SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised pre-training methods for large vision models have shown to enhance performance on downstream supervised tasks. Developing similar techniques for satellite imagery presents significant opportunities as unlabelled data is plentiful and the inherent temporal and multi-spectral structure provides avenues to further improve existing pre-training strategies. In this paper, we present SatMAE, a pre-training framework for temporal or multi-spectral satellite imagery based on Masked Autoencoder (MAE). To leverage temporal information, we include a temporal embedding along with independently masking image patches across time. In addition, we demonstrate that encoding multi-spectral data as groups of bands with distinct spectral positional encodings is beneficial. Our approach yields strong improvements over previous state-of-the-art techniques, both in terms of supervised learning performance on benchmark datasets (up to $\uparrow 7\%$), and transfer learning performance on downstream remote sensing tasks, including land cover classification (up to $\uparrow 14\%$) and semantic segmentation.

1 Introduction

In recent years, self-supervised learning techniques have quickly become the norm for pre-training models on large-scale natural image datasets [1, 2, 3, 4, 5, 6, 7, 8], and have demonstrated strong performance on downstream tasks including image classification [3, 4, 9, 10], image segmentation [3, 11], representation learning [12, 13, 14], image compression [12, 15], and image reconstruction [1]. Unlike supervised learning approaches, self-supervised learning techniques do not require human labeling, making them appealing in settings where unlabeled data are abundant but labeled data are scarce, such as remote sensing data (e.g., satellite imagery). While several large-scale satellite image datasets have been carefully curated in the past few years, including Functional Map of the World (fMoW) [16], BigEarthNet [17], xView [18], SpaceNet [19], annotating these datasets requires specialized skills and is more expensive than traditional computer vision datasets. Moreover, automatic analysis of satellite imagery is often needed for tasks with large societal impact such as poverty or crop yield prediction[20, 21, 22, 23, 24, 25], where acquiring large amounts of labeled data through surveys is impossible or prohibitively expensive. This suggests that self-supervised learning approaches for satellite imagery could be especially valuable.

However, existing self-supervised learning approaches [1, 2, 3, 4, 5, 6] are mainly designed for natural images. As opposed to natural images such as ImageNet [26], satellite imagery is usually associated with meaningful geographical and temporal information, and can consist of multiple spectral bands representing various sensor readings besides visible light (i.e., RGB channels typical in natural images). Depending on the data source, satellite imagery can also vary significantly in resolution. While self-supervised learning methods for satellite imagery exist [27, 28], these approaches cannot learn general representations for both temporal and multi-spectral remote sensing data.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

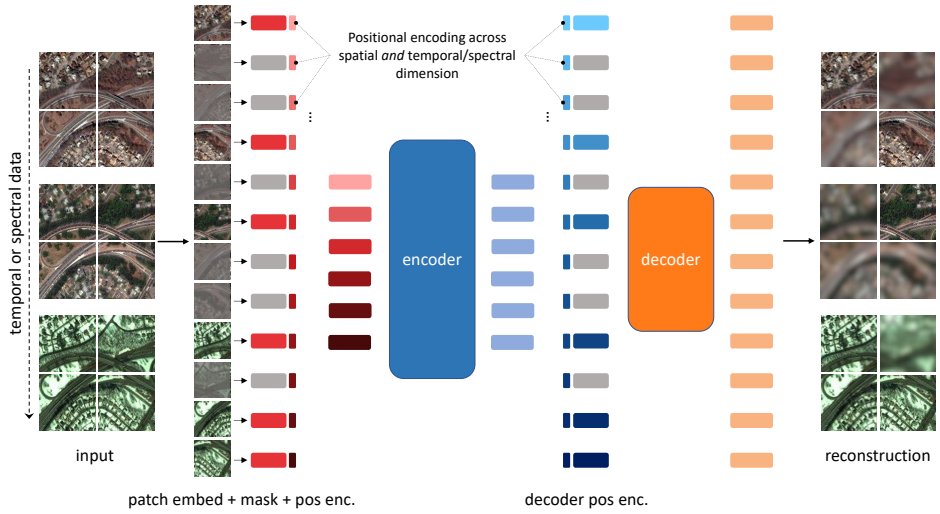


Figure 1: Our proposed SSL framework on temporal and multi-spectral satellite imagery. With carefully-designed masking strategy across image layers of different time or spectrum, and temporal and spectral encoding, our SatMAE serves as a powerful vision learner for remote sensing tasks.

To address this issue, in this paper, we propose **SatMAE**, a self-supervised learning framework based on masked autoencoders (MAEs) [1] which naturally handles temporal and multi-spectral input data. We show that introducing a positional encoding for the temporal/spectral dimension and *independently* masking patches across the temporal/spectral dimension benefits pre-training, allowing the model to learn representations of the data that are more conducive to fine-tuning. Specifically, our contributions are as follows

1. We propose a novel method to leverage temporal or multi-spectral information in satellite imagery to improve self-supervised pre-training with masked autoencoders (see 4).
2. We introduce a new dataset called fMoW-Sentinel based on Sentinel-2 imagery, cross-referenced with fMoW, to serve as a benchmark for training models on multi-spectral satellite imagery (see A.1).
3. We demonstrate the effectiveness of pre-training transformers [29] on satellite imagery, achieving significant improvement over previous state-of-the-art methods on benchmark datasets as well as downstream remote sensing tasks (see 5)

2 Related Work

SSL for Satellite Imagery Self-supervised learning approaches [4, 3, 2, 5, 6] have emerged as promising approaches in remote sensing domains. For instance, [27] and [28] propose to incorporate spatially aligned images over time into contrastive self-supervised learning. Despite the promising results, these two contrastive learning approaches rely heavily on the quality of positive pairs, which is often hard to control. Contrastive Sensor Fusion [30] constructs positive pairs by using co-located images generated by combining different sensor channels. [31, 32, 33] apply off-the-shelf contrastive learning algorithms to satellite images. [33] utilizes image inpainting and transformation prediction as additional pretext tasks. [34] leverages geographical knowledge to aid SSL, which, however, can be difficult to obtain as annotations.

Masked Autoencoder MAE [1] is a powerful self-supervised learning method built upon the Vision Transformer [29] architecture. Instead of constructing a contrastive objective, it proposes the pretext task of reconstructing masked patches of the input, which hardly needs any data augmentation. MAE has achieved state-of-the-arts performance on a wide collection of vision benchmarks. Inspired by the success of [1], many follow-up works extend MAE to different data modalities. VideoMAE [35] proposes video tube masking and reconstruction as a pretext task for video analysis. It masks the

same region across all frames of a given video. GMAE [36] adapts MAE to the domain of graphs. MultiMAE [37] enhance the flexibility of MAE by enabling it to take optional input of different modality and correspondingly adding other training objectives to facilitate multi-modality learning. However, these works fail to handle temporal and multi-spectral input. VideoMAE takes image series as input but requires the frames to be equally-spaced in the temporal dimension, which is not the case for remote sensing data due to the temporal irregularity and discontinuity. In this work, we incorporate temporal and spectral information into a masked autoencoder architecture, and propose a novel self-supervised framework for remote sensing domains.

3 Background

Masked Autoencoder The Masked Autoencoder (MAE) [1] operates on images of the form $I \in \mathbb{R}^{C \times H \times W}$, where H, W are the height and width of the image, respectively, and C is the number of channels (e.g., 3 for RGB images).

The MAE is an autoencoder with asymmetrical encoding and decoding stages. The input image I is resized to a sequence of non-overlapping patches, $S \in \mathbb{R}^{L \times P^2 C}$, where P is the height and width of the patch, and $L = (H/P) \cdot (W/P)$ is the number of patches. Each patch is passed through a patch embedding $f_p : \mathbb{R}^{P^2 C} \mapsto \mathbb{R}^D$ to create a sequence $S' \in \mathbb{R}^{L \times D}$ of embedded patch "tokens". A fraction p_m of the L tokens are masked (usually $p_m = 0.75$) and only the remaining $(1 - p_m)L$ tokens corresponding to "visible" patches are fed to the encoder, a Vision Transformer (ViT) [29] with positional embeddings to denote the spatial location of the patch in the image. The decoder is a series of transformer blocks that operates on all L tokens (with positional embeddings added), where the $p_m L$ encoded visible patches are placed in their original sequence position among $(1 - p_m)L$ masked patches represented by a learnable mask token. The decoder then outputs a reconstructed image $\hat{I} \in \mathbb{R}^{C \times H \times W}$, which is compared to the original input image using the mean-squared error (MSE) loss, computed per-pixel only on the masked patches [1].

Positional encoding Positional encoding is an important feature of transformers that makes the learned representation position-aware. In MAE [1] and many transformers [38, 39], the positional encoding is formulated as:

$$\texttt{Encode}(k, 2i) = \sin \frac{k}{n^{\frac{2i}{d}}}, \ \texttt{Encode}(k, 2i+1) = \cos \frac{k}{n^{\frac{2i}{d}}}$$

Here, k is the position, i is the index of feature dimension in the encoding, d is the number of possible positions, and n is a large constant (normally set to 10000). In MAE, position is defined as the index of the patch along the x and y axes, respectively. Therefore, k ranges from 0 to H/P (or W/P). The final encoding is generated by concatenating the encodings of the x and y coordinates.

4 Method

In this section, we describe the design of SatMAE, and how we utilize temporal (4.1) and multi-spectral (4.2) satellite images as input.

4.1 Temporal SatMAE

We now consider input tensors of the form $I_T \in \mathbb{R}^{T \times C \times H \times W}$, where T denotes the number of images in a temporal sequence. In video data, T is the number of frames, which are usually equally spaced. However, temporal satellite imagery rarely has images at regular intervals. More commonly, several snapshots, or versions, of a given location are taken at irregular times. The length and sample frequency of these sequences of satellite images vary drastically over years and across different regions. For a given spatial location loc, we define T_{loc} as the number of temporally distinct snapshots present in the dataset.

The naïve way to handle an extra dimension would be to reshape I_T to $I'_T \in \mathbb{R}^{TC \times H \times W}$, effectively concatenating the temporal sequence of images along the spectral (i.e. channel) dimension, and then apply the MAE machinery verbatim. This method poses a few difficulties: (i) the model may be unable to generalise to a temporal ordering different to the one used in pre-training, since it can

only understand order through the position of images in the stacked-timeseries (ii) the model cannot reason about the length of time separating two consecutive images in a time sequence, which may be variable when images of a location are sampled at irregular intervals (iii) the model loses access to temporal fine-grained information in deeper layers, as its only direct exposure to encode temporal information is through the initial patch embedding f_p (iv) the resulting model is not temporally-shift invariant (i.e. the model would need to separately learn to detect the same event in two different segments of the temporal sequence).

To address these challenges and to avoid losing temporal information, we resize the temporal sequence I_T to $S_T \in \mathbb{R}^{L_T \times P_T P^2 C}$, where $L_T = L \cdot (T/P_T) = (H/P) \cdot (W/P) \cdot (T/P_T)$, P_T is the "patch size" in the temporal dimension, and L and P are defined in 3. Prior works using transformers for video data suggest using $P_T = 2$, where each "patch" is a cube of shape $2 \times 16 \times 16$ [35, 40, 41]. Since our data has much shorter temporal sequence lengths [16], we let $P_T = 1$ such that $L_T = L \cdot T$. In order to operate on inputs of any temporal order, we re-use the same patch embedding $f_p : \mathbb{R}^{P^2 C} \mapsto \mathbb{R}^D$ for each image in the time series, giving us an embedded sequence of tokens $S'_T \in \mathbb{R}^{L_T \times D}$.

4.1.1 Temporal Encoding

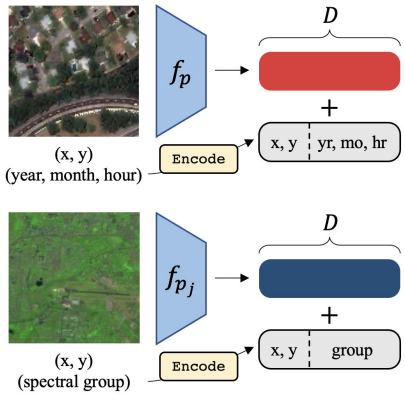


Figure 2: Top: Encoding each temporal patch with a shared patch embedding f_p. Bottom: Encoding each spectral patch with a different patch embedding f_{p_j} for each group j.

For each embedded token in the L_T length sequence, we need to ensure the model retains information about its spatial and temporal position. As shown in many prior works [27, 28], the timestamp of a satellite image is useful for many pre-training or downstream vision tasks. We propose a temporal encoding scheme compatible with the masked autoencoder architecture by treating the temporal dimension similarly to the positional dimensions (see 3).

The timestamp of a satellite image is represented as "year-month-day-hour-minute-second". Instead of passing the entire numerized timestamp into a feature encoder, we propose only keeping the useful parts. Intuitively, the day, minute, and second should be unrelated to the visual appearance of a region. Thus, including these components in the temporal encoding may not be beneficial, and can even be detrimental. In contrast, a landscape may evolve over years due to weather, geology, and human activity. The month reflects season and climate, and the hour reflects daylight and temperature.

Then, the temporal encoding is formulated as:

$$t_{k,i} = \text{CONCAT}[\texttt{Encode}(k_{\text{year}}, i), \texttt{Encode}(k_{\text{month}}, i), \texttt{Encode}(k_{\text{hour}}, i)]$$

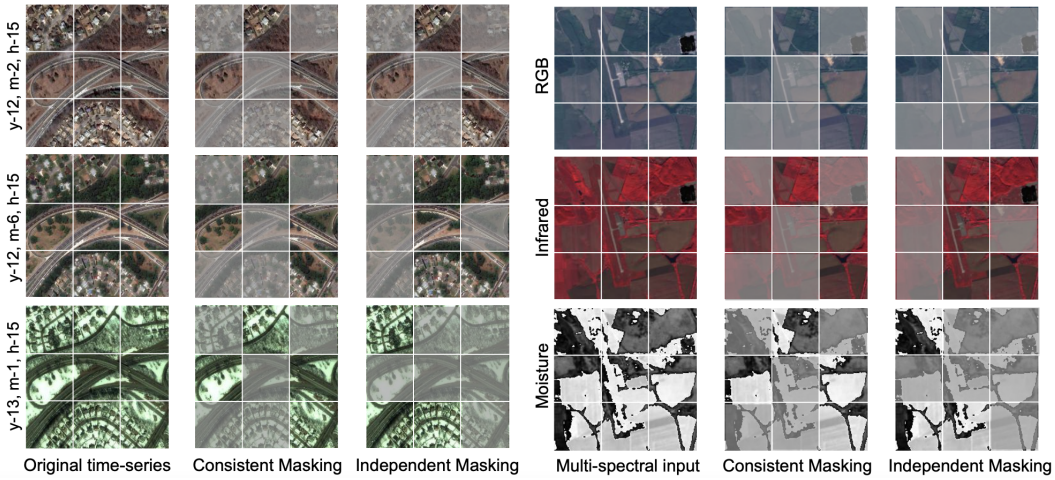
And the final encoding is generated by concatenating the temporal encoding to the positional encoding defined in 3 such that the total length of the encoding is D.

4.1.2 Masking Strategies

With an additional temporal dimension, masking a subset of the L_T tokens needs to be treated with care. As seen in figure 3, there are different ways to mask a temporal series of satellite images.

Consistent Masking Each image is "patchified" separately, but the masked regions are kept consistent across all images (fig. 3a). This approach is also used in VideoMAE [35], with video input.

Independent Masking Each image is "patchified" separately, and masked regions may not be the same across every image. Instead, a fraction p_m of the full sequence of all patch tokens are masked. Another variant is to independently mask the regions of each image, but keep the ratio p_m of masked regions fixed per image. Both variants are equivalent in expectation. Effectively, the model can look at unmasked values of a region that is masked in one image but not in others. While this setting may lead to an easier task for video data since the model can "cheat" and exploit temporal redundancy in videos with high framerates [35], we argue that this form of "cheating" is less feasible in temporal



(a) Temporal data (b) Spectral data

Figure 3: 3a **Temporal masking**: When masking the patches for images in a timeseries, we can choose to keep a patch fully visible or fully masked across time (*consistent masking*), or independently mask all patches (*independent masking*). In both cases, a fraction p_m patches are masked. Here, $T = 3$, and the leftmost column orders the temporal sequence according to the timestamp features. For example, "y-12, m-12, h-15" denotes 12 years from the minimum year (2002), the zero-indexed month 2, and the 15th hour of the day; i.e., roughly 2014, March, 15:00. 3b **Spectral Masking**: The same masking strategies are adapted to groups of the 13 spectral bands in Sentinel-2 images.

satellite imagery, given the strong impact of seasonal variation and changing human activity over periods of time and the much larger time deltas between temporally consecutive images (see fig. 3a).

Independent Masking + Inconsistent Cropping During data pre-processing, we can crop square regions for input inconsistently so that images in the same temporal sequence may be spatially-unaligned. This strategy may help the model learn better representations as it may learn to align images in the sequence across the spatial and temporal dimensions.

4.2 Multi-spectral SatMAE

While MAE does operate on images $I \in \mathbb{R}^{C \times H \times W}$, usually $C = 3$ for RGB images. Satellite data, on the other hand, can often have multiple spectral bands. For example, Sentinel-2 imagery has $C = 13$ bands of 10m, 20m and 60m spatial resolution, each of different wavelengths (see A.2.2). In the following, we discuss and later experimentally compare various ways to encode spectral information.

Stack Channels This method treats the images I with multiple bands as-is, and encodes each patch of C channels to a D dimensional token embedding. Therefore, the sequence of patches $S \in \mathbb{R}^{L \times P^2 C}$ is embedded to a sequence of tokens $S' \in \mathbb{R}^{L \times D}$. We denote this method **SatMAE+Stack**.

Group Channels There are limitations to naively stacking the spectral information, especially that a single convolutional patch embedding may be insufficient to fully capture fine-grained information present in multiple bands of different wavelengths and spatial resolution. We would like the model to preserve information about the different bands through the encoding and decoding stages.

To address this limitation, we propose *grouping* subsets of spectral bands. Given C channels, we form G groups $g_1, g_2, \ldots, g_G$ such that $g_1 + g_2 + \cdots + g_G = C$. This is analogous to slicing the image I in the channel dimension, creating images $I_1, \ldots, I_G$, where $I_j \in \mathbb{R}^{g_j \times H \times W}$. Since for some i, j, it is possible that $g_i \neq g_j$, we use a separate patch embedding $f_{p_j} : \mathbb{R}^{P^2 G_j} \mapsto \mathbb{R}^D$ for each group j, thus allowing the model to best represent each group of channels as token embeddings. Therefore, each group j is first resized from $I_j \in \mathbb{R}^{G_j \times H \times W}$ to $S_j \in \mathbb{R}^{L \times P^2 G_j}$, and then each

patch is embedded with f_{p_j} to produce a sequence of embedded tokens $S'_j \in \mathbb{R}^{L \times D}$. The sequences $S'_1, \ldots, S'_G$ are concatenated to produce the final set of tokens $S' \in \mathbb{R}^{GL \times D}$.

Spectral Encoding Since the tokens in S' now correspond both to a patch location (i, j) in the original input image, as well as a group of channels g, we include an encoding for the group id as well following a similar form as 4.1.1. For a given group index k_g, we have the group encoding

$$g_{k,i} = \texttt{Encode}(k_g, i)$$

The final encoding is a concatenation of the positional encoding $x_{k,i}, y_{k,i}$ and the spectral encoding $g_{k,i}$ such that the total encoding dimension is D (see figure 2). This positional encoding is added to S' before inputting it to the encoder. We denote the combined setting of grouping channels and using a group encoding as **SatMAE+Group**.

Masking Strategies We consider *consistent masking* (denoted **SatMAE+Group+CM**) and *independent masking* (**SatMAE+Group+IM**) as defined in 4.1.2 and as visualized in figure 3b.

5 Experiments

In this section, we start by introducing the datasets we considered for our experiments. Specifically, we collect a new multi-spectral remote sensing image dataset for downstream task evaluation (A.1). We then present our results on various remote sensing transfer-learning and downstream tasks. In summary, our approach demonstrates strong performance on all the tasks we considered, yielding improvements over previous state-of-the-art techniques by up to 6% on supervised learning benchmarks, and up to 14% on remote sensing transfer-learning downstream remote sensing tasks.

For all experiments, we compare with both the current state-of-the-art methods [27, 28] and with supervised learning from scratch using the ViT backbone of our model.

5.1 Datasets for Pre-training

fMoW RGB Functional Map of the World (fMoW) [16] is a dataset of high-resolution satellite image time series across the world, with a task of classification among 62 categories such as airport, shipyard, and zoo.

fMoW Sentinel We create a new dataset based on the fMoW RGB dataset. We collect all 13 frequency bands provided by Sentinel-2 (B1-12 and B8A) for the original fMoW locations, at some of the same times as fMoW images plus some extra times, for a total of 712,874 training images, 84,939 validation images, and 84,966 test images. More details about the preprocessing of the two datasets are included in A.1.

5.2 fMoW RGB (non-temporal)

In this section, we perform experiments on fMoW single image classification task. Following [27], we report both the performance of linear probing and finetuning setting. Table 1 shows that compared to the previous state-of-the-art self-supervised method using a contrastive momentum encoding approach [27, 3], our SatMAE achieved a 6.29% improvement in top 1 classification accuracy. An interesting fact is that without SatMAE pre-training, the ViT-large model could only reach 62.48% at convergence after 50 epochs of finetuning compared to 69.05% achieved by training a ResNet-50 model from scratch . This is likely because the ViT [29] backbone is harder to train than ResNet50 [42], which makes the pre-trained model more valuable in this case.

5.3 fMoW RGB (temporal)

Main experiments We perform image-sequence classification on the temporal version of fMoW RGB to evaluate our temporal SatMAE. The temporal fMoW consists of co-located image sequences with a length of 3. As shown in Table 2, SatMAE surpasses the previous state-of-the-art by 4.48%, and also improves the non-temporal result by 2.06% in terms of top 1 classification accuracy. We can observe from the second last row that this gain is not from the larger model to handle sequences

Method	Backbone	Frozen/Finetune
Sup. (Scratch)	ResNet50	-/69.05
Sup. (IN init.)	ResNet50	-/69.07
GASSL [27]	ResNet50	68.32/71.55
Sup. (Scratch)	ViT-Large	-/62.48
Sup. (IN init.)	ViT-Large	-/76.91
SatMAE (Ours)	ViT-Large	65.94/**77.84**

Table 1: Top 1 Accuracy on fMoW classification. **Frozen** means only performing linear classification on frozen features of the pre-trained model. **Finetune** corresponds to end-to-end finetuning the whole model. **Scratch** means training from scratch, and **IN init.** means using ImageNet pre-trained weights.

Method	Backbone	Top Acc. (1/5)
Sup. (Scratch)	ResNet50	73.24/-
GASSL [27]	ResNet50	74.11/-
SatMAE+Stack	ViT-Large	75.85/88.68
Sup. (Scratch)	ViT-Large	61.89/84.23
SatMAE*	ViT-Large	76.78/92.01
SatMAE	ViT-Large	**81.49/93.26**

Table 2: Classification results on the temporal fMoW RGB dataset. * means copying the same image for three times instead of using image sequences as input. SatMAE+Stack here means stacking the image sequence along the channel space.

Method	Backbone	Top Acc. (1/5)
Sup. Learning*	ResNet152	49.12/75.73
Sup. Learning†	ResNet152	54.46/78.99
SatMAE+Group*	ViT-Large	53.03/77.14
SatMAE+Group†	ViT-Large	51.61/77.26
SatMAE+Group‡	ViT-Large	49.49/76.30
SatMAE+Stack	ViT-Large	57.37/81.63
SatMAE+Group+CM	ViT-Large	57.76/80.49
SatMAE+Group+IM	ViT-Large	**59.30/82.81**

Table 3: Top 1 and Top 5 Accuracy on the fMoW Sentinel validation set. In each row, the full model is finetuned. * denotes finetuning from scratch. † is finetuning from ImageNet weights, ‡ is finetuning from fMoW RGB weights. The remaining rows use fMoW Sentinel for pre-training.

Temp. Enc.	Indep. Mask.	Cons. Crop.	Test Aug.	Top Acc. (1/5)
	✓	✓		78.07/92.73
✓		✓		78.45/91.64
✓	✓			79.90/92.75
✓	✓	✓		79.69/92.09
✓	✓	✓	✓	**81.49/93.26**

Table 4: Ablation studies on different components of temporal SatMAE on the temporal fMoW classification task. The first column is whether using temporal encoding, the second is whether using independent masking, the third is whether cropping consistently, and the last one is whether applying test-time augmentation.

of data. Naively stacking the image sequences in the channel dimension performs even worse than the non-temporal SatMAE. Again, the pre-trained model is crucial for the ViT to achieve high performance over ResNet50.

Ablation studies Table 4 provides a comprehensive ablation study on the components of our temporal SatMAE. We can see that the performance mainly comes from using temporal encoding and adopting the independent masking strategy instead of consistent masking as suggested in VideoMAE [35]. Interestingly, the consistent cropping option slightly hurts the performance, indicating the model does not rely on image sequences to be perfectly spatially-aligned. In addition, applying test-time augmentations similar to [27] brings considerable improvement.

5.4 fMoW Sentinel (Multi-spectral)

In this section, we describe experiments pre-training on fMoW-Sentinel imagery and finetuning models on the image classification task of the fMoW-Sentinel dataset. We pre-train SatMAE+Stack 4.2 and investigate consistent masking 4.1.2 and independent masking 4.1.2, with SatMAE+Group+CM and SatMAE+Group+IM, respectively (see 4.2, 4.2, 4.2). The full models are then finetuned on the fMoW-Sentinel image classification task. For comparison, we also consider finetuning the ResNet-152 model [42] from scratch and from an ImageNet initialisation. We pick the largest model, ResNet-152, for fairer comparison with ViTs.

Model configuration As not all of the 13 Sentinel-2 bands may be useful, in our experiments we drop bands B1, B9 and B10, which correspond to a spatial resolution of 60m. Of the remaining 10 bands, we form three groups: (i) RGB+NIR: B2, B3, B4, B8 (ii) Red Edge: B5, B6, B7, B8A (iii) SWIR: B11, B12. We choose this grouping to ensure each group has bands of the same spatial resolution as well as similar wavelength (see A.2.2). See A.3.2 for pre-training and finetuning details.

Results We present the results in table 3. Our method SatMAE+Group+IM is able to achieve the highest accuracy, outperforming supervised training from scratch by 6.27%. Interestingly,

Method	Backbone	Top 1 Acc.
Sup. (Scratch)	ResNet50	54.46
GASSL [27]	ResNet50	57.63
Sup. (Scratch)	ViT-Large	69.65
SatMAE	ViT-Large	**71.77**

Table 5: NAIP land cover classification results.

Method	Backbone	mIoU
Sup. (Scratch)	ResNet50	75.57
GASSL [27]	ResNet50	**78.51**
Sup. (Scratch)	ViT-Large	74.71
SatMAE	ViT-Large	78.07

Table 6: SpaceNet v2 building segmentation results.

Method	Backbone	Top 1 Acc.
Sup. (Scratch)	ResNet18	63.21
Sup. (IN init.)	ResNet18	86.44
GASSL [27]	ResNet18	89.51
SeCo [28]	ResNet18	93.14
SatMAE*	ViT-Large	95.74
SatMAE*	ViT-Large	98.94
SatMAE+Group	ViT-Large	96.54
SatMAE+Group+IM	ViT-Large	**98.98**

Table 7: EuroSAT land cover classification results. * means we only use the RGB channels of the data.

Method	Backbone	mAP
Sup. (Scratch)	ResNet50	69.49
Sup. (IN init.)	ResNet50	80.04
GASSL [27]	ResNet50	80.20
SeCo [28]	ResNet50	**82.62**
Sup. (Scratch)	ViT-Large	80.07
SatMAE	ViT-Large	82.13

Table 8: BigEarthNet multi-label classification results. Following [28], we choose mean Average Precision (mAP) as the evaluation metric, and use a newer set of class labels.

independent masking performs the best, suggesting that it is beneficial to allow the model to "peek" at regions of unmasked band groups when reconstructing the same region in a masked band group. While ImageNet-initialised backbones also appear to be effective,

5.5 Transfer Learning Experiments

In this section, we use our pre-trained weights to finetune SatMAE on a few downstream tasks on remote-sensing datasets, including land cover classification (5.5), multi-label classification (5.5), and building segmentation (5.5).

Land Cover Classification We perform transfer learning experiments on land cover classification using the NAIP and EuroSAT [43] dataset. NAIP consists of RGB+CIR images of 66 land cover classes obtained by the USDA's National Agricultural Imagery Program, which are split into 100,000 training and 50,000 validation images. EuroSAT is a small dataset containing 27,000 13-band satellite images of 10 classes based on Sentinel-2. We follow [28, 44] for the train/val splits on EuroSAT.

Table 5 and table 7 shows the remarkable improvement of our SatMAE over the state-of-the-arts. Although using the ViT-Large backbone already achieved good results, initiating the model with SAT-MAE pre-trained weights further increased the accuracy by 2%-3%.

Multi-label Land Cover Classification We also use the BigEarthNet [17] dataset for multi-label classification. BigEarthNet consists of 13-band Sentinel-2 images of multiple labels from 19 classes. There are 354,196 images for training and 118,065 images for validation. Following [28], we use a 10% subset of the train set.

Table 8 shows our SatMAE pre-training improves upon the model trained from scratch by over 2%, and achieves comparable results to the state-of-the-art. GASSL and SeCo were actually trained on a larger pre-train dataset (1M Sentinel-2 images v.s. 713k) than our fMoW Sentinel. Therefore we expect further improvement when we pre-train SatMAE with more data and for longer.

Building Segmentation We perform semantic segmentation as downstream task on the SpaceNet v2 dataset [19]in this section. The SpaceNet v2 datasets consists of 6940 high resolution satellite images with segmentation masks for buildings, which are divided into train and test sets of 5000 and 1940 images, respectively.

The results in Table 6 show that our method is able to achieve a larger performance gain compared to the supervised learning from scratch setting compared to [27]. The incompatibility of the transformer backbone with PSANet probably explains why the baseline performance is not as good as that of using ResNet50 as backbone.

5.6 Visualizing reconstruction quality for SatMAE

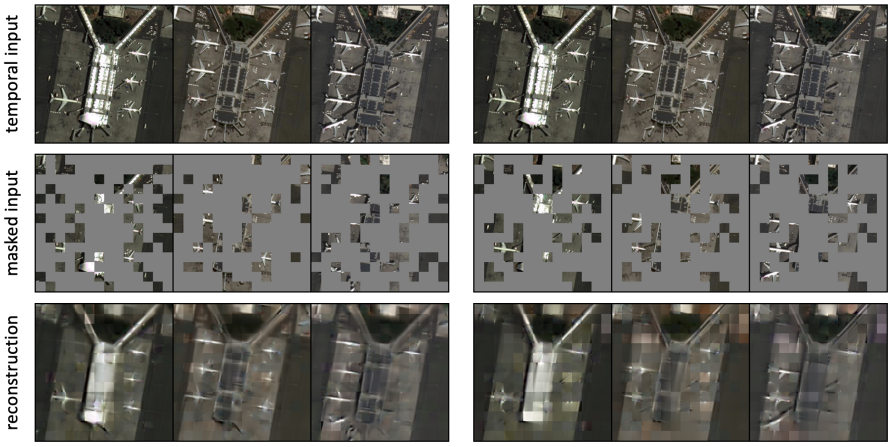


Figure 4: Comparison of reconstruction quality between SatMAE+IM (left) and SatMAE+CM (right).

We show the visualization of the reconstruction quality of two different SatMAE masking strategies in Figure 4. The SatMAE with Independent Masking successfully reconstructs all the airplanes even though their number varies across time. In contrast, the SatMAE with Consistent Masking missed some airplanes in the reconstruction.

6 Conclusion

In this paper, we propose a new SSL framework based on the MAE architecture [1] tailored to remote-sensing data (satellite imagery). Our novel masking strategy in a joint positional, temporal and spectral space, along with the temporal and spectral encoding, enables our model to handle temporal and multi-spectral satellite images as input and learn useful representations. Experiments on the datasets for pre-training and multiple downstream datasets demonstrate the effectiveness of our pre-trained SatMAE model, outperforming previous state-of-the-art results by large margins.

In the future, it would be useful to design more efficient transformer architectures. While SatMAE has a similar number of parameters for both the temporal and multi-spectral setting as a regular ViT, the increased sequence length can strain computational resources. Moreover, it is also worth exploring optimal positional encodings for spectral and temporal data, as well as optimal groups of spectral bands, either by neural-based search methods, or using prior knowledge. Lastly, investigating better architectures for object detection and semantic segmentation using ViTs will be important in generalising SatMAE to further downstream tasks.

Broader Impact

Accurate measurements of economic, social, and environmental phenomena are key inputs into policy decisions made around the world, but the sparsity of labelled data on many outcomes means that such decisions are often not guided by timely or accurate data. We demonstrate how a pre-training framework could relieve the dependence on labelled data for many downstream tasks that use satellite imagery as input. We hope our SatMAE method will help close the gap between SSL performance on natural imagery and on the more challenging satellite imagery, and prompt further attention from the ML community on the usefulness of SSL in satellite-imagery-related tasks.

Better extraction of information from satellite imagery has profound implications for our ability to measure and understand a broad array of social, economic and environmental phenomena that are critical for decision making. Our approach further amplifies the usefulness of the sparse amount of labelled data that exist on key human outcomes, and could enable rapid and accurate extraction of imagery features relevant for critical downstream tasks, including poverty prediction, infrastructure development, and population estimation. Such information could aid governments in more rapid and data-informed decision making and ultimately bring large societal benefits.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We tried our best to be precise.
   (b) Did you describe the limitations of your work? [Yes] We described the limitations in the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed in Appendix B.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We list the url to our code repo in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We list part of them in the experiments section and part of them in Appendix A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We trained on large datasets with small variation across different runs expected. Limited by computation resources, we only ran each experiment once.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We list that in Appendix A.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes, we do that by citing creators of datasets and authors of prior works.
   (b) Did you mention the license of the assets? [Yes] We do that in Appendix A.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We are releasing a new dataset. The instructions and links will be released in our codebase mentioned above.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All data we used are released publicly.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We concluded that no data we are using has such concerns.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.

[2] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[3] 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] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[7] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *CoRR*, abs/2011.00362, 2020.

[8] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[9] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[11] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[12] Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019.

[14] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10364–10374, 2019.

[15] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33:12980–12992, 2020.

[16] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

[17] 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.

[18] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.

[19] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[20] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

[21] Kumar Ayush, Burak Uzkent, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Efficient poverty mapping from high resolution remote sensing images. In *Proc. AAAI Conf. Artif. Intell*, volume 35, pages 12–20, 2021.

[22] Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*, 2020.

[23] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[24] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI conference on artificial intelligence*, 2017.

[25] Anna X Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–5, 2018.

[26] 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, 2009.

[27] 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.

[28] Oscar Mañas, Alexandre Lacoste, Xavier Giro-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.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[30] Aidan M Swope, Xander H Rudelis, and Kyle T Story. Representation learning for remote sensing: An unsupervised sensor fusion approach. *arXiv preprint arXiv:2108.05094*, 2021.

[31] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1182–1191, 2021.

[32] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Multi-modal self-supervised representation learning for earth observation. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3241–3244. IEEE, 2021.

[33] Wenyuan Li, Hao Chen, and Zhenwei Shi. Semantic segmentation of remote sensing images with self-supervised multitask representation learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6438–6450, 2021.

[34] Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[35] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[36] Hongxu Chen, Sixiao Zhang, and Guandong Xu. Graph masked autoencoder. *arXiv preprint arXiv:2202.08391*, 2022.

[37] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[40] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*, 2021.

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[42] 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.

[43] 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.

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