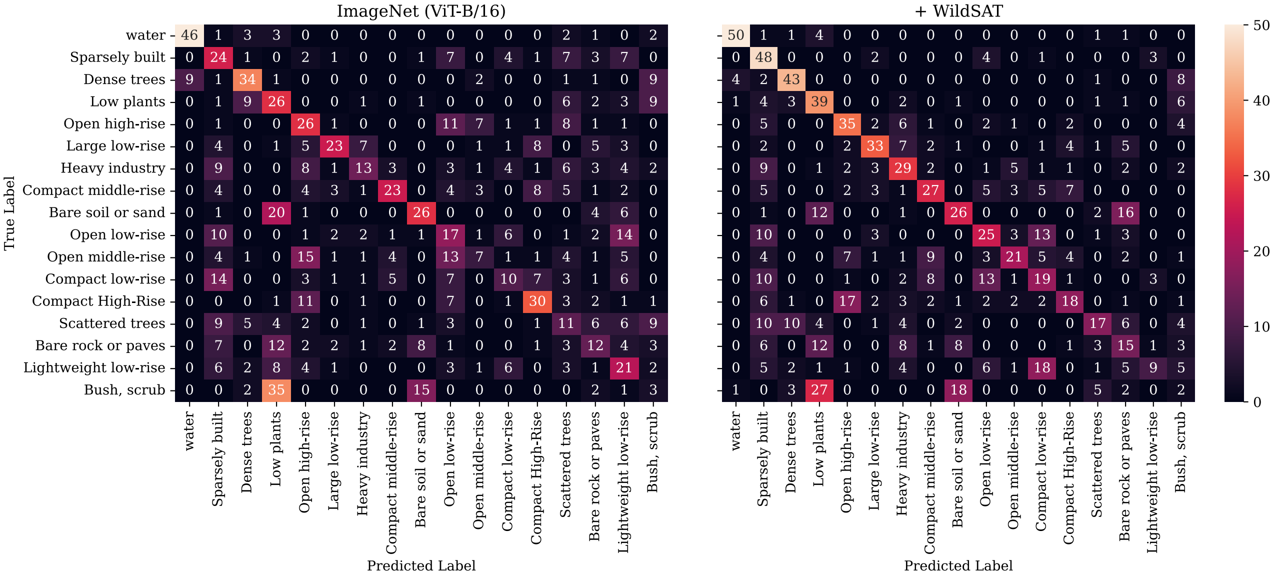

Figure A2. **Confusion matrices comparing the predictions of an ImageNet base model and a WildSAT fine-tuned model on the So2Sat20k dataset [84]**. Both models use a ViT-B/16 architecture. Each matrix displays the result on the provided So2Sat20k test set in GEO-Bench [38].

	EuroSAT [28]			So2Sat20k [84]		
	AnnualCrop	SeaLake	Highway	Water	Low plants	Heavy industry
TaxaBind [63]	1.3	20.4	3.7	0.0	0.0	0.0
GRAFT [47]	33.3	72.1	**55.6**	16.4	0.0	**13.4**
RemoteCLIP [41]	19.6	15.9	5.3	36.5	5.9	0.0
WildSAT (Ours)	**46.5**	**87.4**	0.4	**60.9**	**10.4**	0.0

Table A5. **Zero-shot image classification F1 score on different classes of EuroSAT and So2Sat20k with CLIP-based models**. Since WildSAT is geared towards habitat-related classes, the coverage of zero-shot classification is less effective beyond natural concepts. WildSAT does well on natural classes like 'SeaLake' and 'Water', but struggles on anthropogenic classes like 'Highway' and 'Heavy industry'.

WildSAT outperforms CLIP-based models. Tab. A3 displays the result for each evaluation dataset across different CLIP-based models. All models in the table starts with a pre-trained CLIP ViT-B/16 model [59]. TaxaBind[63] and GRAFT [47] use additional modalities such as ground images, text, and audio to improve model performance on cross-modal tasks such as zero-shot image-text retrieval. However, we show that while these same models do well on zero-shot tasks, they tend to "forget" some of the original image representations, with linear probing performance on downstream datasets lower than that of the original CLIP model. With WildSAT applied to CLIP, we show that we can outperform not only other CLIP-based models, *i.e*. GRAFT and TaxaBind, but also outperform the standard CLIP model across all the datasets in the evaluation. We hypothesize we can prevent "forgetting" by applying parameter efficient fine-tuning on out-of-domain pre-trained models such as CLIP. We further show this in Tab. A9.

WildSAT reduces errors on habitat-related classes. Fig. A2 shows the confusion matrix of a sample result on the So2Sat20k test set. The matrix on the left shows the result of a base ImageNet pre-trained model, and the right matrix shows the result when WildSAT is applied. We show that WildSAT improvements on the true positive counts along the diagonal are largely due to fewer false positives on habitat-related classes. Looking at the second row of both matrices (class 'Sparsely built'), the true positive count doubled from 24 to 48. A lot of this improvement comes from less false positives on 'Scattered trees' (from 7 false positives to 0), 'Bare rock or paves' (from 3 false positives to 0), and 'Dense trees' (from 1 false positive to 0)—all of which are habitat-related attributes. Similar trends can be observed on other classes as well such as in 'Dense Trees' and 'Low Plants' where we see higher true positive counts with WildSAT.

	Cashew1k [79]				SAcrop3k [3]			
	Accuracy		IoU		Accuracy		IoU	
	Base	+WS	Base	+WS	Base	+WS	Base	+WS
ImageNet [14]	91.6%	91.9%	70.3%	70.6%	60.7%	62.3%	24.3%	25.0%
MoCov3 [7]	92.4%	93.2%	71.4%	73.3%	60.7%	60.8%	22.9%	24.9%
SeCo [48]	86.7%	93.2%	62.6%	73.3%	59.2%	59.4%	22.3%	22.8%
SatlasNet [5]	82.5%	91.8%	55.2%	71.0%	56.1%	57.1%	19.4%	20.5%
Random	69.8%	92.9%	40.1%	72.6%	54.7%	56.3%	18.0%	20.3%
	84.6%	**92.6%**	59.9%	**72.2%**	58.3%	**59.2%**	21.4%	**22.7%**

Table A6. **Downstream satellite image segmentation results**. WildSAT can also improve on satellite segmentation tasks across different models and datasets. All models use the frozen pre-trained encoders with a convolutional-based decoder trained for each of the downstream tasks. 'Base' refers to the original models specified as the encoder and '+WS' refers to the same models further trained on the species observation data.

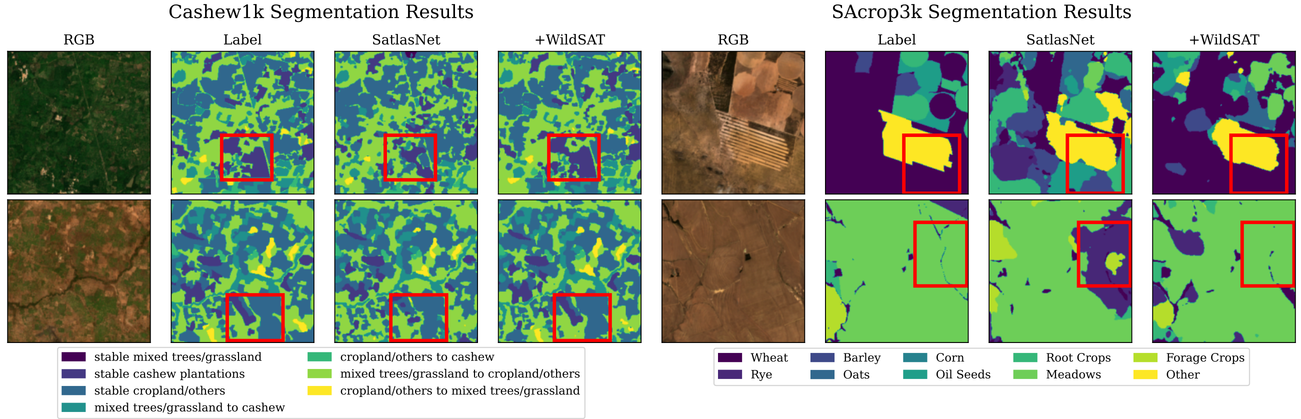
Cashew1k Segmentation Results
RGB
Label
SatlasNet
+WildSAT
SAcrop3k Segmentation Results
RGB
Label
SatlasNet
+WildSAT
stable mixed trees/grassland
stable cashew plantations
stable cropland/others
mixed trees/grassland to cashew
cropland/others to cashew
mixed trees/grassland to cropland/others
cropland/others to mixed trees/grassland
Wheat
Rye
Barley
Oats
Corn
Oil Seeds
Root Crops
Meadows
Forage Crops
Other

Figure A3. **Visualization of segmentation results using SatlasNet without and with WildSAT**. WildSAT can more accurately identify classes 'stable cashew plantations' and 'stable cropland/others' in Cashew1k [79], and improve on the identification of 'Rye' and 'Meadows' in SAcrop3k [3]. Some areas of improvement are highlighted in red boxes.

WildSAT is effective on zero-shot classification of habitat-related classes. Tab. A5 shows zero-shot image classification results on classes from EuroSAT and So2Sat20k. WildSAT is geared towards classes related to nature and animal habitats, and thus see improvements in those classes (*e.g*. 'AnnualCrop', 'SeaLake', 'Water', 'Low plants'). While WildSAT underperforms on anthropogenic classes (*e.g*. 'Highway', 'Heavy industry') compared to models like TaxaBind and GRAFT, future work can integrate wildlife data and anthropogenic labels to create a more comprehensive and balanced representations.

Satellite image segmentation results. Tab. A6 shows results of applying different ResNet50 encoders on satellite image segmentation. For each encoder, a convolution-based decoder is added while adopting the U-Net architecture [61]. To accurately evaluate the features learned with and without WildSAT, we freeze the encoder and fine-tune only the decoder on the downstream dataset. Tab. A6 shows that WildSAT can provide rich representations leading to better segmentation performance. Fig. A3 visualizes how WildSAT improves the segmentation outputs on Cashew1k and SAcrop3k. Further training on WildSAT leads to more accurate differentiation between the crop types.

C.2. Additional Ablations

WildSAT improves models by covering model-specific gaps. Tab. A7 displays an ablation study conducted on two different types of models: a ResNet50 SeCo [48] model and a ViT-B/16 ImageNet [14] model. SeCo is pre-trained with a contrastive objective on time augmented satellite images (*i.e*. satellite images from the same location, but from different seasons). This objective is similar to the $\mathcal{L}_{\text{img}}$ term (Eqn. 1 in the main paper) in the loss function of our WildSAT framework. Thus, we see from Tab. A7, that simply adding the image augmentation term ('img-a') only slightly improves the average performance across the downstream datasets (from 70.1% to 71.8%). This small improvement could be attributed to the additional examples related to habitats that are possibly not as well-represented in the SeCo dataset. However, if we add other modalities such as text and location (in addition to the satellite image augmentation), we see a larger improvement with

ResNet50 SeCo [48]

Loss Terms	loc	env	text	img-a	UCM [78]	AID [76]	RESISC45 [8]	So2Sat20k [84]	Average
Base Model					86.1%	74.3%	80.2%	39.9%	70.1%
$\mathcal{L}_{\text{loc}}$	✓				84.0%	76.2%	81.1%	43.5%	71.2%
$\mathcal{L}_{\text{loc}}$	✓	✓			84.1%	76.3%	83.0%	38.7%	70.5%
$\mathcal{L}_{\text{txt}}$			✓		82.8%	74.4%	79.6%	39.5%	69.1%
$\mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓		✓		84.3%	72.7%	78.5%	41.3%	69.2%
$\mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓	✓	✓		84.0%	75.8%	81.8%	40.0%	70.4%
$\mathcal{L}_{\text{img}}$				✓	83.3%	75.0%	82.9%	46.0%	71.8%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{loc}}$	✓			✓	85.7%	77.9%	86.7%	46.2%	74.2%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{loc}}$	✓	✓		✓	85.3%	77.1%	85.7%	48.2%	74.0%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}$			✓	✓	86.6%	77.0%	84.7%	48.6%	74.2%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓		✓	✓	86.8%	78.1%	86.0%	49.0%	75.0%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓	✓	✓	✓	88.8%	79.6%	86.3%	46.0%	**75.2%**

ViT-B/16 ImageNet [14]

Loss Terms	loc	env	text	img-a	UCM [78]	AID [76]	RESISC45 [8]	So2Sat20k [84]	Average
Base Model					93.2%	84.4%	88.2%	41.8%	76.9%
$\mathcal{L}_{\text{loc}}$	✓				95.2%	85.8%	89.3%	43.4%	78.4%
$\mathcal{L}_{\text{loc}}$	✓	✓			94.7%	86.2%	88.8%	44.2%	78.5%
$\mathcal{L}_{\text{txt}}$			✓		96.1%	85.1%	88.9%	42.2%	78.1%
$\mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓		✓		95.6%	86.5%	89.6%	39.6%	77.8%
$\mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓	✓	✓		95.1%	86.2%	89.7%	45.0%	79.0%
$\mathcal{L}_{\text{img}}$				✓	97.1%	87.1%	91.5%	53.7%	82.3%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{loc}}$	✓			✓	97.1%	88.1%	91.7%	54.0%	82.7%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{loc}}$	✓	✓		✓	96.8%	88.6%	92.1%	52.7%	82.6%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}$			✓	✓	96.9%	87.7%	92.0%	54.3%	82.7%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓		✓	✓	97.1%	87.9%	91.9%	53.4%	82.6%
$\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{loc}}$	✓	✓	✓	✓	97.5%	88.9%	93.0%	55.2%	**83.6%**

Table A7. **Ablation of the various components of the WildSAT framework.** We ablate on SeCo [48], a self-supervised pre-training method that applies contrastive learning on seasonal augmentations of images, and on ImageNet [14], a supervised pre-training on ImageNet. Through the different modalities in WildSAT, we can improve model-specific gaps.

an average performance of 74.2% and 75.2%, respectively. In contrast, an ImageNet pre-trained model benefits from satellite image augmentations ($\mathcal{L}_{\text{img}}$ or 'img-a') since it was trained on a different domain. Simply adding the image augmentation term improved average performance from 76.9% to 82.3%. Further adding other modalities such as text and location also pushes the performance higher to 83.6%. These results support our hypothesis that WildSAT can complement and further improve different architectures by leveraging the different modalities.

Location encoder ablation. Tab. A8 shows an ablation study conducted on our choice of the location encoder. We use a ResNet50 SeCo encoder as the base model, and report accuracy on downstream classification datasets. All rows in the table use WildSAT with satellite images and text ($\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}$), which are matched based on the location—*i.e.* location is implicitly used in all the results, and we ablate which encoder to use for explicitly including location as an input. We replace the location encoder in our WildSAT framework with one of the following: no model (*i.e.* use the position encoded latitude and longitude and/or the raw environmental covariates vector), SatCLIP [35], or SINR [12]. We use the SatCLIP pre-trained location encoder that takes the latitude and longitude as an input. For SINR, we explore the two variants of using (1) just the location ('loc') or (2) both the location ('loc') and environmental covariates ('env'). We find that the best average performance uses SINR with both the location and environmental covariates.

ResNet50 SeCo [48]									
No Model		SatCLIP	SINR		UCM	AID	RESISC45	So2Sat20k	Average
loc	env	loc	loc	env	[78]	[76]	[8]	[84]	
					86.6%	77.0%	84.7%	48.6%	74.2%
✓					87.3%	78.1%	85.5%	48.3%	74.8%
	✓				86.3%	76.3%	84.9%	47.2%	73.7%
✓	✓				87.0%	77.4%	85.1%	48.4%	74.5%
		✓			86.0%	78.2%	85.5%	50.0%	74.9%
			✓		86.8%	78.1%	86.0%	49.0%	75.0%
			✓	✓	88.8%	79.6%	86.3%	46.0%	**75.2%**

Table A8. **Ablation of the location encoder.** These runs assume both the text ($\mathcal{L}_{\text{txt}}$) and the image augmentation($\mathcal{L}_{\text{img}}$) are already part of the model, which implicitly uses location (since images and text are matched based on location). We ablate the explicit addition of location as an input through different location encoders. We explore using no model (*i.e.* directly just using the latitude/longitude or environmental covariates), SatCLIP [35], and SINR [12]. The last row of the table corresponds to our WildSAT setup.

	Encoder	PEFT	UCM [78]	AID [76]	RESISC45 [8]	FMoW [9]	EuroSAT [28]	So2Sat20k [84]	BEN20k [65]
ResNet50	ImageNet1K [60]		91.3%	82.0%	85.6%	42.1%	95.0%	47.0%	56.1%
ResNet50	ImageNet1K [60]	✓	93.6%	86.7%	90.1%	46.0%	96.0%	46.6%	57.5%
ViT-B/16	CLIP [59]		82.1%	71.0%	75.3%	34.9%	93.4%	50.4%	49.0%
ViT-B/16	CLIP [59]	✓	96.3%	88.0%	93.0%	53.6%	97.1%	49.7%	59.1%
ResNet50	SatlasNet [5]		90.1%	79.4%	85.4%	42.4%	95.4%	44.8%	56.4%
ResNet50	SatlasNet [5]	✓	86.9%	76.8%	82.0%	35.7%	94.1%	41.6%	51.6%
ResNet50	SeCo [48]		88.8%	79.6%	86.3%	42.8%	95.5%	46.0%	57.3%
ResNet50	SeCo [48]	✓	86.7%	77.3%	83.2%	37.3%	94.0%	44.4%	54.8%

Table A9. **Ablation of parameter efficient fine-tuning (PEFT) when applied with WildSAT.** Models pre-trained on out-of-domain datasets (*e.g.* ImageNet, CLIP) that are fine-tuned with PEFT can perform better on downstream tasks by preserving original representations from the base model. In contrast, models pre-trained on in-domain datasets (*e.g.* SatlasNet, SeCo) show limited improvement from PEFT since the fine-tuning is in the same domain as the pre-training (*i.e.* satellite images)—fine-tuning all layers has better performance.

Parameter efficient fine-tuning (PEFT) preserves out-of-domain pre-training representations. Tab. A9 displays the effect of fine-tuning all parameters of a given base model compared to fine-tuning specific layers (*i.e.* applying PEFT). We compare the effect on out-of-domain pre-trained models (*e.g.* ImageNet [60], CLIP [59]), and in-domain pre-trained models (*e.g.* SatlasNet [5], SeCo [48]). We find that out-of-domain pre-trained models have better downstream performance by applying scale and shift fine-tuning [21, 40], or by applying DoRa [49]. By fine-tuning specific layers, the models retain some of the original representations learned from the pre-training (*i.e.* ImageNet or CLIP) so that performance does not deteriorate compared to the base models. This has a significant impact on large models such as ViT, since fine-tuning all weights alters many parameters and risks shifting them in suboptimal directions. On the other hand, while applying PEFT for in-domain pre-trained models SatlasNet [5] and SeCo [48] improves performance compared to the base model, we see better performance when directly fine-tuning all the layers. This may be because the model has already undergone pre-training on satellite images, making additional pre-training on similar data from WildSAT result in a non-disruptive shift.

C.3. Zero-shot Retrieval

In Fig. A4, we display more zero-shot retrieval examples. The first row of examples demonstrates retrieval of general landscapes such as 'rainforest' or 'mountains'. The second row demonstrates retrieval of wildlife habitats. We enumerate each of the wildlife examples below including their expected habitats. All the enumerated habitats are consistent with the retrieved satellite images.

Description of the wildlife examples from the second set of rows in Fig. A4:

1. 'house sparrow' is a small, common bird typically found in urban areas.
2. 'albatross' is a large bird commonly found in the sea.
3. 'sandpiper' is a small bird that dwells in the coast.
4. 'horned lark' is a bird species found in open land such as on farmland, on prairies, and in deserts.

5. ‘cactus’ is a type of plant commonly found in the desert.
6. ‘rock pigeon’ is a bird commonly found in urban and residential areas.
7. ‘virginia rail’ is a bird found in freshwater and brackish marshes, and sometimes salt marshes in winter.
8. ‘american marten’ is a North American mammal that is found in forests, and broadly distributed in North America from Alaska and Canada to New York.

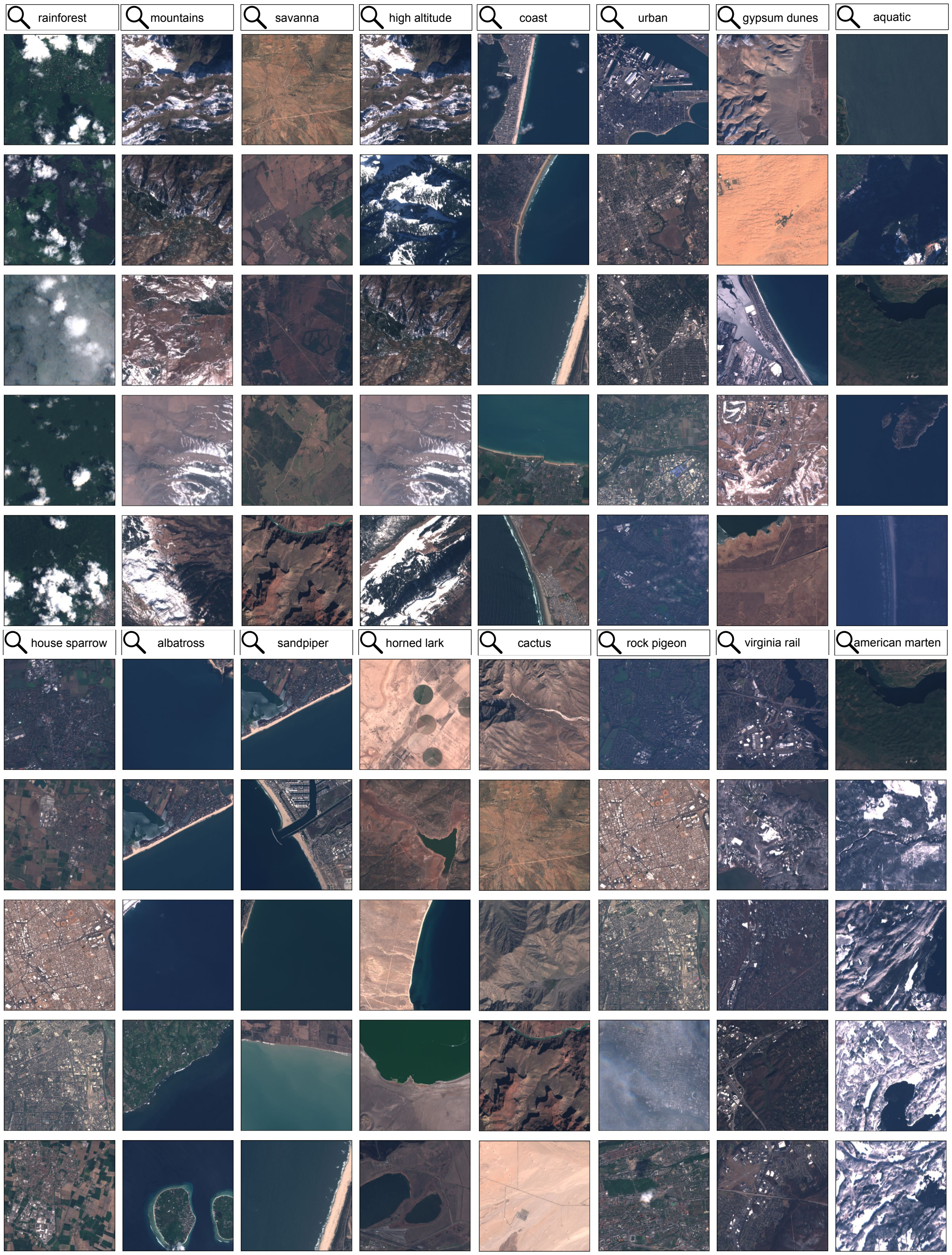

Figure A4. **Additional zero-shot results for text-based satellite image retrieval**.