Learning Representations of Satellite Images From Metadata Supervision

Jules Bourcier[1,2], Gohar Dashyan[1], Karteek Alahari[2], and Jocelyn Chanussot[2]

[1] Preligens, Paris, France
[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France

Abstract. Self-supervised learning is increasingly applied to Earth observation problems that leverage satellite and other remotely sensed data. Within satellite imagery, metadata such as time and location often hold significant semantic information that improves scene understanding. In this paper, we introduce Satellite Metadata-Image Pretraining (SatMIP), a new approach for harnessing metadata in the pretraining phase through a flexible and unified multimodal learning objective. SatMIP represents metadata as textual captions and aligns images with metadata in a shared embedding space by solving a metadata-image contrastive task. Our model learns a non-trivial image representation that can effectively handle recognition tasks. We further enhance this model by combining image self-supervision and metadata supervision, introducing SatMIPS. As a result, SatMIPS improves over its image-image pretraining baseline, SimCLR, and accelerates convergence. Comparison against four recent contrastive and masked autoencoding-based methods for remote sensing also highlight the efficacy of our approach. Furthermore, our framework enables multimodal classification with metadata to improve the performance of visual features, and yields more robust hierarchical pretraining. Code and pretrained models will be made available at: `https://github.com/preligens-lab/satmip`.

Keywords: Self-supervised and multimodal learning · Remote sensing

1 Introduction

In recent years, self-supervised learning (SSL) has become a staple pretraining paradigm in computer vision, and has received much attention in the domain of remote sensing and Earth observation (EO) [58]. This marked interest can be attributed to two broad reasons. First, for a wide variety of high-impact EO tasks, ranging from crop-yield prediction to urban planning [11,14,19,31,47,50,54], labels are scarce and difficult to obtain, while unlabeled satellite imagery is abundantly available. This makes SSL eminently practical. Second, the diversity of remote sensors yields unique challenges of specialization, context awareness, and multimodal fusion, with rich spatial, temporal, and spectral contexts. This calls for the development of tailored representation learning methods, in order to address limitations of existing generic vision models [46,53].

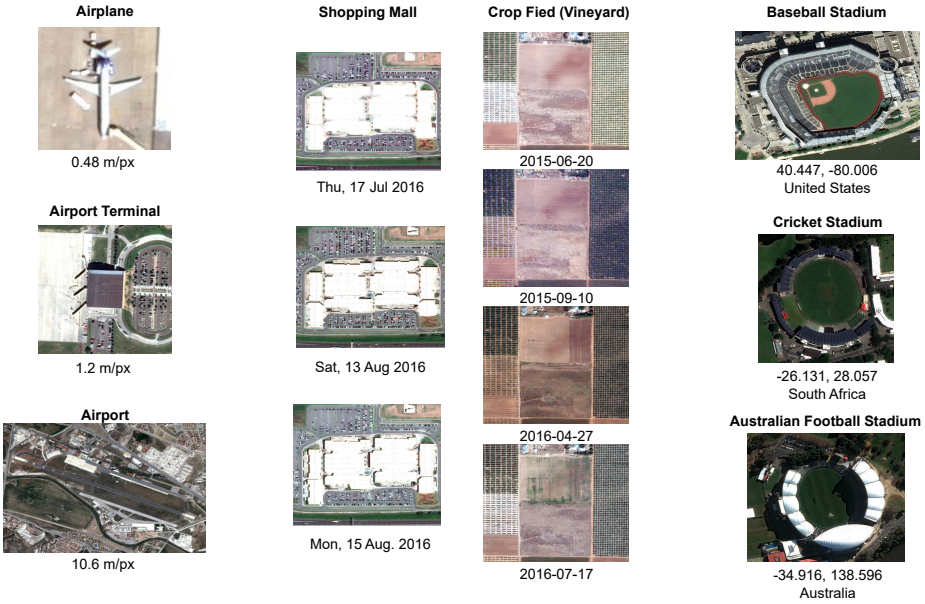


Fig. 1: Examples of satellite images where metadata can help recognizing objects. Ground sample distance (GSD) (a), which determines the area occupied by pixels, provides information on size and scale (e.g., airport/airport terminal/airplane); time (b) and location (c) can help understand the functional nature of man-made structures having different appearance depending on time (e.g., shopping mall's parking lot are fuller on the weekend, crop fields undergo cycles) or place (*e.g.*, some sports facilities correlated with regions on Earth), respectively.

EO satellites usually generate a rich set of metadata associated with the images they capture. They provide information about the size, scale, time of acquisition, as well as numerous other image properties. Some of these metadata, such as the ground sampling distance (GSD), timestamp, and geographic location, can be highly explanatory of semantic content present in images, as illustrated in Fig. 1.

Recent works have proposed SSL objectives using satellite metadata, such as predicting geolocation from images as a pretext task [2], using spatial neighbors or spatially-aligned images over time as positive views in contrastive learning [2, 26, 36, 42], or extending Vision Transformers (ViTs) [16] with positional encodings that integrate information from timestamps and spectral groups [12] or the GSD [44]. These methods have successfully improved contrastive/siamese learning algorithms [6, 7, 22] or masked autoencoders (MAEs) [4, 21]. However, these methods utilize different metadata fields via specialized model architectures or tasks. To our knowledge, there has not been a unified and flexible approach for incorporating heterogeneous metadata into a pretraining algorithm.

In this work, we propose a simple and effective model for learning visual representations from satellite metadata supervision. Our model, Satellite Metadata-Image Pretraining (SatMIP), encodes pairs of images and metadata as separate modalities and aligns them in a deep embedding space via a contrastive task, inspired by language-image pretraining [27, 43]. Through this task, we aim to learn a visual encoder that embeds metadata information, and their latent semantic characteristics, into image features. It requires metadata only during the pretraining phase, and not necessarily during transfer to downstream tasks. We pretrain ViTs backbones with SatMIP on the Functional Map of the World (fMoW) dataset [11], using GSD, timestamp, and geolocation, among other metadata fields. Through extensive experiments on various downstream classification datasets, we observe that the visual encoders pretrained with SatMIP generate non-trivial representations that generalize to downstream recognition tasks, showing that learning a joint embedding between images and metadata makes a meaningful pretext task.

To go one step further, we combine SatMIP with the image SSL method SimCLR [7], introducing SatMIPS. By co-solving an image-image and a metadata-image contrastive task with an efficient "coupled" architecture, SatMIPS benefits from both sources of supervision, and improves over it's SimCLR baseline, yielding better representations while converging faster. Moreover, on several downstream tasks, it outperforms multiple existing MAE and contrastive-based pretraining methods involving metadata for remote sensing. Furthermore, on downstream tasks with metadata, SatMIP allows deploying metadata features in tandem with visual features, which can further improves the classification performance. In addition, we also show that metadata supervision yields stronger results with hierarchical pretraining [45].

- We propose SatMIP, a novel self-supervised pretraining task and model for remote sensing inspired by CLIP [43], which aligns images with their metadata in a joint embedding space.
- We further propose SatMIPS, an evolution of the SLIP [38] architecture, which combines image-image and metadata-image contrastive learning.
- We conduct extensive experiments involving various downstream classification datasets, demonstrating the effectiveness and efficiency of our approach.

2 Related Work

Using geospatial metadata for visual representation learning. Satellite images systematically convey metadata that can be leveraged for free within SSL tasks similarly to pseudo-labels. One strategy involves employing metadata estimation as a pretext task: [61] proposes self-supervised time and location estimation tasks for learning geotemporal image features, while [2] solves a location classification as a subsidiary task to contrastive SSL [7, 22]. Contrary to predicting metadata information directly, our approach aligns image and metadata into a common embedding space.

Another avenue is the implicit incorporation of metadata information into existing pretext tasks, by using it to enrich the set of positive or negative instances and thereby learn the invariances driven by these augmentations. Building on this idea, some works use neighboring images in space as positives [26, 28, 42], while others use spatially-aligned images over time as positives [2, 35, 36]. Another line of recent works have extended masked autoencoders (MAE) [21] for remote sensing, and incorporated metadata information into positional encodings in ViTs. [12] proposes a spectral and temporal reconstruction task, and embeds timestamps and spectral bands into the positional encodings; [44] solves a super-resolution task and embeds GSD into positional encodings to incorporate scale information; [25] further extends this idea to multiple spectral bands, GSDs and sensors. In contrast to our model that is agnostic to network architectures or vision SSL frameworks, these approaches bake into ViTs and are specifically tailored to MAEs.

Embedding metadata. Recent studies aim to directly encode metadata along with images and perform a form of metadata-image pretraining. Close to our work, in [29, 34, 56], location encoders are learned through a contrastive image-location pretraining task, aiming to be deployed on downstream tasks involving location. In contrast to our approach, these works employ a two-step approach of training an image encoder, then a location encoder on frozen image features. Our objective being to train a visual encoder intended for downstream tasks without necessarily relying on metadata, we adopt the opposite approach by training a visual encoder with metadata supervision. Furthermore, instead of relying on location-specific encoder architectures as them, we employ a generic Transformer [55] that can be fed heterogeneous types of metadata beyond location.

Closer to our approach, [64] uses EXIF metadata and images in contrastive pretraining for learning to extract low-level camera properties of images for forensics tasks. Our approach differs in that we propose metadata-image pretraining for high-level representation learning. Furthermore, we enhance our metadata-image objective by concurrently solving an image-image contrastive task. Other work such as [20, 24] used medical images and biodata records as tabular features in contrastive learning for increasing performances on downstream visual diagnostic tasks. Such metadata is the result of a supervised and very expensive collection process that does not scale to large datasets.

Language *vs*. metadata supervision. Language-image pretraining (LIP) has emerged as a significant advance that bridges the gap between natural language and image representation learning. The works of CLIP [43] and ALIGN [27] showed that the straightforward pretext task of predicting which caption corresponds to a given image is an effective way to learn image representations on large-scale noisy (image, text) pairs. SLIP [38] extends CLIP via a multi-task learning framework combining it with image-image contrastive pretraining [7], showing that both objectives are synergistic. LIP has also recently garnered at-

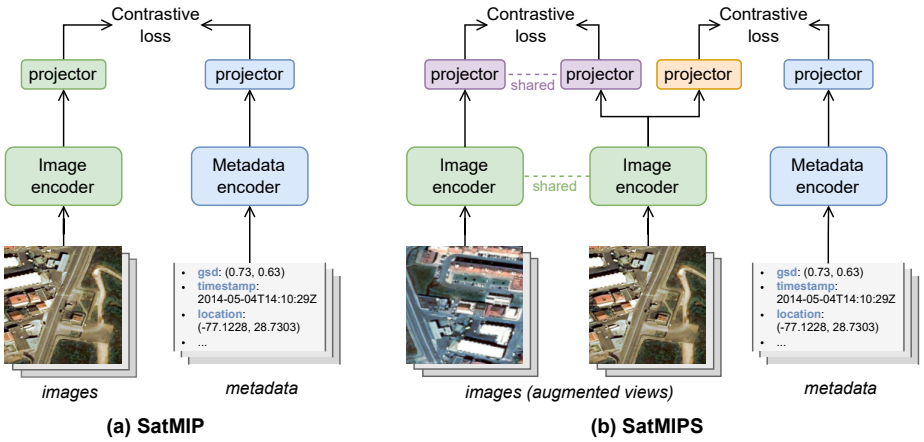


Fig. 2: (a) **Satellite Metadata-Image Pretraining (SatMIP)** learns a joint embedding between images and associated metadata. Batches of inputs are processed by a dual encoder and projection head, optimized through a contrastive loss, as in CLIP [43]. After pretraining, only the image encoder can be transferred to downstream tasks, or both can be used to perform bimodal (image, metadata) recognition. (b) **SatMIP with Self-Supervision (SatMIPS)** combines metadata with SimCLR-style [7] self-supervision: it learns both a joint embedding between augmented views of an image as well as an image and its metadata. The image encoder is shared between branches unlike the projection heads, and one image view is coupled between the two objectives for efficiency.

tention within the field of remote sensing [32,66]. However, while image captions are available in huge web-crawled multimedia datasets such as YFCC100M [51] or LAION-5B [48], such human-provided captions are scarce for remotely sensed images. On the other hand, metadata are automatically produced by sensors and are therefore widely available, which means that the number of image-metadata pairs can follow the ever-growing number of satellite images. Seeking to harness metadata for complementing vision, our SatMIP (respectively SatMIPS) model is analogous to CLIP (respectively SLIP), but with metadata as an input modality instead of language.

3 Method

We aim to learn a visual representation of remotely sensed images that embeds the semantic information contained within metadata that is obtained directly from the imaging sensor. To this end, we introduce SatMIP, a pretraining strategy that learns a joint embedding between an image and its metadata. Then, we introduce SatMIPS, which leverages both image self-supervision and metadata supervision. The architectures of SatMIP and SatMIPS are presented in Fig. 2.

3.1 SatMIP: Contrastive pretraining of metadata and image embeddings

Contrastive pretext task. We assume that we have access to an unlabeled dataset as $\mathbb{X} = \{(\mathbf{v}_i, \mathbf{m}_i)\}_{i=1}^{N}$, where $(\mathbf{v}_i, \mathbf{m}_i) \sim p(\mathcal{I}, \mathcal{M})$ are associated images and metadata pairs, sampled from their respective spaces $\mathcal{I}$, $\mathcal{M}$. The metadata space $\mathcal{M}$ can be composed of a set of numerical variables (*e.g.*, GSD, location coordinates, or look angle) and categorical variables (*e.g.*, sensor name). We define two neural network encoders, one for images $e^{\mathcal{I}} : \mathcal{I} \rightarrow \mathbb{R}^d$ and one for metadata $e^{\mathcal{M}} : \mathcal{M} \rightarrow \mathbb{R}^d$. Each encoder $e^{\mathcal{I}/\mathcal{M}}$ is composed of a backbone $f^{\mathcal{I}/\mathcal{M}}$ and a projection head $g^{\mathcal{I}/\mathcal{M}}$. Given a sampled batch of K image and metadata pairs, we compute embeddings of images $\mathbf{z}_i^{\mathcal{I}} = e^{\mathcal{I}}(\mathbf{v}_i)$ and metadata $\mathbf{z}_i^{\mathcal{M}} = e^{\mathcal{M}}(\mathbf{m}_i)$. Following CLIP [43], we use a contrastive loss by considering matching images and metadata as positives and non-matching images and metadata across the batch as negatives. Let us define the generic contrastive loss function [41]:

$$\mathcal{L}^{\text{clr}}(a_i, b_i) = -\log \frac{\exp\left(s(a_i, b_i)/\tau\right)}{\sum_{j=1}^{K} \exp\left(s(a_i, b_j)/\tau\right)}, \tag{1}$$

where (a_i, b_i) are two vectors of equal dimension, s is the cosine similarity and τ is a parameter than adjusts the dynamic range. We define the loss of SatMIP as the symmetrized contrastive loss between images and metadata embeddings:

$$\mathcal{L}_i^{\text{MI}}(\mathbf{z}_i^{\mathcal{I}}, \mathbf{z}_i^{\mathcal{M}}) = \frac{1}{2}\left(\mathcal{L}^{\text{clr}}(\mathbf{z}_i^{\mathcal{I}}, \mathbf{z}_i^{\mathcal{M}}) + \mathcal{L}^{\text{clr}}(\mathbf{z}_i^{\mathcal{M}}, \mathbf{z}_i^{\mathcal{I}})\right). \tag{2}$$

Through this objective, the weights of encoders $e^{\mathcal{I}}$ and $e^{\mathcal{M}}$ are optimized simultaneously to embed feature vectors of the matching image and metadata nearby in a common latent space. The objective can be naturally interpreted as one of classifying the correct metadata from the image, and correct image from the metadata. Intuitively, the image encoder will learn implicit neural features from the metadata, and vice versa. Note that there does not exist a simple 1:1 mapping between images and metadata, because metadata can match many image variations and vice versa (*e.g.*, due to the non-deterministic nature of weather). This prevents the model from solely overfitting the pretext task. In addition, we apply data augmentation to the images which futher regularizes the task.

Transfer to downstream tasks. Our primary goal with SatMIP is to learn a visual representation from the image backbone $f^{\mathcal{I}}$, to be transferred to downstream tasks. In this case, the metadata encoder is used only as a proxy for pretraining and is then discarded. Alternatively, provided that a downstream task provides a subset of metadata fields used to train the metadata encoder, we can encode metadata alongside images with the dual backbone ($f^{\mathcal{I}}$, $f^{\mathcal{M}}$) and fuse their embeddings as input to a supervised model. Several techniques exist to fuse multimodal embeddings [1, 3]. We use the simplest strategy of concatenating both vectors and fitting a parametric classifier on top to learn an optimal combination of features for a given task.

3.2 SatMIPS: Combining self- with metadata supervision

We further introduce SatMIPS which combines the previously described SatMIP with the SSL method SimCLR [7].

SimCLR learns a joint embedding between two augmented views of the same image with a contrastive loss, which makes it very similar to CLIP. Let the image encoder be $e = g \circ f : \mathcal{I} \to \mathbb{R}^c$. Given a batch of positive views $\{(\mathbf{v}_i, \mathbf{v}'_i)\}_{k=1}^K$, it computes embeddings $\mathbf{z}_i = e(\mathbf{v}_i)$ and $\mathbf{z}'_i = e(\mathbf{v}'_i)$ and employs the following symmetrized contrastive loss to align the embeddings of matching views:

$$\mathcal{L}_i^{\text{Sim}}(\mathbf{z}_i, \mathbf{z}'_i) = \frac{1}{2}\left(\mathcal{L}^{\text{clr}}(\mathbf{z}_i, \mathbf{z}'_i) + \mathcal{L}^{\text{clr}}(\mathbf{z}'_i, \mathbf{z}_i)\right) \tag{3}$$

Multi-task framework. [38] proposed SLIP as a combination of CLIP and SimCLR through multi-tasking. Following their framework, we express the SatMIPS task as a linear combination of the two SimCLR and SatMIP objectives. We share the visual backbone $f = f^{\mathcal{I}}$ between both models and optimize the sum of their loss:

$$\mathcal{L}_i^{\text{MI+Sim}}(\mathbf{z}_i^{\mathcal{I}}, \mathbf{z}_i^{\mathcal{M}}, \mathbf{z}_i, \mathbf{z}'_i) = \mathcal{L}_i^{\text{MI}} + \lambda \mathcal{L}_i^{\text{Sim}} \tag{4}$$

where λ is a hyperparameter that weights the promicence of the SimCLR loss relative to the SatMIP loss. We find that $\lambda = 1$, *i.e.*, equal weighting works well (we show this in additional ablations in the supplementary material).

Efficient view coupling. The main issue of SLIP is that it increases the number of images processed from 1 to 3, resulting in approximately 3× more activations [38] that increase training time and memory footprint significantly. To alleviate this cost, in SatMIPS, we couple one of the image view $\mathbf{v}_i$ between SimCLR and SatMIP. The output of the backbone f on this shared view is directed through the specific projection heads of SimCLR (g) and SatMIP ($g^{\mathcal{I}}$). This design is driven by our tests which showed that SatMIP works well with the same strong augmentation policy as SimCLR (*cf*. Sec. 4.7). Thanks to view coupling, we can largely reduce the overhead without an impact on downstream performance (*cf*. Sec. 4.7).

We have selected SimCLR as the image SSL method in SatMIPS for its simplicity and conceptual similarity with SatMIP. However, the general design of SatMIPS is agnostic to this choice, and the metadata-image objective could be blended with another SSL method.

4 Experiments and results

We evaluate our SatMIP and SatMIPS models by studying the performance of their learned representations on a set of remote sensing downstream classification tasks. We conduct experiments to benchmark the quality of representations

under k-nearest neighbors (kNN) and linear probing classification, the rate of convergence of pretraining, and the application of hierarchical pretraining. We then perform an ablation study of important components of our models.

4.1 Datasets

Pretraining. To pretrain our models we use the training set of the Functional Map of the World (fMoW) dataset [11], similar to previous work [2,12,44]. It consists of 363k global, very high-resolution images and associated metadata obtained by MAXAR optical satellites. We use the fMoW-RGB product, composed of the RGB pansharpened images. The metadata is composed of a diverse set of metadata fields, including, among others, GSD, timestamp, location, location-derived information such as UTM zone and country, cloud cover, and various imaging angles; we describe the full metadata considered in the supplementary material. We exclude any field that is obtained through manual annotation such as areas of interests and land use categories. Unless otherwise specified, in our SatMIP experiments, we used a combination of three fields: the GSD, timestamp, and location, described in Tab. 1 and visualized in the supplementary material. As we can see, the fields span an extensive range of values. We preprocess each source image by cropping over the annotated area of interest and resizing to 224×224 pixels, and we transform the GSD and location fields to match the preprocessed images.

Evaluation. To evaluate the performance of pretraining methods, we use a diverse set of 7 remote sensing RGB image classification datasets: (1) The labeled version of fMoW with 62 classes of functional land use (sharing the same training data as for pretraining); (2) RESISC45 [10] for land use/land cover classification of multi-sensor (satellite & aerial) images; (3) Optimal31 [57] for land use/land cover classification of multi-sensor images; (4) UC Merced [60] for land use classification of very-high resolution aerial images; (5) FGSC23 [63] for fine-grained ship classification in high-resolution multi-sensor images; (6) EuroSAT [23] for land use/land cover classification of Sentinel-2 images; (7) So2Sat [67] for local climate zone classification of Sentinel-2 images. We report macro-averaged F1 score on fMoW and FGSC23, and top-1 accuracy on the other datasets.

4.2 Setup

Baseline and state-of-the-art. We adopt SimCLR [7] as the natural baseline to compare SatMIP and SatMIPS to. We also compare to existing SSL models for remote sensing that are pretrained on fMoW-RGB: Geo, TP, and Geo-TP from [2], which are originally based on MoCo [8,22]; SatMAE [12], Scale-MAE [44], and SatMAE++ [40], which are based on MAE [21]. We reproduce contrastive methods on top of SimCLR for an even comparison with our models, while for MAE-based models, we take pretrained weights available on their official repositories,.

Table 1: The three fields of metadata we use to train our models, and descriptive statistics of values in the fMoW training set.

Field	Description	Min	Median	Max
Ground sample distance	Physical distance between pixel centers, in x and y directions (m)	0.08, 0.06	0.76, 0.60	23.13, 22.35
Timestamp	Date and time of acquisition (UTC)	2002-01-28 07:04:18	2015-08-06 10:08:02	2017-07-12 08:25:25
Location	Latitude and longitude of the image centroid (degrees)	-54.9320, -179.8810	37.9951, 7.0395	71.6118, 179.0439

Implementation details. *Visual encoders:* Unless otherwise noted, we use the MoCoV3 [9] version of a ViT-Small with patch size 16, consisting of 21.7M parameters.

Metadata encoders: To encode metadata, we experimented with two different types of encoders: (a) a textual encoder, which first converts metadata to text and then tokenizes it into as a sequence, processed like the language modality in CLIP [43] (inspired by [64] which applied this approach to EXIF metadata); (b) a tabular encoder, which natively supports numerical fields. For experiments targeting image-only classification, we use a textual encoder, while we use a tabular encoder for experiments involving bimodal (visual and metadata) classification. Both approaches work within our models, but perform differently depending on the downstream use case: we present a comparative study in the supplementary material. We use Transformer models [55] composed of 3 layers with 8 attention heads and a width of 512. For the textual approach, we use a BERT-style Transformer [15], while for the tabular approach, with use the FT-Transformer model [17].

Projection heads: In SatMIP, we use a linear layer for the visual and textual encoders. In SimCLR and SatMIPS, the projection head for the image-image objective is a 3-layer MLP.

Data augmentation: For all models, the inputs image views are generated with the augmentation policy of [7], with the addition of vertical flips and rotations [5,65]. We do not apply augmentation to the metadata.

Pretraining: We use a global batch size of 1024 for all models and train with the AdamW [33] optimizer for 200 epochs, unless otherwise noted. All other hyperparameters of SimCLR and SatMIP (S) are provided in the supplementary material.

Evaluation schemes. The keyword we apply for evaluating the models is *practicality*, of both the learned representation and the pretraining algorithm. To measure the achievement of these criteria, we evaluate the representation quality by fitting kNN or linear classifiers on frozen features extracted from the training

Table 2: kNN classification performance on various downstream datasets of SSL methods pretrained on fMoW. We compare SatMIP and SatMIPS against the baseline SimCLR and existing contrastive and MAE-based methods, under a consistent evaluation. For reference, Random indicates a feature extractor with random weights. We **bold** the best accuracies per dataset. R45: RESISC45, O31: Optimal31, UCM: UC Merced, F23: FGSC-23, Euro: EuroSAT, So2: So2Sat, Acc.: Accuracy (Top-1).

Model	ViT size	Epochs	fMoW F1	R45 Acc.	O31 Acc.	UCM Acc.	F23 Acc.	Euro Acc.	So2 Acc.
Random	S	–	5.1±0.1	32.2±0.4	28.5±0.6	44.0±0.9	26.7±1.4	70.0±0.1	33.3±0.2
Contrastive-based									
SimCLR [7]	S	200	61.1±0.6	88.5±0.4	86.0±1.2	95.0±0.4	57.3±2.1	94.3±0.6	**56.9**±0.5
SimCLR-Geo [2]	S	200	59.0±0.3	88.9±0.5	87.0±1.0	95.2±0.6	60.4±0.9	95.5±0.2	**57.5**±0.4
SimCLR-TP [2]	S	200	65.2±0.3	90.4±0.5	87.6±1.1	97.6±0.4	61.2±2.3	95.00±0.2	**57.3**±0.7
SimCLR-Geo-TP [2]	S	200	**65.8**±0.6	**91.3**±0.5	**89.7**±1.0	**97.1**±0.4	**63.0**±0.4	**95.6**±0.2	**57.2**±0.6
MAE-based									
SatMAE [12]	L	800	46.3	75.2	69.6	86.2	44.7	91.3	53.7
Scale-MAE [44]	L	800	51.4	85.9	81.6	89.0	48.2	**96.1**	**56.7**
SatMAE++ [40]	L	800	38.0	77.1	67.8	84.9	46.9	93.1	51.5
SatMIP	S	200	55.2±0.2	87.5±0.1	84.8±0.6	95.2±0.8	56.4±0.2	**95.7**±0.5	55.9±0.2
SatMIPS	S	200	62.3±0.0	89.7±0.2	87.9±0.2	94.9±0.7	60.8±0.6	95.1±0.1	**57.1**±0.5

set of downstream tasks, following regular protocols [6,59]. We measure resource efficiency by comparing performance across amounts of pretraining epochs and total training time.

4.3 Quality of visual representations

We present the results of our kNN classification experiments on fMoW and the four downstream datasets in Tab. 2. First, we observe that SatMIP learns non-trivial representations: it clearly outperforms random features, and even all MAE-based methods on all datasets except Scale-MAE on Euro and So2. It also competes with SimCLR on UCM and outperforms it on Euro. This validates that using metadata as supervision is effective for learning high-level semantic representations. Second, we see that SatMIPS outperforms SimCLR on all datasets, with the exception of UCM and So2 where they tie. This shows that image and metadata self-supervision interact constructively in SatMIPS to improve the quality of the shared features. Moreover, SatMIPS outperforms SimCLR-Geo on fMoW, R45, O31, and F23, and is comparable on other datasets. This tends to indicate that integrating metadata into a joint embedding objective is more effective than directly predicting metadata, as is done by Geo. Still, SatMIPS is mostly outperformed by SimCLR-TP and Geo-TP: although adding temporal positives makes a stronger extension than adding metadata supervision, we note that both methods could potentially be combined. However, this is beyond the scope of this work.

4.4 Classification on image and metadata features

Table 3: Classification with varying modalities on fMoW and EuroSAT: performance of image and combined image and metadata features learned via our models, using linear probing.

Model	fMoW F1			EuroSAT Acc.		
	Image	Image+Meta.	Δ	Image	Image+Meta.	Δ
SatMIP	$59.3_{\pm 0.3}$	$63.1_{\pm 0.1}$	$+3.8_{\pm 0.2}$	$94.5_{\pm 0.6}$	$95.5_{\pm 0.1}$	$+1.0_{\pm 0.4}$
SatMIPS	$65.8_{\pm 0.1}$	$68.6_{\pm 0.2}$	$+2.8_{\pm 0.2}$	$95.8_{\pm 0.3}$	$96.4_{\pm 0.2}$	$+0.6_{\pm 0.3}$

We investigate how the metadata modality can provide further benefits when used in downstream classification tasks. We perform bimodal classification using the combined representations of image and metadata encoders, on fMoW (using GSDs, timestamps and locations) and also on EuroSAT, for which we use the supplied GSDs and locations. We concatenate the features from each modality, and fit a linear classifier on the combined features. Results are presented in Tab. 3. On fMoW, we observe that for both SatMIP and SatMIPS, bimodal features provide a substantial performance improvement compared to features from images alone. On EuroSAT, bimodal features also achieve modest improvements. These results show that, in addition to forming useful supervision for pretraining, the learned metadata features are complementary to visual features for downstream tasks and can further improve performance.

4.5 Convergence speed analysis

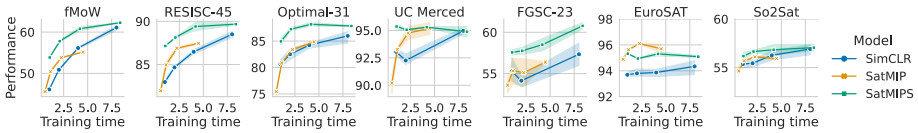


Fig. 3: kNN classification performance of various datasets as a function of total pretraining time for SatMIP, SimCLR and SatMIPS. Error bars indicate mean ± std. dev. over 3 runs. Data points correspond to 25, 50, 100 and 200 epochs in order of increasing time. Training times are relative to SimCLR at 25 epochs. Performance metric is accuracy except for fMoW and FGSC23 where it is F1 score.

To investigate the resource efficiency of pretraining procedures, we compare the performance obtained for different amounts of pretraining epochs and the resulting training time of SimCLR, SatMIP and SatMIPS. Results are presented

in Fig. 3. First, we note that SatMIP trains faster than SimCLR (by about 44%), and while it generally underperforms the baseline at 200 epochs, it generally does better when comparing at equivalent training time. Second, we observe that SatMIPS converges faster than SimCLR: the performance of the former at 100 epochs equates or surpasses the one of the latter at 200 epochs, on all the datasets. Even though SatMIPS is marginally slower to train than SimCLR for an equal amount of iterations/epochs (by about 5%), the higher convergence rate makes it a more efficient method. This shows that our metadata-image objective makes pretraining efficient, in addition to being effective.

4.6 Hierarchical pretraining

Table 4: kNN classification performance employing hierarchical pretraining (HPT) [45], with base pretraining on YFCC [51] with SLIP [38], and further pretraining each on fMoW with each model. Abbreviations share those of Tab. 2.

Model	HPT from SLIP-YFCC	fMoW F1	R45 Acc.	O31 Acc.	UCM Acc.	F23 Acc.	Euro Acc.	So2 Acc.	Avg. Δ
SimCLR		$61.1_{\pm 0.6}$	$88.5_{\pm 0.4}$	$86.0_{\pm 1.2}$	$95.0_{\pm 0.4}$	$57.3_{\pm 2.1}$	$94.3_{\pm 0.6}$	$56.9_{\pm 0.5}$	
	✓	$63.7_{\pm 0.1}$	$90.0_{\pm 0.2}$	$88.5_{\pm 0.8}$	$94.8_{\pm 0.0}$	$55.1_{\pm 1.3}$	$94.2_{\pm 0.3}$	$57.1_{\pm 0.4}$	+0.6
SatMIP		$55.2_{\pm 0.2}$	$87.5_{\pm 0.1}$	$84.8_{\pm 0.6}$	$95.2_{\pm 0.8}$	$56.4_{\pm 0.2}$	$95.7_{\pm 0.5}$	$55.9_{\pm 0.2}$	
	✓	$61.2_{\pm 0.4}$	$90.5_{\pm 0.2}$	$88.2_{\pm 0.1}$	$96.5_{\pm 0.2}$	$58.2_{\pm 0.4}$	$96.0_{\pm 0.4}$	$57.1_{\pm 0.7}$	**+2.4**
SatMIPS		$62.3_{\pm 0.04}$	$89.7_{\pm 0.2}$	$87.9_{\pm 0.2}$	$94.9_{\pm 0.7}$	$60.8_{\pm 0.6}$	$95.1_{\pm 0.1}$	$57.1_{\pm 0.5}$	
	✓	$66.3_{\pm 0.2}$	$91.4_{\pm 0.3}$	$89.7_{\pm 0.1}$	$96.2_{\pm 0.3}$	$59.9_{\pm 0.3}$	$95.9_{\pm 0.1}$	$57.9_{\pm 0.5}$	+1.4

We consider the compatibility of models with hierarchical pretraining (HPT) as it is known to be a practical way to increase the performance of SSL models, especially in remote sensing [37,45,62]. To do so, we initialize the ViT backbones with openly-available base weights pretrained with SLIP, on the large generalist YFCC15M [43,51] dataset. Results are presented in Tab. 4. We see that, on average, HPT provides greater performance improvements with SatMIP and SatMIPS. With SatMIP, the gains are also more steady across datasets, while HPT gives negative results on some datasets with SimCLR (UCM and F23) and SatMIPS (F23). Thus, SatMIPS advantage over SimCLR is reinforced, but more surprisingly, SatMIP outperforms SimCLR on all datasets except fMoW and O31, showing that the performance gap between metadata supervision and image self-supervision is largely closed when leveraging HPT.

4.7 Ablation study

We ablate key components of SatMIP and SatMIPS. For these experiments, we pretrain only for 25 epochs on fMoW. Additional ablations are provided in the supplementary material.

Table 5: Influence of the choice of metadata fields in SatMIP. We report downstream linear probing performance on fMoW and RESISC45, with "(I)" meaning classification on image features, while "(I,M)" means bimodal classification with combined (image, metadata) features. Time.: Timestamp, Loc.: Location. We highlight the defaults in blue and **bold** the best accuracies.

Metadata fields				Performance		
GSD	Time.	Loc.	12 other fields	$\text{fMoW}_{(I)}$ F1	$\text{fMoW}_{(I,M)}$ F1	$\text{R45}_{(I)}$ Acc.
✓				$37.4_{\pm 0.6}$	$40.5_{\pm 0.9}$	$71.3_{\pm 0.4}$
	✓			$43.7_{\pm 0.1}$	$46.5_{\pm 0.1}$	$74.3_{\pm 0.3}$
		✓		$47.1_{\pm 0.2}$	$53.3_{\pm 0.2}$	$76.1_{\pm 0.7}$
	✓	✓		$46.3_{\pm 0.1}$	$51.1_{\pm 0.1}$	$76.0_{\pm 0.2}$
✓		✓		$\mathbf{51.3_{\pm 0.7}}$	$\mathbf{58.3_{\pm 0.7}}$	$77.9_{\pm 0.3}$
✓	✓			$49.4_{\pm 0.2}$	$54.2_{\pm 0.3}$	$76.9_{\pm 0.4}$
✓	✓	✓		$\mathbf{50.7_{\pm 0.4}}$	$\mathbf{57.7_{\pm 0.5}}$	$\mathbf{78.5_{\pm 0.2}}$
✓	✓	✓	✓	$\mathbf{50.7_{\pm 0.1}}$	$\mathbf{58.0_{\pm 0.2}}$	$\mathbf{78.6_{\pm 0.1}}$

Choice of metadata in SatMIP. To study how each metadata field impacts the quality of representations learned by SatMIP, we ablate the set of fields used for pretraining and measure the performance of visual and bimodal classification on downstream tasks. We compare different combinations of the three GSD, timestamp and location fields, and using an "extended" set of fields composed of these three, and 12 other fields available in fMoW, detailed in the supplementary material. We present these results in Table 5. Regarding sets of single fields, we see that GSD performs the worst, followed by timestamp and location. This suggests that location contains the most useful semantic information for supervision. fMoW being a globally distributed dataset with a high variability of locations (*cf*. supplementary material), it likely contributes to the importance of location in SatMIP. Nevertheless, all single fields yield non-trivial performances, and are useful in bimodal classification on fMoW, demonstrating SatMIP's ability to incorporate metadata flexibly.. Overall, combining multiple fields together tends to improve performance over using single fields. On fMoW image and bimodal classification, SatMIP reaches best performance with GSD and location-only for both visual and bimodal classification, and enlarging to timestamp and other fields does not improve performance further. On R45 however, we observe best performance with GSD, timestamp and location. These results suggests that the fields of GSD, timestamp and location contain most of the useful information present in satellite metadata, and that SatMIP can constructively combine the complementary information present in these heterogeneous fields.

Coupling in SatMIPS. We justify the SatMIPS design choice of coupling image views between SimCLR and SatMIP. First, we ablate the image augmentation policy in SatMIP, comparing the strong policy used for the SimCLR

Table 6: Impact of image augmentation in SatMIP and view coupling in SatMIPS. Training time and GPU memory consumption are reported relative to SimCLR. We highlight the defaults in green/blue and **bold** the best numbers per model.

Model	Augmentation	Coupling?	Training time (rel.)	Memory /GPU (rel.)	fMoW F1	R45 Acc.
SimCLR	SimCLR-Sat	–	1.	1.	**46.1**$_{\pm 0.0}$	**83.0**$_{\pm 0.3}$
SimCLR	crop	–	1.	1.	33.4$_{\pm 0.2}$	77.5$_{\pm 0.2}$
SatMIP	SimCLR-Sat	–	0.56	0.62	**45.6**$_{\pm 0.4}$	**81.9**$_{\pm 0.2}$
SatMIP	crop	–	0.56	0.62	44.6$_{\pm 0.4}$	80.0$_{\pm 0.4}$
SatMIPS	SimCLR-Sat	no	1.53	1.58	**54.5**$_{\pm 0.1}$	**87.5**$_{\pm 0.1}$
SatMIPS	SimCLR-Sat	yes	**1.05**	**1.11**	53.9$_{\pm 0.1}$	87.1$_{\pm 0.0}$

baseline (SimCLR-Sat) with the light random resized cropping policy commonly used for LIP [38, 43]. We observe that a strong policy performs better, showing that it provides useful regularization while not altering the correctness of the metadata enough to perturb the contrastive task. Consequently, we adopt a common augmentation policy between SimCLR and SatMIP, and evaluate the impact on view coupling in SatMIPS on resource efficiency and performance. We observe that coupling reduces the training time and memory usage by about 30% relatively to decoupling, and has very little impact on the representation performance. Thanks to coupling, SatMIPS' training time is only 5% higher than that of SimCLR, and memory usage is also contained.

5 Conclusion

In this paper, we proposed a new self-supervised model for harnessing semantic information specific to satellite imagery metadata. We considered metadata as a complimentary modality to images, and demonstrated that SatMIP successfully learns useful visual and metadata representations. Our results have shown that metadata supervision is a strong competitor to traditional image-based SSL objectives, and that, within a multi-task framework, they are highly synergistic.

Our work on metadata supervision is focused on experiments with RGB images, but it could be applied to other remote sensing modalities, such as multispectral or radar images, since it does not make assumptions about the visual encoder. Also, we expect such representations to benefit from combining diverse sensors, thanks to the increased visual and metadata diversity. In addition, explicit information about the spectral bands could be included into the metadata encoding (*e.g.*, wavelengths or calibration parameters [30]) for learning spectrally-aware representations.

Additionally, our evaluation is focused on classification tasks on frozen features. One more avenue for future work is to study the behavior of models when finetuned, and exploring the transferrability our models to dense prediction tasks.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work has been supported in part by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003), and has also benefited from access to the HPC resources of IDRIS under the allocation 2024-AD011013097R2 made by GENCI.

References

1. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Proceedings, Part I. pp. 427–443. Springer (2020)
2. Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10181–10190 (2021)
3. Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M.: Visual and textual deep feature fusion for document image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 562–563 (2020)
4. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022)
5. Bourcier, J., Dashyan, G., Chanussot, J., Alahari, K.: Evaluating the label efficiency of contrastive self-supervised learning for multi-resolution satellite imagery. In: Image and Signal Processing for Remote Sensing XXVIII. vol. 12267, pp. 152–161. SPIE (2022)
6. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9630–9640 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
9. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9620–9629 (2021)
10. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
11. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180 (2018)
12. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems **35**, 197–211 (2022)
13. Corley, I., Robinson, C., Dodhia, R., Ferres, J.M.L., Najafirad, P.: Revisiting pretrained remote sensing model benchmarks: resizing and normalization matters. arXiv preprint arXiv:2305.13456 (2023)

14. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 172–181 (2018)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)
18. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
19. Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M.: Creating xBD: A dataset for assessing building damage from satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 10–17 (2019)
20. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023)
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
23. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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)
24. Huang, W.: Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2473–2482 (2023)
25. Irvin, J., Tao, L., Zhou, J., Ma, Y., Nashold, L., Liu, B., Ng, A.Y.: USat: A unified self-supervised encoder for multi-sensor satellite imagery. arXiv preprint arXiv:2312.02199 (2023)
26. Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., Ermon, S.: Tile2vec: Unsupervised representation learning for spatially distributed data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3967–3974 (2019)
27. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
28. Kang, J., Fernandez-Beltran, R., Duan, P., Liu, S., Plaza, A.J.: Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. IEEE Transactions on Geoscience and Remote Sensing **59**(3), 2598–2610 (2020)

29. Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M.: Towards global, general-purpose pretrained geographic location encoders. In: NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models (2023)
30. Kuester, M., Ochoa, T.: Absolute radiometric calibration is an essential tool to imagery science, but what is it? https://blog.maxar.com/tech-and-tradecraft/2020/absolute-radiometric-calibration-is-an-essential-tool-to-imagery-science-but-what-is-it (Feb 2020)
31. Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y., McCord, B.: xview: Objects in context in overhead imagery. arXiv preprint arXiv:1802.07856 (2018)
32. Li, X., Wen, C., Hu, Y., Yuan, Z., Zhu, X.X.: Vision-language models in remote sensing: Current progress and future trends. IEEE Geoscience and Remote Sensing Magazine (2024)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
34. Mai, G., Lao, N., He, Y., Song, J., Ermon, S.: Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In: International Conference on Machine Learning. PMLR (2023)
35. Mall, U., Hariharan, B., Bala, K.: Change-aware sampling and contrastive learning for satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5261–5270 (2023)
36. Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9414–9423 (2021)
37. Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C.: Towards geospatial foundation models via continual pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16806–16816 (2023)
38. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European Conference on Computer Vision. pp. 529–544. Springer (2022)
39. Neumann, M., Pinto, A.S., Zhai, X., Houlsby, N.: In-domain representation learning for remote sensing. In: International Conference on Learning Representations Workshops. pp. 1–20 (2020)
40. Noman, M., Naseer, M., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.S.: Rethinking transformers pre-training for multi-spectral satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27811–27819 (2024)
41. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
42. Pantazis, O., Brostow, G.J., Jones, K.E., Mac Aodha, O.: Focus on the positives: Self-supervised learning for biodiversity monitoring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10583–10592 (2021)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
44. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099 (2023)

45. Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillory, D., Metzger, S., et al.: Self-supervised pretraining improves self-supervised pretraining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2584–2594 (2022)
46. Rolf, E., Klemmer, K., Robinson, C., Kerner, H.: Mission critical – satellite data is a distinct modality in machine learning. arXiv preprint arXiv:2402.01444 (2024)
47. Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X.: Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. arXiv preprint arXiv:1906.07789 (2019)
48. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
49. Stewart, A.J., Robinson, C., Corley, I.A., Ortiz, A., Ferres, J.M.L., Banerjee, A.: Torchgeo: deep learning with geospatial data. In: Proceedings of the 30th international conference on advances in geographic information systems. pp. 1–12 (2022), https://github.com/microsoft/torchgeo
50. Sumbul, G., Charfuelan, M., Demir, B., Markl, V.: Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 5901–5904. IEEE (2019)
51. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
52. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
53. Tuia, D., Roscher, R., Wegner, J.D., Jacobs, N., Zhu, X.X., Camps-Valls, G.: Toward a collective agenda on ai for earth science data analysis. IEEE Geoscience and Remote Sensing Magazine **9**(2), 88–104 (2021)
54. Van Etten, A., Lindenbaum, D., Bacastow, T.M.: Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232 (2018)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
56. Vivanco Cepeda, V., Nayak, G.K., Shah, M.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems **36** (2024)
57. Wang, Q., Liu, S., Chanussot, J., Li, X.: Scene classification with recurrent attention of vhr remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **57**(2), 1155–1167 (2018)
58. Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X.: Self-supervised learning in remote sensing: A review. IEEE Geoscience and Remote Sensing Magazine **10**(4), 213–247 (2022)
59. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)

60. Yang, Y., Newsam, S.: 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. pp. 270–279 (2010)
61. Zhai, M., Salem, T., Greenwell, C., Workman, S., Pless, R., Jacobs, N.: Learning geo-temporal image features. In: British Machine Vision Conference (2019)
62. Zhang, T., Gao, P., Dong, H., Zhuang, Y., Wang, G., Zhang, W., Chen, H.: Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. Remote Sensing **14**(22), 5675 (2022)
63. Zhang, X., Lv, Y., Yao, L., Xiong, W., Fu, C.: A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 1271–1285 (2020)
64. Zheng, C., Shrivastava, A., Owens, A.: Exif as language: Learning cross-modal associations between images and camera metadata. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6945–6956 (2023)
65. Zheng, X., Kellenberger, B., Gong, R., Hajnsek, I., Tuia, D.: Self-supervised pretraining and controlled augmentation improve rare wildlife recognition in uav images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 732–741 (2021)
66. Zhou, Y., Feng, L., Ke, Y., Jiang, X., Yan, J., Yang, X., Zhang, W.: Towards vision-language geo-foundation model: A survey. arXiv preprint arXiv:2406.09385 (2024)
67. Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., 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)

Supplementary Material

A Datasets details

We report properties of the used datasets in Tab. S1, including number of train and test samples, number of classes, image resolution, GSD range, location extent and the sensors comprising each dataset.

Table S1: Details of classification datasets used in experiments.

Dataset	Num. train samples	Num. test samples	Num. classes	Resolution (px)	GSD (m)	Location extent	Sensor(s)
fMoW [11]	363,572	53,043	62	224×224	0.06–23	207 countries / 400 UTM zones	WorldView-2, WorldView-3, QuickBird-2, GeoEye-1
RESISC45 [10]	18,900	6,300	45	256×256	0.2–30+	global	various (Google Earth)
Optimal31 [57]	930	930	31	256×256	0.5–8	global	various (Google Earth)
UC Merced [60]	1260	840	21	256×256	0.3	Contiguous USA	NAIP
FGSC-23 [63]	3256	824	23	variable (40–800)	0.4–2	*Unknown*	various (Google Earth), GaoFen-1
EuroSAT [23]	16,200	5,400	10	64×64	10	34 European countries	Sentinel-2
So2Sat [67]	352,366	48,307	17	32×32	10	42 cities distributed globally	Sentinel-2

Spectral bands. We perform all experiments on three-bands RGB images. For EuroSAT and So2Sat which provide additional spectral bands, we retain only the RGB bands.

Data splits. We use one train and one test split for all datasets. For fMoW, we use the official train and validation splits as our train and test splits respectively, following [2,12,44]. For RESISC45, UC Merced, EuroSAT and So2Sat, we use the train and test splits available in TorchGeo [49], which are taken from [39]. For UC Merced, we use the combined test and val splits as our test set, to inflate its size. For So2Sat, we use the "Culture-10" version of the dataset. For Optimal31, we randomly split the full dataset (1,860 samples) between train and test with a 50/50 ratio.

fMoW preprocessing. Our preprocessing of fMoW aligns with previous works [2, 11, 12]. We use the fMoW-RGB dataset product composed pansharpened color images converted to 8-bit JPEGs files, and JSON metadata files. We preprocess the dataset using the standard method: each image is cropped around an area of interest (AOI) and resized to 224×224 pixels. Resized cropping affects the associated GSD and location. We transform the GSD height and width according to the size ratio of the cropped image to the resized cropped image. We replace the location polygon with the one encompassing the AOI. Other metadata fields are not affected.

Resizing and normalization. For the evaluation of fMoW-pretrained models, we follow [13] for resizing and normalizing of images. We resize to the resolution used for pretraining (224×224 pixels) or keep the original size if it is higher. Doing so tends to give optimal performance for all the compared models on all datasets, except for Scale-MAE, which we evaluate using a resolution of 128×128 pixels on all datasets as it gives better performance[3]. For pretraining and evaluation, we perform channel-wise standardisation with mean and standard deviation statistics computed on the training set of each dataset [13].

B Pretraining details

Visual encoders. We follow [38] for the configurations of the ViT backbones. We use the ViT-S variant from MoCoV3 [9] with 12 heads per attention layer (*vs*. 6 in original ViT-S [52]). We use a patch size of 16, and learnable positional embeddings. The output representation that is passed to the projection head for pretraining, and used for downstream tasks, is the `CLS` token of the last layer.

Textual metadata encoders. For our experiments with a textual representation of metadata, we apply the following processing. Following [64], we format different fields as key-value pairs of strings, and concatenate each key-value pair together to form a composite string using the syntax "$\texttt{key}_1\texttt{: value}_1\texttt{, ..., key}_n\texttt{: value}_n$". Afterwards, we tokenize the text using the CLIP's Byte Pair Encoding (BPE) tokenizer. We then feed the sequence of tokens to a BERT-style [15] Transformer encoder with 3 layers, width 512, 8 attention heads per layer, and a FFN size factor of 4. We use learnable positional embeddings. We use the "pre-norm" variant of Transformer following [43]. The output representation that is passed to the projection head for pretraining, and used for downstream tasks, is the `CLS` token of the last layer.

Tabular metadata encoders. For our experiments with metadata as tabular features, we decompose the metadata into atomic numerical or categorical fields; the only field for which this is not straightforward is timestamp, which

[3] this is consistent with the results of [44]

we convert into year, month, day, hour, and weekday. The numerical features are further standardized by removing the mean and scaling to unit variance. We concatenate both numerical and categorical vectors and pass as input to a FT-Transformer [17] composed of a feature tokenizer (see [17] for details) and a Transformer with 3 layers, a width of 192, 8 attention heads per layer and a FFN size factor of 4/3. We use the "pre-norm" variant of Transformer, and remove the first normalization from the first layer following [17]. The output representation that is passed to the projection head for pretraining, and used for downstream tasks, is the `CLS` token of the last layer.

Projection heads. We follow [38] for the configuration of projection heads. The projection head for the metadata-image loss in SatMIP(s) is a linear layer specific to each modality that map each representation to a 512-dim embedding. The projection head for the image-image loss in SimCLR and SatMIPS is a MLP composed of 3 4096-dim hidden layers, interposed with BatchNorm and ReLU, and outputs 256-dim embeddings.

Temporature scaling in contrastive loss. Following [38], the temperature τ is set to 0.1 for the image-image loss in SimCLR and SatMIPS, while it is set to a learnable parameter in the metadata-image loss in SatMIP(S).

Augmentation. We use the same augmentation policy across image inputs in SimCLR/SatMIP(S). Borrowing from [5], we opt for a modified version of the standard SimCLR policy for satellite images, composed of: random resized cropping with a scale ratio sampled uniformly in $[0.2, 1.0]$ and target size 224 px, color jittering with $p = 0.8$, grayscaling with $p = 0.2$, Gaussian blurring with $p = 0.5$, horizontal flipping with $p = 0.5$; vertical flipping with $p = 0.5$, and rotation with $p = 0.75$ by an angle sampled uniformly in $\{90, 180, 270\}$. In the ablation of Tab. 6 in the main paper, the "crop" policy is random resized cropping with a scale ratio sampled between $[0.5, 1.0]$ [38].

Training. Most of our training hyperparameters are reused from [38], and we translate their recipes of CLIP and SLIP to our SatMIP and SatMIPS, respectively. We perform stochastic gradient descent with the AdamW [33] optimizer (with $(\beta_1, \beta_2) = (0.9, 0.98)$ and $\epsilon = 1e-8$). We use a global batch size of 1024, and a cosine learning rate decay, with 1 epoch of linear warmup [7]. We apply the linear scaling rule [18] to set the initial learning rate: $\text{lr} = \text{lr}_{\text{base}} \cdot \text{bs}/256$, with bs the batch size and lr_{base} a base learning rate. We train the models with mixed precision. Base learning rate and weight decay have different values depending on the model, given in the following table:

Model	SimCLR	SatMIP	SatMIPS
base learning rate	2e-4	1.875e-4	3.75e-4
weight decay	0.1	0.5	0.5

In SatMIPS loss, we set the value of λ to 1. We show the impact of λ in Tab. S4.

Code and compute environment. Our implementation of SimCLR, SatMIP and SatMIPS is based on the official code of SLIP[4]. We use PyTorch 2.1. Trainings are performed on compute nodes with 4 Nvidia V100-32GB or 8 Nvidia A100-40GB. kNN evaluations are performed on one V100-32GB.

C kNN classification details

After pretraining, the representation we evaluate is the `CLS` token output of ViT backbones. We use a weighted kNN classifier following standard practice [6,44,59]. We freeze the pretrained model and extract the representations of the training and testing set examples. We classify each test sample by performing a weighted vote among the top k training samples sorted by decreasing cosine similarity. We do not use any data augmentation. We sweep the number of neighbors k in the set $\{1, 5, 20, 100\}$ for each model and dataset combination, and report optimal results. For all contrastive-based models, we select $k = 100$ on fMoW and $k = 5$ on the other datasets. For MAE-based models, we select $k = 20$ on fMoW and $k = 5$ on the other datasets. We enable mixed precision for feature extraction and calculating the pairwise similarities between samples.

D Linear probing classification details

For linear probing, we fit a logistic regression classifier on the training set embeddings, using L-BFGS optimizer with 200 maximum iterations, and no regularization. For bimodal classification, we first extract image and metadata features and concatenate both [CLS] token embeddings, standardize the feature to zero mean and unit variance, and fit a logistic regression classifier.

E Description of fMoW metadata

In Tab. S2, we detail the full set of 15 metadata fields from fMoW that we considered throughout our experiments. Recall that by default, we used the subset of GSD, timestamp, and location fields (row (1), (4), and (5) in Tab. S2, respectively). This full set of 15 fields was used in the ablation in Tab. 5 of the main paper.

Additionally, we visualize the distribution of the main metadata fields:

- *GSD:* In Fig. S1, we observe that the vast majority of GSD width and height values are concentrated between 0.3 m and 2 m. The distribution has a long tail of higher GSD values ranging up to 23 m.

[4] https://github.com/facebookresearch/SLIP

Table S2: Details of the full set of 15 metadata fields selected from the fMoW dataset for our experiments. Colors designate two types of metadata: (a) sensor: fields that are determined by the sensor's characteristics and/or it's relative position to the target); (b) environment: fields that are determined by the environment (*i.e.*, the geotemporal context). Refer to the fMoW paper [11] for a detailed documentation.

Field	Description	Example value
Ground sample distance	GSD of panchromatic band in the raw image strip, in meters. Transformed according to resized cropping (*cf*. Sec. A).	`[0.3749, 0.2916]`
Multispectral ground sample distance	GSD of multispectral bands in the raw image strip, in meters. We include the average of width and height. Transformed according to resized cropping.	`1.3365`
Pixel size	Size of a pixel in longitude and latitude units in the panchromatic band, in degrees. Transformed according to resized cropping.	`[3.27e-06, 2.54e-06]`
Timestamp	ISO UTC timestamp of acquisition down to the second.	`2016-07-02 T12:40:44Z`
Location	Longitude (-180–180) and latitude (-90–90) of the image centroid, in degrees. Transformed according to cropping (*cf*. Sec. A).	`[-43.246798, -22.982982]`
UTM zone	Provides a number for the UTM zone (1–60), along with a letter representing the latitude band ("C"–"X").	`23K`
Country code	ISO alpha-3 country code.	`BRA`
Cloud cover	Percentage of the raw image strip that is completely obscured by clouds (0–100).	`0`
Scan direction	Direction in which the sensor is pointed during take, relatively to the orbital path. Equals "Forward" if taken ahead of the orbital path and "Reverse" if taken behind.	`Reverse`
Wavelengths	Approximate central wavelength of the red, green and blue bands. [b]	`[661, 545, 477]`
Target azimuth	Azimuth angle of the sensor to the center of the image strip, from north, clockwise, in degrees (0–360).	`0.58`
Sun azimuth	Azimuth angle of the sun to the center of the image strip, from north, clockwise, in degrees (0–360).	`67.86`
Sun elevation	Elevation angle of the sun from the horizon, in degrees (0–90).	`61.34`
Off-nadir angle	The off-nadir angle of the sensor to the center of the image strip, in degrees (0–90).	`43.92`
Sensor platform	Name of the sensor, among: WorldView-2, WorldView-3, QuickBird-2, and GeoEye-1.	`GEOEYE01`

[a] Note that all sensors capture at the same wavelengths, so this field is constant throughout the dataset, making it inoperative.

– *Location:* In Fig. S2, we see that locations span a global distribution across all five continents. However, we note an overall bias towards the global North, while some regions, such as Subsaharan Africa and South Asia, are underrepresented.
– *Timestamp:* In Fig. S3, we see that dates are unequally spread in the full 12-years time range, with the majority being 2014 and 2017. Months and weekdays, however, are more uniformly distributed.

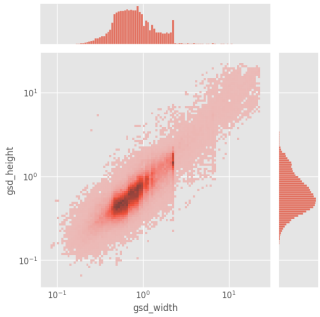


Fig. S1: Distribution of ground sampling distances (width and height) in the fMoW training set. Note the log scale.

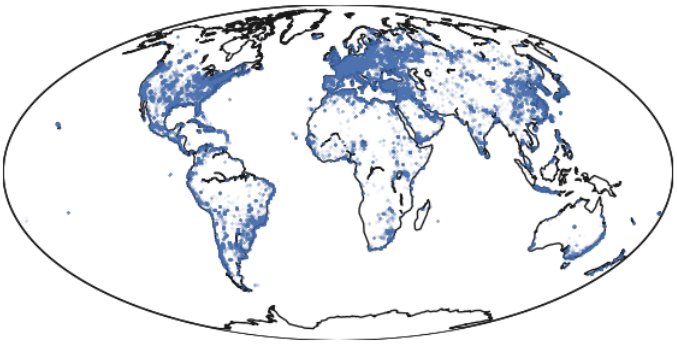
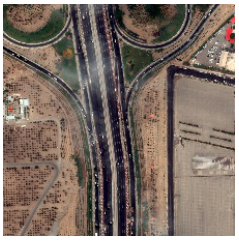


Fig. S2: Distribution of geographic locations in the fMoW training set.

F Examples of images and metadata

Tab. S3 presents sample images and metadata pairs from the fMoW dataset, that we used for metadata-image pretraining within SatMIP and SatMIPS. Metadata is here shown as text form.

Table S3: Sample images from the fMoW dataset with their metadata as a formatted text, using the full set of 15 fields described in Tab. S2. We also report the number of resulting text tokens (excluding start-of and end-of-text tokens), and the class the sample belongs to.

	Image	Metadata (text)	Class
(1)		ground_sample_distance: [11.4546, 9.1344], multispectral_ground_sample_distance: 41.1896, pixel_size: [1.03e-04, 8.23e-05], timestamp: 2015-09-21T15:30:08Z, location: [-73.310776, -3.785814], utm_zone: 18M, country_code: PER, cloud_cover: 14, scan_direction: Reverse, wavelengths: [661, 545, 477], target_azimuth: 39.12, sun_azimuth: 77.28, sun_elevation: 70.44, off_nadir_angle: 27.70, sensor_platform: GEOEYE01	Airport
(2)		ground_sample_distance: [2.0638, 1.8120], multispectral_ground_sample_distance: 7.7561, pixel_size: [1.86e-05, 1.64e-05], timestamp: 2016-07-02T07:33:42Z, location: [51.253020, 35.711928], utm_zone: 39S, country_code: IRN, cloud_cover: 3, scan_direction: Forward, wavelengths: [661, 545, 477], target_azimuth: 312.30, sun_azimuth: 126.03, sun_elevation: 71.28, off_nadir_angle: 22.31, sensor_platform: WORLDVIEW03_VNIR	Interchange
(3)		ground_sample_distance: [1.8129, 1.9807], multispectral_ground_sample_distance: 7.5664, pixel_size: [1.65e-05, 1.80e-05], timestamp: 2014-05-27T03:05:01Z, location: [120.572210, 14.984249], utm_zone: 51P, country_code: PHL, cloud_cover: 6, scan_direction: Reverse, wavelengths: [661, 545, 477], target_azimuth: 85.94, sun_azimuth: 58.19, sun_elevation: 76.37, off_nadir_angle: 25.91, sensor_platform: WORLDVIEW02	Crop Field
(4)		ground_sample_distance: [2.1283, 1.3499], multispectral_ground_sample_distance: 6.9531, pixel_size: [1.89e-05, 1.20e-05], timestamp: 2005-12-21T17:59:22Z, location: [-105.222951, 39.749676], utm_zone: 13S, country_code: USA, cloud_cover: 2, scan_direction: Forward, wavelengths: [661, 545, 477], target_azimuth: 289.08, sun_azimuth: 164.96, sun_elevation: 25.32, off_nadir_angle: 25.12, sensor_platform: QUICKBIRD02	Educational Institution
(5)		ground_sample_distance: [0.3191, 0.4155], multispectral_ground_sample_distance: 1.4697, pixel_size: [3.24e-06, 4.22e-06], timestamp: 2017-04-26T10:08:48Z, location: [6.462309, 13.403795], utm_zone: 32P, country_code: NGA, cloud_cover: 0, scan_direction: Reverse, wavelengths: [661, 545, 477], target_azimuth: 341.90, sun_azimuth: 84.84, sun_elevation: 69.78, off_nadir_angle: 17.99, sensor_platform: GEOEYE01	Single-Unit Residential

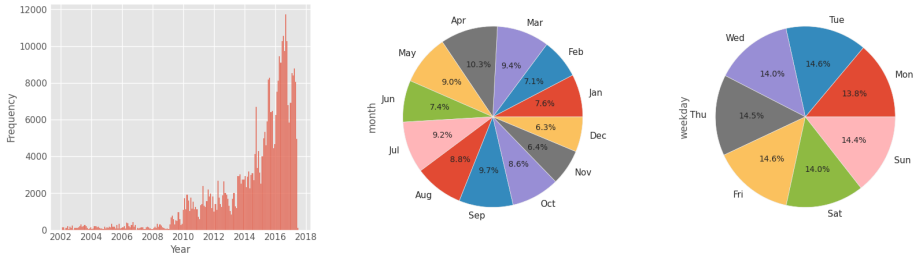


Fig. S3: Distribution of timestamps' years, months and weekdays in the fMoW training set.

G Additional ablations

We present additional ablations of our SatMIP and SatMIPS models.

G.1 Multi-task balancing in SatMIPS loss

In Tab. S4, we ablate the value of the hyperparameter λ, which balances the metadata-image and image-image objectives. We pretrain on fMoW for 25 epochs with a textual metadata encoder. We observe that performance is not significantly impacted by the choice of λ, provided that the value is greather than 0 (or it is equivalent to SatMIP). SatMIPS can benefit equally from both objectives regarless of their weighting.

Table S4: Impact of the multi-task loss balancing factor λ in SatMIPS. Note that $\lambda = 0$ is equivalent to SatMIP as the SimCLR objective is null.

λ	fMoW F1	R45 Acc.	F23 Acc.	So2 Acc.
0	$45.6_{\pm 0.4}$	$81.9_{\pm 0.2}$	$54.1_{\pm 0.4}$	$54.7_{\pm 0.5}$
0.5	$53.7_{\pm 0.3}$	$86.8_{\pm 0.03}$	$57.0_{\pm 0.4}$	$56.0_{\pm 0.1}$
1.0	$53.9_{\pm 0.1}$	$87.1_{\pm 0.02}$	$57.6_{\pm 0.5}$	$56.2_{\pm 0.6}$
2	$53.9_{\pm 0.4}$	$86.7_{\pm 0.1}$	$58.4_{\pm 1.1}$	$56.3_{\pm 0.2}$

G.2 Textual *vs*. tabular metadata encoders in SatMIP

We present an extensive comparison of the two approaches we adopt for encoding metadata within SatMIP: using a text encoder (BERT-style Transformer on textualized inputs), and a tabular encoder (FT-Transformer on featurized inputs). Our choice for using a textual encoder was motivated by [64], who demonstrated

the flexibility and effectiveness of textual encoding on EXIF tags. Nevertheless, we may hypothesize that a textual representation should be ill-suited for numerical fields such as location or GSD: as it treats them as sequences of digit tokens, it must limited understanding of those fields. Using vectorized features as input to a tabular encoder must be more suited for numerical fields by definition. First, in Tab. S5, we compare the kNN classification performance of SatMIP(S) trained with both type of encoders on the various datasets as well as their efficiency. We observe that for SatMIP, surprisingly, a textual encoder tends to perform better, with higher accuracies on 5 out of the 7 datasets. For SatMIPS, their performance is on par overall, except in favor of the textual encoder on one dataset (O31). These results indicate that a textual encoder tends to be more effective, although the tabular encoder is competitive. However, this observation may just be due to the choice of hyper-parameters, as we mostly reused the hyperparameters from SLIP [38] with minimal tuning, and SLIP uses a textual encoder on language captions. Therefore, we cannot draw any definitive conclusions. However, we note that the tabular encoder is more memory efficient, as it requires way less tokens to train (about 10× for GSD, timestamp and location as inputs).

Then, in Sec. G.2, we compare the linear probing performance of SatMIP(S) trained with both encoders using multiple modalities. We observe than when metadata features are used as input to classification, the tabular encoder peforms much better than the textual encoder, which is in contrast to using image features alone. This clearly shows that numerical fields understanding is important for deploying metadata encoders. The textual encoder might be good at solving the image-metadata matching task from token sequences, but it is limited in it's ability to generalize to new data on downstream tasks. This shows that a tabular encoder should be favored when considering multimodal classification with SatMIP(S).

Table S5: kNN classification performance with a tabular encoder *vs*. a textual encoder for metadatas. 200 epochs pretraining on fMoW. Training time and memory usage are relative to the baseline, SimCLR.

Model	Encoder	fMoW F1	R45 Acc.	O31 Acc.	UCM Acc.	FGSC-23 F1	Euro Acc.	So2 Acc.	Train. time	Mem. /GPU
SatMIP	Textual	$55.2_{\pm 0.2}$	$87.5_{\pm 0.1}$	$\mathbf{84.8}_{\pm \mathbf{0.6}}$	$\mathbf{95.2}_{\pm \mathbf{0.8}}$	$\mathbf{56.4}_{\pm \mathbf{0.2}}$	$\mathbf{95.7}_{\pm \mathbf{0.5}}$	$\mathbf{55.9}_{\pm \mathbf{0.2}}$	0.56	0.62
	Tabular	$\mathbf{55.8}_{\pm \mathbf{0.3}}$	$87.2_{\pm 0.3}$	$82.4_{\pm 0.9}$	$94.3_{\pm 0.3}$	$55.3_{\pm 1.3}$	$94.2_{\pm 0.4}$	$55.2_{\pm 0.5}$	0.56	**0.52**
SatMIPS	Textual	$62.4_{\pm 0.1}$	$89.7_{\pm 0.1}$	$\mathbf{88.1}_{\pm \mathbf{1.1}}$	$95.2_{\pm 0.6}$	$58.8_{\pm 0.5}$	$94.8_{\pm 0.1}$	$57.3_{\pm 0.1}$	1.05	1.11
	Tabular	$62.5_{\pm 0.4}$	$89.6_{\pm 0.2}$	$86.5_{\pm 0.9}$	$95.6_{\pm 0.4}$	$59.2_{\pm 1.1}$	$94.9_{\pm 0.2}$	$57.2_{\pm 0.4}$	1.05	**1.01**

Table S6: Linear probing classification on fMoW using multiple modalities: image, metadata, or both, after pretraining on fMoW for 25 or 200 epochs.

		fMoW F1	
Model	**Modality**	**25 epochs**	**200 epochs**
SatMIP	Image	49.5$_{\pm 0.3}$	57.6$_{\pm 0.5}$
	Meta	18.7$_{\pm 0.6}$	19.5$_{\pm 0.3}$
	Image+Meta	**54.0**$_{\pm \mathbf{0.5}}$	**59.3**$_{\pm \mathbf{0.6}}$
SatMIPS	Image	57.7$_{\pm 0.1}$	66.3$_{\pm 0.4}$
	Meta	19.4$_{\pm 1.2}$	19.6$_{\pm 0.2}$
	Image+Meta	**60.5**$_{\pm \mathbf{0.9}}$	**67.0**$_{\pm \mathbf{1.5}}$

(a) Textual metadata encoder

		fMoW F1	
Model	**Modality**	**25 epochs**	**200 epochs**
SatMIP	Image	50.7$_{\pm 0.4}$	59.3$_{\pm 0.3}$
	Meta	27.8$_{\pm 0.1}$	27.8$_{\pm 0.2}$
	Image+Meta	**57.7**$_{\pm \mathbf{0.5}}$	**63.1**$_{\pm \mathbf{0.1}}$
SatMIPS	Image	59.5$_{\pm 0.1}$	65.8$_{\pm 0.1}$
	Meta	27.9$_{\pm 0.1}$	27.8$_{\pm 0.1}$
	Image+Meta	**64.6**$_{\pm \mathbf{0.2}}$	**68.6**$_{\pm \mathbf{0.2}}$

(b) Tabular metadata encoder

H Additional results

We report additional results corresponding to the experiments presented in the main paper.

We have analyzed the time and memory efficiency of method relatively to our baseline (SimCLR). In Tab. S8, we report the absolute numbers, corresponding to the 200 epochs pretraining runs in Fig. 3 of the main paper.

Table S8: Absolute training time and peak memory usage of the different pretraining methods, for 200 epochs, with ViT-S backbone and a batch size of 1024 distributed over 4 Nvidia V100-32GB. The reported training times are averages of 3 runs.

Model	Training time (minutes)	Peak memory usage (GiB)
SimCLR	392	17.3
SatMIP	222	10.7
SatMIPS	414	19.2

I CO_2 emissions related to experiments

Experiments performed throughout this project consumed a total of 5,123 hours of V100-SXM2-32GB compute and 9358 hours of A100-SXM4-80GB compute. We performed our experiments on the Jean Zay HPC cluster from IDRIS, located in Orsay, France. As reported by our HPC cluster monitoring tool, the experiments amount to a total of 0,504 T CO_2eq.