Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters

Isaac Corley*
University of Texas at San Antonio
San Antonio, TX, USA
isaac.corley@utsa.edu

Caleb Robinson*
Microsoft AI for Good Research Lab
Redmond, WA, USA
caleb.robinson@microsoft.com

Rahul Dodhia
Microsoft AI for Good Research Lab
Redmond, WA, USA
rahul.dodhia@microsoft.com

Juan M. Lavista Ferres
Microsoft AI for Good Research Lab
Redmond, WA, USA
jlavista@microsoft.com

Peyman Najafirad
University of Texas at San Antonio
San Antonio, TX, USA
peyman.najafirad@utsa.edu

Abstract

Research in self-supervised learning (SSL) with natural images has progressed rapidly in recent years and is now increasingly being applied to and benchmarked with datasets containing remotely sensed imagery. A common benchmark case is to evaluate SSL pre-trained model embeddings on datasets of remotely sensed imagery with small patch sizes, e.g., 32 × 32 pixels, whereas standard SSL pre-training takes place with larger patch sizes, e.g., 224 × 224. Furthermore, pre-training methods tend to use different image normalization preprocessing steps depending on the dataset. In this paper, we show, across seven satellite and aerial imagery datasets of varying resolution, that by simply following the preprocessing steps used in pre-training (precisely, image sizing and normalization methods), one can achieve significant performance improvements when evaluating the extracted features on downstream tasks – an important detail overlooked in previous work in this space. We show that by following these steps, ImageNet pre-training remains a competitive baseline for satellite imagery based transfer learning tasks – for example we find that these steps give +32.28 to overall accuracy on the So2Sat random split dataset and +11.16 on the EuroSAT dataset. Finally, we report comprehensive benchmark results with a variety of simple baseline methods for each of the seven datasets, forming an initial benchmark suite for remote sensing imagery.[1]

1. Introduction

With increasing frequency, self-supervised learning (SSL) models, foundation models, and transfer learning methods have been applied to remotely sensed imagery [6, 10, 11, 18, 19, 25, 31, 33, 35, 36, 39, 40, 42, 49, 53, 55, 56]. As such, rigorous benchmarks are needed to identify the strengths and weaknesses in the proposed methods.

A commonly used benchmark in any transfer learning setup is the use of embeddings from a model that is pre-trained on the ImageNet (ILSVRC2012) dataset [13] – due to both the ease of implementation [9, 34] and strong performance when generalizing to unseen data [27]. However, even with fully convolutional neural networks, the size of image inputs to the model is an important factor that should be controlled for at test/inference time. Common large-scale benchmarks libraries like PyTorch Image Models (timm) [57] and OpenCLIP [28] provide benchmark results trained at varying image sizes and evaluate at the same sizes as opposed to the original dataset size. Plainly put, models that are pretrained on ImageNet images that have been resized and cropped to a fixed image size (traditionally 224 × 224 or 256 × 256), will produce the most relevant embeddings for transfer learning when they are given the same

*Equal contribution

[1] Experimental code, datasets, and model checkpoints are available at github.com/isaaccorley/resize-is-all-you-need

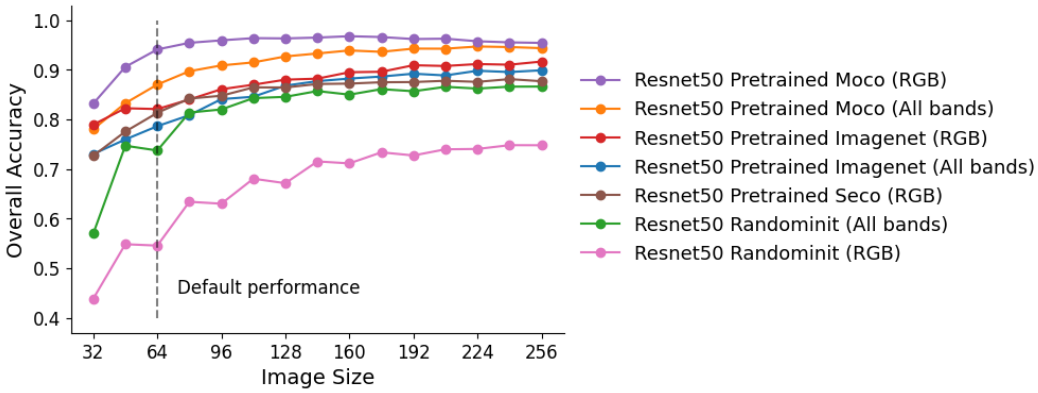

Figure 1. The effect of input image size on EuroSAT downstream performance (overall accuracy) across different ResNet models. By default, EuroSAT images are 64 × 64 pixels, however resizing to larger image sizes before embedding increases downstream accuracy under a KNN ($k = 5$) classification model in all cases.

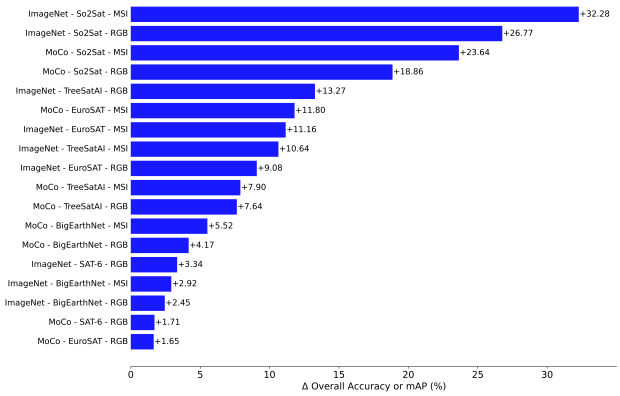

Figure 2. Difference in downstream task metrics, Overall Accuracy (OA) (multiclass) or mean Average Precision (mAP) (multilabel), after resizing images to 224 × 224 from the original, smaller, image size. ImageNet pre-trained models, for example, often are trained with 224 x 224 inputs and therefore do not produce useful embeddings with smaller image patches.

image size at test time.

Satellite missions such as Sentinel-2 [15] and Landsat-8 [45] capture imagery over the Earth's surface at relatively low spatial resolutions, e.g. 10-60 meters/pixel, compared to the resolution of objects in natural imagery. Because of this, it is common for labeled datasets of remotely sensed imagery to contain images of smaller sizes, e.g. 32 × 32 [59], than traditional image classification datasets. Thus, if images from these datasets are used as-is with ImageNet pretrained models, then the results will be sub-optimal.

A similar story can be told with image normalization methods. A standard preprocessing method for ImageNet pre-trained models is to normalize all values in an image to a $[0, 1]$ range then perform channel-wise standardization with ImageNet statistics. However, as remotely sensed imagery usually has a higher bit-depth (or color-depth) than images in standard vision datasets (12 or 16-bit depth vs. 8-bit depth), different image normalizations methods are usually applied. For example, a common method used with Sentinel-2 imagery is to divide all values by 10,000 (to convert the raw sensor values to reflectance values) then use these as inputs in a network [35, 56]. If images that are normalized with one method are used with a network that is pre-trained under a different normalization method, then the results will also be sub-optimal.

We demonstrate that it is vital to consider how an embedding model was trained when using it for transfer learning on downstream remote sensing tasks. For example, through simple bilinear upsampling of input images from 64 × 64 to 224 × 224 on the EuroSAT RGB dataset [26], we find that accuracy of the embeddings generated by a ImageNet pretrained ResNet-50 [22] increases from 0.82 to 0.91. Similarly, performing a channel-wise standardization instead of re-scaling the image values to represent reflectance results in a performance increase from 0.66 to 0.91 (when combined with resizing to 224 × 224). **Performing these steps correctly gives simple baselines, like ImageNet pre-training, results that are competitive with previously published methods.** Additionally, we benchmark several simple methods, including MOSAIKS [44] and a simple image statistic based feature extraction method, and find that they beat ImageNet and/or remote sensing SSL pretraining methods on several datasets.

While not particularly surprising, our results form a set of strong baselines that can be used to benchmark future methods for self-supervised learning with remotely sensed imagery against. Further, our experimental setup is open-sourced and can be easily appended to as the community focuses on different geospatial machine learning tasks.

Table 1. Results on the EuroSAT dataset [26] for multiclass classification using KNN ($k = 5$). We report Overall Accuracy (OA) for both RGB and all MSI bands. We compare to fine-tuned performance of several SSL methods taken from their respective papers. *The Scale-MAE result uses a KNN-5 and is comparable to the other KNN results.

Model	Weights	Size	RGB	MSI
ResNet50	MoCo	64	*94.11*	81.85
		224	**95.76**	**93.65**
ResNet50	ImageNet	64	82.09	78.65
		224	91.17	89.81
ResNet50	Random	64	59.92 ± 0.34	75.10 ± 0.23
		224	73.76 ± 0.53	87.19 ± 0.81
RCF	Random	64	78.85 ± 0.33	87.56 ± 0.35
		224	76.90 ± 0.33	87.41 ± 0.12
RCF	Empirical	64	81.47 ± 0.08	*91.10 ± 0.11*
		224	77.88 ± 0.08	90.14 ± 0.15
Image Stat.	-	64	76.94	89.56
ViT-L	Scale-MAE [42]	64	96.00*	-
ResNet18	GASSL [2]	64	89.51	-
ResNet18	SeCo [35]	64	93.14	-
ViT-L	SatMAE [10]	224	98.94	-

Table 2. Results on the SAT-6 dataset [3] for multiclass classification using KNN ($k = 5$). We report Overall Accuracy (OA) and compare to the fully-supervised performance of DeepSAT and DeepSATv2 models taken from their respective papers.

Model	Weights	Size	OA
ResNet50	MoCo	34	98.15
		224	*99.86*
ResNet50	ImageNet	34	96.55
		224	**99.89**
ResNet50	Random	34	91.64 ± 0.66
		224	98.57 ± 0.08
RCF	Random	34	99.40 ± 0.06
		224	99.29 ± 0.07
RCF	Empirical	34	99.65 ± 0.02
		224	98.85 ± 0.06
Image Stat.	-	28	99.60
DeepSat [3]	Sup.	28	93.92
DeepSatv2 [32]	Sup.	28	99.84

Our main contributions are as follows:

- We propose a set of strong baseline methods for remote sensing scene classification – including an ImageNet pretrained ResNet-50, random convolutional features (RCF), and a simple image statistic feature extraction method – that outperform self-supervised pretrained models on several datasets. We have implemented these methods into the open source TorchGeo library [46].
- We present a set of benchmark results across seven geospatial machine learning datasets commonly used as downstream tasks for testing pre-trained model performance with our baseline methods.
- We demonstrate the importance of proper resizing and normalization of images for optimal performance and fair comparisons in geospatial machine learning benchmarks.

1.1. Related Work

Recent works have shown that while many new deep learning architectures claim to achieve state-of-the-art performance due to their proposed novel model design, they in fact only do so because of inconsistencies in training strategies and hyperparameters when comparing to baselines and prior methods. Bello et al. [4] explored that by simply retraining with recent training techniques and tricks, the original ResNet [22] architecture significantly outperforms its own previous baselines and reaches a competitive top-1 ImageNet accuracy. Du et al. [16] concluded the same findings for 3D ResNets [52] for video recognition tasks. Goyal et al. [21] examined the similar effects for numerous architectures in the 3D point cloud classification field. Finally, Musgrave et al. [37] repeat the same idea for metric learning methods. In other words, when all models are on the same playing field, performance gains from past methods over strong baselines tend to become insignificant.

Previous papers that explore the effect of resizing inputs on performance in convolutional neural networks include Richter et al. [43] and Touvron et al. [51]. Both papers investigate different experimental setups by varying training and testing at different image sizes and empirically show that increasing the image size during inference improves performance which begins to saturate around an image size of 256×256. However, both works strictly explore natural images only with ImageNet pretraining as opposed to remotely sensed imagery, as is the objective of this paper. Wang et al. [56] provide the closest evidence of this case for remote sensing data by performing a short experiment reporting linear probing results showing a boost in performance while increasing the input image size.

2. Methods

In this study we extract feature representations (or embeddings) from remotely sensed image datasets using a variety of methods (described below) while varying the image preprocessing steps. Specifically, we vary the image size that is passed through to the feature extractor using Pytorch's [41] `torch.nn.functional.interpolate` implementa-

Table 3. Results on the So2Sat dataset [59] for multiclass classification using KNN ($k = 5$). We report Overall Accuracy (OA) for both RGB and all MSI bands and for both the *Random* and *Culture-10* splits. We compare to both fully-supervised and linear probing results for several SSL methods.

			Random		Culture-10	
Model	**Weights**	**Size**	**RGB**	**MSI**	**RGB**	**MSI**
ResNet50	MoCo	34	75.07	72.51	51.45	49.36
		224	**93.93**	**96.15**	**56.03**	**53.54**
ResNet50	ImageNet	34	66.21	56.18	47.76	42.11
		224	*92.99*	88.46	*54.53*	*50.32*
ResNet50	Random	34	46.19 ± 0.19	55.06 ± 0.35	29.10 ± 0.30	35.47 ± 0.18
		224	71.74 ± 1.87	84.10 ± 0.32	34.16 ± 0.23	45.68 ± 0.50
RCF	Random	34	72.67 ± 0.45	89.40 ± 0.14	30.92 ± 0.11	45.23 ± 0.33
		224	74.22 ± 0.44	89.72 ± 0.11	31.19 ± 0.21	45.36 ± 0.36
RCF	Empirical	34	71.00 ± 0.32	*95.37 ± 0.06*	35.32 ± 0.45	47.63 ± 0.10
		224	51.66 ± 0.46	95.20 ± 0.02	27.36 ± 0.24	44.98 ± 0.16
Image Stat.	-	32	83.84	91.09	38.36	47.93
ResNet50	MoCo [56]	224	-	-	-	61.80
ResNet50	DINO [5]	224	-	-	-	57.00
ViT-S	DINO [5]	224	-	-	-	62.50
ViT-S	MAE [24]	224	-	-	-	60.00
ResNet50	Sup. [56]	224	-	-	-	57.50
ViT-S	Sup. [56]	224	-	-	-	59.30

tion with bilinear interpolation, and we vary the image normalization method between channel-wise standardization (i.e. the default practice for most ImageNet pretrained models), converting the input image values into a reflectance value (i.e. the default practice for most Sentinel-2 pretrained models), min-max normalization, or method specific normalizations (e.g. the percentile normalization from [35]). In datasets that have multispectral information we run experiments using only the RGB channels, as well as all the channels (MSI)[2].

We extract feature representations using the following methods:

ResNet-50 Random init. [22] A vanilla ResNet-50 with random weight initialization (following the default torchvision settings). The features generated by this and the following two ResNet-50 models are produced by the final global average pool operation and are 2048-dimensional.

ResNet-50 ImageNet [13] A ResNet-50 that is pretrained on ImageNet with images of size 224 × 224 (default torchvision pretrained weights).

ResNet-50 SSL4EO [56] A ResNet-50 that is pretrained using the MoCo-v2 [7, 23] self-supervised learning method on the SSL4EO dataset with 224x224 images.

RCF (Random) [44] A feature extraction method that consists of projecting the input to a lower dimensional space using random convolutional features (RCF). We use the implementation from TorchGeo with 512 convolutional filters and a 3x3 kernel size. In the results we refer to this method as RCF with random weights.

MOSAIKS / RCF (Empirical) [44] A feature extraction method similar to RCF but that initializes the weights using ZCA whitened patches sampled randomly from the training set. We use the implementation from TorchGeo with 512 convolutional filters and a 3x3 kernel size. In the results we refer to this method as RCF with empirical weights.

Image Statistics A hand crafted baseline method that consists of simply computing per-channel pixel statistics from the imagery. Given an image we compute the mean, standard deviation, minimum, and maximum value for each band and concatenate these into a simple $4c$-dimensional feature representation, where c is the number of input channels.

2.1. Evaluation

For evaluating the representation performance of a pretrained model it is common to perform "linear probing" on a given

[2]Note that for processing multispectral (MSI) imagery through ImageNet pretrained ResNets, we repeat the RGB weights in the first convolutional layer to account for the additional input bands. For SSL4EO MSI pretrained ResNets, we zero-pad channels to account for any bands not made available in datasets.

Table 4. Results on the BigEarthNet dataset [47] for 19-class multilabel classification using KNN ($k = 5$). We report overall F1 score, and overall mean average precision (mAP). For reference, we compare to the fully supervised S-CNN as well as fine-tuned results from the GASSL, SeCo, and SatMAE SSL methods.

			RGB		MSI	
Model	**Weights**	**Size**	**F1**	**mAP**	**F1**	**mAP**
ResNet50	MoCo	120	*68.99*	*70.65*	63.61	64.64
		224	**72.56**	**74.81**	68.33	70.17
ResNet50	ImageNet	120	65.38	66.62	62.61	62.96
		224	67.47	69.07	65.04	65.88
ResNet50	Random	120	52.34 ± 0.22	52.63 ± 0.19	60.48 ± 0.34	61.17 ± 0.50
		224	57.05 ± 1.02	57.61 ± 1.13	64.94 ± 0.25	66.31 ± 0.32
RCF	Random	120	54.48 ± 0.26	53.94 ± 0.26	69.98 ± 0.20	72.01 ± 0.28
		224	54.37 ± 0.28	53.74 ± 0.23	70.06 ± 0.21	72.12 ± 0.29
RCF	Empirical	120	57.40 ± 0.22	57.22 ± 0.23	$\mathit{73.31 \pm 0.14}$	$\mathit{76.18 \pm 0.19}$
		224	53.36 ± 0.23	52.90 ± 0.22	$\mathbf{73.41 \pm 0.13}$	$\mathbf{76.29 \pm 0.15}$
Image Stat.	-	120	61.67	62.00	69.42	71.29
S-CNN	BigEarthNet [47]	120	67.59	-	70.98	-
ResNet50	GASSL [2]	120	-	80.20	-	-
ResNet50	SeCo [35]	120	-	82.62	-	-
ViT-L	SatMAE [10]	224	-	82.13	-	-

downstream task by training a linear model on the representations generated by the pre-trained model and measuring the performance of this linear model. However, this method is implemented very differently between papers – some papers use data augmentation [56] while others don't, and others use a variety of different optimizers (SGD, Adam, LARS), regularization methods[3], and learning rates / learning rate schedules. Therefore, for fair evaluation we fit a K-Nearest-Neighbors (KNN) model [12] to extracted features from various datasets, setting $k = 5$, as performed similarly in [42, 53].

3. Datasets

The datasets used throughout our experiments were selected particularly due to their original image sizes being small to show the effects of resizing. These datasets are commonly benchmarked without resizing which makes them perfect candidates for quantifying the effects of size vs performance. We also select datasets which are from both low-resolution satellite sources as well as high resolution aerial imagery.

EuroSAT The EuroSAT dataset [26] is a land cover classification dataset of patches extracted from MSI Sentinel-2 [15] imagery. The dataset contains 27,000 64 × 64 10m spatial resolution images with 13 bands and labels for 10 land cover categories. We use the dataset splits defined in Neumann et al. [38].

[3]For example, by default the Adam optimizer in PyTorch will not apply L2 regularization on the weights of the model (weight decay), while scikit-learn linear models are trained with L2 regularization by default.

SAT-6 The SAT-6 dataset [3] is a land cover classification dataset of patches extracted from aerial imagery from the National Agriculture Imagery Program (NAIP) [17]. The dataset contains 405,000 28 × 28 RGBN patches at 1m spatial resolution and labels for 6 land cover categories. We use the train and test splits provided with the dataset.

So2Sat The So2Sat dataset [59] is a local climate zone (LCZ) classification dataset of patches extracted from Sentinel-1 and Sentinel-2 imagery. For our experiments we only utilize the Sentinel-2 bands. The dataset contains 400,673 MSI patches with 10 bands and at 10m spatial resolution. Each patch is of size 32 × 32 and contains a single label from 17 total LCZ categories. We use the train and test splits from the Random and Culture-10 sets provided with the dataset.

BigEarthNet The BigEarthNet dataset [47] is a multi-label land cover classification dataset of patches extracted from MSI Sentinel-2 imagery. The dataset contains 590,326 120 × 120 10m spatial resolution images with 12 bands and labels for 19 land cover categories. We use the splits provided with the dataset and defined in [48].

TreeSatAI The TreeSatAI dataset [1] is a multi-sensor, multilabel tree species classification dataset of patches extracted from aerial and MSI Sentinel-1 [50] and Sentinel-2 imagery. For our experiments we only utilize the Sentinel-2 bands. The dataset contains 50,381 10m spatial resolution images with 12 spectral bands, which

Table 5. Results on the TreeSatAI dataset [1] for multilabel classification using KNN ($k = 5$). We report overall F1 score and mean average precision mAP. We compare to the fully-supervised LightGBM performance and fine-tuned Presto SSL method.

			RGB		MSI	
Model	**Weights**	**Size**	**F1**	**mAP**	**F1**	**mAP**
ResNet50	MoCo	34	29.21	29.93	37.65	36.24
		224	37.68	*37.57*	45.18	44.14
ResNet50	ImageNet	34	27.69	27.30	32.07	30.69
		224	**40.37**	**40.58**	42.00	41.33
ResNet50	Random	34	29.37 ± 0.42	29.08 ± 0.18	36.47 ± 0.34	34.73 ± 0.15
		224	35.42 ± 0.33	34.75 ± 0.43	49.09 ± 0.83	48.48 ± 0.89
RCF	Random	34	33.15 ± 0.21	32.15 ± 0.09	52.24 ± 0.35	51.83 ± 0.33
		224	32.37 ± 0.20	31.29 ± 0.18	52.49 ± 0.17	51.99 ± 0.43
RCF	Empirical	34	31.70 ± 0.06	31.13 ± 0.17	$\mathbf{56.00 \pm 0.04}$	$\mathbf{56.08 \pm 0.25}$
		224	28.93 ± 0.47	28.50 ± 0.23	$\mathit{55.60 \pm 0.13}$	$\mathit{55.77 \pm 0.29}$
Image Stat.	-	20	*38.39*	37.19	51.97	51.56
LightGBM [29]	-	20	-	-	52.52	61.66
ViT	Presto [53]	9	-	-	50.32	67.78

are available in 6 × 6 or 20 × 20 sizes, and labels for 20 tree species categories. We use the train and test splits provided with the dataset.

UC Merced The UC Merced (UCM) dataset [58] is a land use classification dataset that consists of 2,100 256 × 256 pixel aerial RGB images over 21 target classes. We use the train/val/test splits defined in Neumann et al. [38].

RESISC45 The RESISC45 dataset [8] is a scene classification dataset that consists of 45 scene classes and 31,500 256 × 256 pixel aerial RGB images extracted from Google Earth. We use the dataset splits defined in Neumann et al. [38].

4. Results and Discussion

4.1. Fair Comparisons to ImageNet Pretraining

As stated in Section 1.1, prior research has shown the significance of resizing images during testing for ImageNet pretrained models. To emphasize this, we perform a short experiment comparing features extracted from the EuroSAT [26] dataset using a ResNet-18 pretrained with both the Seasonal Contrast (SeCo) method [35] and ImageNet. For fair evaluation, we compute downstream task results at the original image size 64 × 64 and resized to 224 × 224 with KNN and linear probe methods.

For linear probing we utilize the exact same experimental setup and script as in [35] while only adding a resize transformation. As seen in Table 6, depending on the model used for evaluation, one pretraining method can appear better than another. Furthermore, while increasing the image size improves performance for both methods, it does not improve equally. When reading the linear probing results in [35], one would assume that the SSL pretrained model clearly outperforms ImageNet pretraining. However, as we can see, this is not the case, and further investigation are needed. Further, in Table 8, we observe that an ImageNet pretrained model outperforms the best reported results in SatMAE [10] in the same experimental setup.

Table 6. Comparison of SeCo [35] vs. ImageNet pretraining on the EuroSAT validation set. We show Overall Accuracy results for both KNN and linear probe at different image sizes.

Size	**Weights**	**KNN ($k = 3$)**	**KNN ($k = 10$)**	**Linear Probe**
64	SeCo	84.04	84.11	**93.14**
	ImageNet	**85.39**	**85.20**	86.44
224	SeCo	86.57	85.63	**96.30**
	ImageNet	**90.54**	**90.63**	93.13

4.2. Image Size vs. Performance

Figure 1 shows how the performance of a variety ResNet-50 models varies with input image size on the EuroSAT dataset when using just the RGB bands vs. all spectral bands as input. We observe in all cases that the default dataset image size (64 × 64 pixels) does not result in optimal performance. For example, resizing from 64 × 64 to 256 × 256 results in a 10 point increase in accuracy in a ResNet-50 that is pretrained on ImageNet. In Tables 1-5 we report performance from each method at the native resolution of the dataset and after resizing each image to 224x224 and observe performance improvements across all methods in nearly all cases.

To visualize the effects of resizing (and standard normalization), in Figure 3 we show t-SNE [54] plots of EuroSAT RGB features extracted using a ResNet-50 pretrained on Im-

Table 7. Results on the RESISC45 dataset [8] for multiclass classification using KNN ($k = 5$). We report Overall Accuracy (OA) and compare to performance metrics of various remote sensing SSL methods taken from their respective papers. *The Scale-MAE result uses a KNN-5 and is comparable to the other KNN results.

Model	Weights	Size	OA
ResNet-50	MoCo	256	*73.24*
ResNet-50	ImageNet	256	**77.48**
ResNet-50	Random	256	36.30 ± 0.25
RCF	Random	256	42.29 ± 0.12
RCF	Empirical	256	36.15 ± 0.36
Image Stat.	-	256	34.03
ViT-L	Scale-MAE [42]	256	85.0 *
ViT-L	SatMAE [10]	256	77.1*
ViT-L	ConvMAE [20]	256	78.8*

Table 8. Results on the UC Merced dataset [58] for multiclass classification using KNN ($k = 5$). We report Overall Accuracy (OA) and compare to the linear probing performance of the Scale-MAE, SatMAE, and ConvMAE methods taken from their respective papers. *The Scale-MAE result uses a KNN ($k = 5$) and is comparable to the other KNN results.

Model	Weights	Size	OA
ResNet50	MoCo	256	*85.50*
ResNet50	ImageNet	256	**90.70**
ResNet50	Random	256	47.94 ± 1.07
RCF	Random	256	52.14 ± 0.24
RCF	Empirical	256	56.90 ± 0.63
Image Stat.	-	256	47.90
ViT-L	Scale-MAE [42]	256	85.1*
ViT-L	SatMAE [10]	256	84.2*
ViT-L	ConvMAE [20]	256	81.7*

ageNet. The the plot shows that EuroSAT classes are clearly separable at an input size of 224 x 224 while only partially separable at 32 x 32. Additionally, when resizing but not using any normalization, there are no clear clusters corresponding to the dataset classes. While we use a NVIDIA DGX server with 2x A100 GPUs to increase the speed of our benchmarks, we note that none of these methods actually require a GPU to perform inference or KNN classification on extracted features.

4.3. Benchmarks

We perform thorough benchmarks using the methods described in Section 2 on each dataset from Section 3, using the evaluation metric common to that dataset, in Tables 1 through 8. In each experiment we fit a non-parametric k-nearest neighbor model with $k = 5$ to the train set. For deterministic methods we report a single value calculated over the test set for each dataset, while for stochastic methods we report the average ± the standard deviation of the metric calculated over the test set over 5 runs with different random seeds. We bold the best performing of the baseline methods by column and italicize the second best performing method. Additionally, we show several fine-tuning, linear probing, and fully-supervised baselines from original dataset papers or other SSL remote sensing papers. Note that we perform these comparisons not with the goal of outperforming them but for transparency of the difference in performance in representation ability to the state-of-the-art. Finally, we note that our evaluation method is the same as that of Reed et al. [42] and indicate this with an asterisk where appropriate.

For the EuroSAT experiments we show results from GASSL [2], SeCo [35], and SatMAE [10] self-supervised methods that use fine-tuning on top of the pretrained network (as reported by SatMAE). We note that methods which use a (ViT) [14] model are unable to accept input images with varying sizes and therefore we only report performance from their original training image size.

For the SAT-6 experiments we compare to the performance of the DeepSat [3] model proposed in the original SAT-6 dataset paper as well as the DeepSatv2 [32] model from a follow-up paper.

For the UC Merced experiments, we compare to the performance of SatMAE [10], Scale-MAE [42], and ConvMAE [20] as reported in the Scale-MAE paper.

Our results show the following:

- SSL4EO MoCo-v2 pretrained weights have the best overall performance across downstream tasks. They rank in the top-2 methods by performance for 6 out of the 7 RGB datasets, and 3 out of 5 MSI datasets.
- The Scale-MAE pretrained model performs the best in the EuroSAT and RESISC45 datasets, however is outperformed by ImageNet pretraining in the UCM dataset.
- The image statistic baseline outperforms ImageNet pretrained models on all but one of the MSI datasets (and it is 0.25% lower than ImageNet in this case).
- MOSAIKS (i.e. RCF with empirical weights) is a very strong baseline on the MSI datasets and ranks in the top 2 methods by performance for 4 out of the 5 MSI datasets.
- In SAT-6 experiments, all methods except for the randomly initialized ResNet-50 achieve greater than 99% accuracy. Even the image statistic baseline achieves a 99.6% overall accuracy. This suggests that the dataset is too simple to be used as a benchmark for comparing models as it will be difficult to observe statistically significant changes in accuracy between 99.6% (any result worse than this would suggest a model that is less expressive than simply

Figure 3. t-SNE [54] plots of EuroSAT test set embeddings extracted using a ResNet50 pretrained on ImageNet with different preprocessing. (left to right: 32 × 32 with normalization, 224 × 224 without normalization, 224 × 224 with normalization)

extracting image statistics) and 100%. Nevertheless, future work could explore this dataset in other settings, such as few-shot learning.

- Resizing images does not result in significantly changed downstream performance with the RCF methods (as compared to the ResNet based models). We hypothesize that this method is largely scale invariant – however leave further experiments (such as varying convolutional size with input size, etc.) to future work.
- For 2 out of 5 datasets with MSI bands, adding the additional MSI bands degrades ResNet-50 ImageNet pretrained performance. However, in all cases, adding MSI information increases the ResNet-50 random init. performance. This further highlights the difference in distributions between ImageNet, natural imagery, and remotely sensed imagery.
- In the So2Sat dataset, switching from the Random set to the Culture-10 set decreases the accuracy of RCF methods more than the pre-trained models. We hypothesize that this is because the Culture-10 set tests geographic generalization, and RCF will only be able to use color/texture from the train set while the pre-trained models could potentially group similar patches across sets to similar feature representations.

5. Best Practices

To recap, below is a list of best practices we believe all remote sensing pre-training research should include in their analyses. While these may seem obvious, it is critical to follow these guidelines to produce accurate and transparent benchmarks for understanding the strengths and weaknesses of methods proposed to the community.

1. **Always compare to simple baseline**: Performance across datasets can be misleading, therefore always compare a simple and effective baseline. We recommend an ImageNet pretrained model, random convolutional features, and image statistics.
2. **Resize & Normalize**: Resize and normalize inputs to the same parameters as during training, e.g., when comparing to ImageNet pretrained models, normalize to the range $[0, 1]$, normalize to scale inputs to $\mu = 0$ and $\sigma = 1$, and resize inputs to 224 × 224.
3. **Prefer KNN over Linear Probing and Fine-tuning**: Linear probing has the potential to overstate feature representation ability due to the numerous hyperparameters and ways to perform linear probing experiments. Additionally, while fine-tuning compares pretrained weights as an initialization, this tends to not be the purest indicator for representation ability and has been shown to underperform for out-of-distribution downstream tasks [30].

References

[1] Steve Ahlswede, Christian Schulz, Christiano Gava, Patrick Helber, Benjamin Bischke, Michael Förster, Florencia Arias, Jörn Hees, Begüm Demir, and Birgit Kleinschmit. Treesatai benchmark archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data*, 15(2):681–695, 2023. 5, 6

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 3, 5, 7

[3] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2015. 3, 5, 7

[4] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34:22614–22627, 2021. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[6] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 1

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Im-

proved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[8] 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. 6, 7

[9] François Chollet et al. Keras. https://keras.io, 2015. 1

[10] 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. 1, 3, 5, 6, 7

[11] Isaac Corley and Peyman Najafirad. Supervising remote sensing change detection models with 3d surface semantics. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3753–3757. IEEE, 2022. 1

[12] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13 (1):21–27, 1967. 5

[13] 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. 1, 4

[14] 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. *arXiv preprint arXiv:2010.11929*, 2020. 7

[15] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. 2, 5

[16] Xianzhi Du, Yeqing Li, Yin Cui, Rui Qian, Jing Li, and Irwan Bello. Revisiting 3d resnets for video recognition. *arXiv preprint arXiv:2109.01696*, 2021. 3

[17] USDA Farm Service Agency (FSA). National Agriculture Imagery Program (NAIP). USDA Geospatial Data Gateway, 2015. 5

[18] Anthony Fuller, Koreen Millard, and James R Green. Satvit: Pretraining transformers for earth observation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1

[19] Anthony Fuller, Koreen Millard, and James R Green. Transfer learning with pretrained remote sensing transformers. *arXiv preprint arXiv:2209.14969*, 2022. 1

[20] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 7

[21] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pages 3809–3820. PMLR, 2021. 3

[22] 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, 3, 4

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

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 4

[25] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu. Self-supervised audiovisual representation learning for remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 116:103130, 2023. 1

[26] 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. 2, 3, 5, 6

[27] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 1

[28] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1

[29] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017. 6

[30] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 8

[31] Alexandre Lacoste, Evan David Sherwin, Hannah Kerner, Hamed Alemohammad, Björn Lütjens, Jeremy Irvin, David Dao, Alex Chang, Mehmet Gunturkun, Alexandre Drouin, et al. Toward foundation models for earth monitoring: Proposal for a climate change benchmark. *arXiv preprint arXiv:2112.00570*, 2021. 1

[32] Qun Liu, Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat v2: feature augmented convolutional neural nets for satellite image classification. *Remote Sensing Letters*, 11(2):156–165, 2020. 3, 7

[33] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023. 1

[34] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 1

[35] 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, 3, 4, 5, 6, 7

[36] Georgii Mikriukov, Mahdyar Ravanbakhsh, and Begüm Demir. Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing. *arXiv preprint arXiv:2201.08125*, 2022. 1

[37] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 681–699. Springer, 2020. 3

[38] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 5, 6

[39] Maxim Neumann, André Susano Pinto, Xiaohua Zhai, and Neil Houlsby. Training general representations for remote sensing using in-domain knowledge. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 6730–6733. IEEE, 2020. 1

[40] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023. 1

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 3

[42] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*, 2022. 1, 3, 5, 7

[43] Mats L Richter, Wolf Byttner, Ulf Krumnack, Anna Wiedenroth, Ludwig Schallner, and Justin Shenk. (input) size matters for cnn classifiers. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30*, pages 133–144. Springer, 2021. 3

[44] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):4392, 2021. 2, 4

[45] David P Roy, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Dennis Helder, James R Irons, David M Johnson, Robert Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145: 154–172, 2014. 2

[46] Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. Torchgeo: deep learning with geospatial data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–12, 2022. 3

[47] 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. 5

[48] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3): 174–180, 2021. 5

[49] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1

[50] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012. 5

[51] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019. 3

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[53] Gabriel Tseng, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023. 1, 5, 6

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6, 8

[55] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1

[56] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. *arXiv preprint arXiv:2211.07044*, 2022. 1, 2, 3, 4, 5

[57] Ross Wightman, Nathan Raw, Alexander Soare, Aman Arora, Chris Ha, Christoph Reich, Fredo Guan, Jakub Kaczmarzyk, mrT23, Mike, SeeFun, contrastive, Mohammed Rizin, Hyeongchan Kim, Csaba Kertész, Dushyant Mehta, Guillem Cucurull, Kushajveer Singh, hankyul, Yuki Tatsunami, Andrew Lavin, Juntang Zhuang, Matthijs Hollemans, Mohamed Rashad, Sepehr Sameni, Vyacheslav Shults, Lucain, Xiao Wang, Yonghye Kwon, and Yusuke Uchida.

rwightman/pytorch-image-models: v0.8.10dev0 Release, 2023. 1

[58] 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. 6, 7

[59] 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. 2, 4, 5