



**HAL**  
open science

# Spatiotemporal self-supervised pre-training on satellite imagery improves food insecurity prediction

Ruben Cartuyvels, Tom Fierens, Emiel Coppieters, Marie-Francine Moens,  
Damien Sileo

## ► To cite this version:

Ruben Cartuyvels, Tom Fierens, Emiel Coppieters, Marie-Francine Moens, Damien Sileo. Spatiotemporal self-supervised pre-training on satellite imagery improves food insecurity prediction. *Environmental Data Science*, 2023, 2, pp.e48. 10.1017/eds.2023.42 . hal-04411802

**HAL Id: hal-04411802**

**<https://hal.science/hal-04411802v1>**


Submitted on 23 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

METHODS PAPER  

# Spatiotemporal self-supervised pre-training on satellite imagery improves food insecurity prediction

Ruben Cartuyvels , Tom Fierens, Emiel Coppeters, Marie-Francine Moens and Damien Sileo

Department of Computer Science, KU Leuven, Leuven, Belgium

**Corresponding author:** Ruben Cartuyvels; Email: [ruben.cartuyvels@kuleuven.be](mailto:ruben.cartuyvels@kuleuven.be)

**Received:** 01 February 2022; **Revised:** 25 September 2023; **Accepted:** 06 November 2023

**Keywords:** deep learning; food insecurity; remote sensing; unsupervised pre-training

## Abstract



Global warming will cause unprecedented changes to the world. Predicting events such as food insecurities in specific earth regions is a valuable way to face them with adequate policies. Existing food insecurity prediction models are based on handcrafted features such as population counts, food prices, or rainfall measurements. However, finding useful features is a challenging task, and data scarcity hinders accuracy. We leverage unsupervised pre-training of neural networks to automatically learn useful features from widely available Landsat-8 satellite images. We train neural feature extractors to predict whether pairs of images are coming from spatially close or distant regions on the assumption that close regions should have similar features. We also integrate a temporal dimension to our pre-training to capture the temporal trends of satellite images with improved accuracy. We show that with unsupervised pre-training on a large set of satellite images, neural feature extractors achieve a macro F1 of 65.4% on the Famine Early Warning Systems network dataset—a 24% improvement over handcrafted features. We further show that our pre-training method leads to better features than supervised learning and previous unsupervised pre-training techniques. We demonstrate the importance of the proposed time-aware pre-training and show that the pre-trained networks can predict food insecurity with limited availability of labeled data.

## Impact Statement

This study shows that satellite images and deep learning can be used to drastically improve predictions of food insecurity compared to existing predictors in countries or regions where food insecurity is mainly caused by agricultural or weather-related factors. Vast amounts of unlabeled and publicly available satellite image data can be used to pre-train a neural network using the method proposed in this study. This further improves predictions, but also decreases the amount of labeled food insecurity data needed for training in order to obtain accurate predictions. This is useful since accurate food insecurity data to train models might be hard or costly to obtain. To increase the impact of this work, it would be valuable to research how to improve forecasts of food insecurity in the future, which remains hard.

## 1. Introduction

Satellite imagery has been a precious source of information for many different fields and for many years. Satellite images are, for instance, essential for weather prediction, agricultural observations,

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

oceanography, cartography, biodiversity monitoring, and many more. Since the first orbital satellite images obtained in 1959, there are now over 150 earth observation satellites in orbit. With an abundance of satellite imagery available both across time and space, many studies (Mohanty et al., 2020) have searched for efficient ways to process this data to gain useful insights. In recent years, deep convolutional neural networks (CNNs) have increasingly been used to analyze such imagery (Jean et al., 2016; Kussul et al., 2017; Nevavuori et al., 2019; Yeh et al., 2020). However, training deep neural networks from scratch in a supervised way requires a large amount of labeled data, which is costly to obtain.

A variety of contrastive self-supervised pre-training methods has been proposed to deal with this problem (Jean et al., 2019; Ayush et al., 2021a; Kang et al., 2021; Manas et al., 2021). These methods pre-train neural networks on large amounts of unlabeled satellite imagery so they learn useful parameters. They typically are contrastive, which means they maximize the mutual information between pairs of similar samples (tiles of satellite imagery) while minimizing mutual information between dissimilar pairs. The learned parameters can then be used as a starting point for the supervised training of different downstream tasks, for which little labeled data might be available, and often prove to be more effective than using randomly initialized neural networks. Yet, existing methods completely ignore the temporal dimension of satellite imagery (Jean et al., 2019; Kang et al., 2021) or learn only highly time-invariant and highly spatially variant representations in a non-flexible manner (Ayush et al., 2021a; Manas et al., 2021).

This is problematic since downstream tasks may range from being highly variant to highly invariant against spatial, or independently, temporal distance. For instance, models for weather or rainfall forecasting might benefit from sensitivity to changes that typically occur on a timescale of days, while for land cover classification, it might be beneficial to abstract those exact same changes away and focus on changes occurring over years. This study explores the use of relational pre-training (Patacchiola and Storkey, 2020), a state-of-the-art contrastive pre-training method for satellite imagery. During pre-training, both similarities between the same satellite image tile over time and similarities between geographically neighboring image tiles are taken into account. Importantly, this framework allows the implementer to easily and independently specify the degree of temporal and spatial sensitivity needed for a certain downstream task by choosing thresholds that determine which pairs in the contrastive pre-training are considered similar and which pairs dissimilar.

We use freely available LANDSAT-8 imagery, from which we construct representations that serve as an input to predict food insecurity in Somalia. Although several studies explore the use of satellite imagery for predicting poverty, food insecurity is a relatively unexplored topic. Yet, in 2019, as much as 8.9% of the world's population was undernourished, and 10.10% lived in severe food insecurity (Roser and Itchie, 2019). Existing early-warning systems, as Andree et al. (2020) note, suffer from high false-negative rates. Therefore, automating and improving warning systems can be of great humanitarian value.

Our hypothesis is that useful information can be drawn from satellite imagery to predict Famine Early Warning Systems (FEWS) Integrated Phase Classification (IPC) food insecurity scores (Korpi-Salmela et al., 2012) due to, for instance, environmental changes and increasing droughts.

Our research questions are:

1. Can pre-trained representations of satellite images improve food insecurity prediction accuracy?
2. How do different temporal and spatial relationship prediction settings as pre-training influence downstream task performance?

We analyze the effect of relational pre-training on satellite imagery representations by comparing different temporal and spatial similarity thresholds. We compare the performance of our pre-trained model with a pre-trained baseline and with fully supervised networks for a range of training set sizes. We include the predictions of our model in the input for an existing food crises predictor (Andree et al., 2020) to test if

this improves performance. We test out-of-domain food insecurity prediction in regions that weren't included in pre-training data.

Our findings suggest that using spatially and temporally linked images as positive pairs for relational pre-training can outperform (1) a randomly initialized network without pre-training, (2) pre-training on standard data augmentations as in Patacchiola and Storkey (2020), (3) a network that has been pre-trained on ImageNet (Deng et al., 2009), and (4) a strong contrastive baseline pre-trained on the same satellite imagery (Jean et al., 2019). Our pre-trained model also outperforms a random forest classifier based on previously used manually selected features (Andree et al., 2020). We show that our pre-trained model needs little labeled data to learn to make good predictions and that the model's predictions are not reducible to predicting the season of the acquisition of a satellite image. We compare the importance of the input LANDSAT-8 bands. We find that forecasting future food insecurity remains hard for all of the considered methods.

## 2. Related work

### 2.1. Self-supervised image representation learning

Large amounts of labeled data are needed for training neural networks in a supervised way. Since labeled data are scarce and expensive to collect compared to unlabeled data, specific methods have been developed to leverage unlabeled data. A model can be trained in two stages.

First, the model is trained on a large, unlabeled dataset in a self-supervised manner. In *self-supervised learning* (SSL), pseudo-labels are constructed automatically from the unlabeled data, which reframes unsupervised learning as supervised learning and allows the use of standard learning techniques like gradient descent (Dosovitskiy et al., 2016; Zhang et al., 2017). The goal of this first stage is to obtain a neural network that produces informative, general-purpose representations for input data (Bengio et al., 2013).

In a second stage, models are finetuned for specific downstream tasks, for which (often little) labeled data are available, in a supervised way. By leveraging large amounts of unlabeled data, pre-trained models often outperform their counterparts that have only been trained on the smaller labeled dataset in a supervised manner (Schmarje et al., 2021). Such techniques are used in many machine learning (ML) application domains, like in natural language processing, image recognition, video-based tasks, or control tasks (Mikolov et al., 2013; Devlin et al., 2019; Florensa et al., 2019; Rouditchenko et al., 2019; Han et al., 2020; Liu et al., 2023; Qian et al., 2021).

Our work uses *contrastive learning* for learning image representations in a self-supervised way (Chopra et al., 2005; Le-Khac et al., 2020; Jaiswal et al., 2021). We train a model to project samples into a feature space where positive pairs are close to and negative pairs are far from each other. Contrastive pre-training has been used to learn image representations with great success recently, for instance, by van den Oord et al. (2018), Wu et al. (2018), Chen et al. (2020), Grill et al. (2020), He et al. (2020), Misra and van der Maaten (2020), and Patacchiola and Storkey (2020), who used different notions of distance and different training objectives.

Our study builds upon the relational reasoning framework for contrastive pre-training proposed by Patacchiola and Storkey (2020), but adapts it to satellite images by using spatial and temporal information to define similar and dissimilar image pairs, instead of images and data augmentations.<sup>1</sup> We chose this approach because of its state-of-the-art results and interpretability. We are not the first to use spatial and temporal information for contrastive pre-training: for instance, Qian et al. (2021) proposed a method for pre-training video representations, but their application domain of videos of daily human actions was quite different from our setting, their contrastive samples were video fragments, and they did not use spatial neighborhoods for defining positive or negative pairs.

<sup>1</sup> Data augmentations are generations of new samples from a base sample, with a transformation, such as a random crop or color distortion.

## 2.2. Learning representations of satellite imagery

The large amounts of publicly available remote-sensing data from programs such as LANDSAT (Williams et al., 2006) and SENTINEL (The European Space Agency, 2021) make this an interesting area of application for self-supervised pre-training techniques. Additionally, metadata like the spatial location or timestamps of images can be used to construct the distributions from which positive and negative pairs for contrastive learning are sampled.

Deep learning has, for instance, been used on satellite imagery for land cover and vegetation type classification (Kussul et al., 2017; Rustowicz et al., 2019; Vali et al., 2020), various types of scene classification (Cheng et al., 2017), object or infrastructure recognition (Li et al., 2017, 2020), and change detection (Kotkar and Jadhav, 2015; Chu et al., 2016; Gong et al., 2016; de Jong and Bosman, 2019).

Several studies have proposed the use of self-supervised pre-training on satellite images. Jean et al. (2019) proposed a triplet loss that pulls representations of spatially close tiles toward each other and pushes representations of distant tiles away from each other. Wang et al. (2020b) additionally used language embeddings from geotagged customer reviews. Kang et al. (2021) and Ayush et al. (2021a) also defined positive pairs based on geographical proximity and used momentum contrast (He et al., 2020) for a larger set of negative samples.

However, most recent works ignore the additional information that could be obtained from the temporal dimension of satellite images: satellites usually gather images of the same locations across multiple points in time. Ayush et al. (2021a) take images of the same location from distinct points in time as positive pairs, which causes their representations to be inevitably time-invariant. Manas et al. (2021) also proposed a contrastive pre-training method for satellite imagery using the temporal dimension. Both studies obtained representations that are maximally spatially variant, in the sense that only tiles of exactly the same location are considered similar. Neither method allowed to flexibly set different thresholds and consequently obtain different degrees of temporal and spatial variance.

In this work, we apply the state-of-the-art relational reasoning method of Patacchiola and Storkey (2020) to satellite images for the first time. This allows us to flexibly define and test several rules for positive/negative pair sampling, including rules that define images of the same location from distinct moments as dissimilar, and images from the same moments of nearby but not exactly the same locations as similar, which could result in relatively more space-invariant but time-variant representations, compared to Ayush et al. (2021a) and Manas et al. (2021).

## 2.3. Food insecurity prediction

Numerous studies have attempted to predict socioeconomic variables from satellite images. The predicted variables most often concern poverty, economic activity, welfare, or population density (Townsend and Bruce, 2010; Jean et al., 2016; Goldblatt et al., 2019; Hu et al., 2019; Bansal et al., 2020; Yeh et al., 2020; Ayush et al., 2021b; Burke et al., 2021). Some studies used additional data sources such as geotagged Wikipedia articles (Sheehan et al., 2019; Uz Kent et al., 2019). Researchers have also predicted crop yields from satellite images (Wang et al., 2018; Nevavuori et al., 2019), which is closer to food insecurity prediction, with the main difference being that food insecurity might also be caused by different factors such as political instability.

Other studies also predicted food insecurity using ML, but none used satellite imagery directly.<sup>2</sup> The World Bank has published two studies that predicted food insecurity from data in addition to defining a food insecurity score. Wang et al. (2020a) used a panel vector-autoregression (PVAR) model to model food insecurity distributions of 15 Sub-Saharan African countries on longer time horizons. Andree et al. (2020), on the other hand, used a random forest (Breiman, 2001) to model food insecurity on a shorter time horizon, with multiple handcrafted features as input: (1) structural factors such as spatial and population

<sup>2</sup> Andree et al. (2020) use the normalized difference vegetation index (NDVI) as an input feature, which is computed from satellite images as the normalized difference of images in two different spectral bands, but this simple scalar value cannot convey as much information as an entire image.

trends, ruggedness, and land use shares, (2) environmental factors such as the normalized difference vegetation index (NDVI), rainfall, and water balance equation, (3) violent conflict information, and (4) food price inflation. We mainly compare with the shorter-term predictions of Andree et al. (2020).

Lentz et al. (2019) predicted different food insecurity scores for Malawi from various input variables using linear and log-linear regression models.

### 3. Spatiotemporal SSL

Contrastive learning methods enable representation learning without annotated data. Instead, they rely on the intrinsic structure of data. For example, different patches from the same image are likely to be similar to each other and dissimilar to patches from other images. Training image representations to enable this discrimination should lead to useful image features.

Here we leverage the contrastive framework proposed by Patacchiola and Storkey (2020), who formulated this principle with an explicit relation prediction. They mapped each image  $I$  to augmentations  $\mathcal{A}(I)$  (e.g., patches) and jointly trained an image encoder  $\phi$  and a relation prediction network  $\rho$  to predict whether augmentations come from the same image:

$$\hat{y} = \rho(\phi(\mathcal{A}(I_i)), \phi(\mathcal{A}(I_j))), \tag{1}$$

where  $\hat{y}$  should be close to 1 when  $i = j$  and close to 0 otherwise. We use the same loss as Patacchiola and Storkey (2020):

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N} \sum_{i=1}^N w_i \text{CrossEntropy}(y_i, \hat{y}_i), \tag{2}$$

where  $w_i$  is the focal factor that modulates the loss according to the prediction confidence through a hyperparameter  $\gamma$ :

$$w_i = \frac{1}{2} [(1 - y_i)\hat{y}_i + y_i(1 - \hat{y}_i)]^\gamma. \tag{3}$$

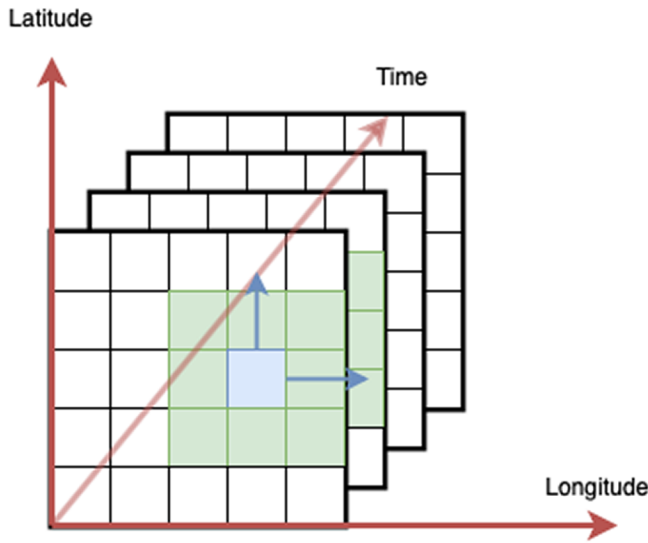
If  $\gamma > 1$ , uncertain predictions have a greater effect on the training loss.

Patacchiola and Storkey (2020) only rely on standard spatial image augmentations (horizontal flip, random crop-resize, conversion to grayscale, and color jitter). Where for natural images it makes sense to assume different images will be semantically different, since they are likely to depict different objects or scenes, satellite image tiles could be seen as a patchwork that forms a single large image, evolving over time, from the same object, that is, the Earth. The division of satellite imagery into smaller image tiles follows arbitrary boundaries determined by, for example, latitude/longitude coordinates, and not actual semantic boundaries, hence the resulting neighboring tiles are not necessarily likely to vary semantically. However, as spatial distance between satellite image tiles or time between when satellite images are taken increases, so does the likelihood that what is depicted changes semantically. Therefore, we define new augmentations based on temporal and spatial distances, and we consider far-away patches as if they came from a different image. In the next section, we evaluate this idea and compare different similarity criteria. We call our pre-training method spatiotemporal SSL (SSSL).

For this purpose, we introduce different thresholds  $D_g$  and  $D_t$  of respectively geographic distance and temporal distance in order to define positive pairs.  $D_g$  is measured in degrees of longitude/latitude, and  $D_t$  in months. Let  $x_i$  be a sampled anchor observation characterized by time  $t_i$ , latitude  $lat_i$ , and longitude  $lon_i$ . Positive (similar) pairs include images  $x_j$  for which the following constraints apply:

$$t_i - D_t < t_j < t_i + D_t, \tag{4a}$$

$$lat_i - D_g < lat_j < lat_i + D_g, \tag{4b}$$



**Figure 1.** SSSL: for one sample, positive samples are those that are closer in time to the image than the temporal threshold, and are closer in space to the sample than the spatial threshold.

$$\text{lon}_i - D_g < \text{lon}_j < \text{lon}_i + D_g. \tag{4c}$$

Figure 1 illustrates this constraint. When  $D_t$  is arbitrarily high, and  $D_g \approx 0$ , the positive pairs are similar to the positive pairs defined by Ayush et al. (2021a). This would result in spatially variant but time-invariant representations and reduce the effect of seasonality or other temporal trends on the representation. When  $D_t < f$ , on the other hand, with  $f$  the temporal frequency of the imagery, positive pairs are purely location-based. This is similar to the strategy of Tile2Vec (Jean et al., 2019). In addition to fixed thresholds of  $D_g$  as in Eqs. (4b)–(4c), we also define spatial positive pairs as images that correspond to the same predefined area (administrative unit, AU).

#### 4. IPC score prediction

We approach the downstream task of predicting food insecurity as a classification problem of satellite tiles  $I$  (or a collection thereof) into one out of five possible IPC scores  $s$  corresponding to different levels of food insecurity. We chose to approach IPC score prediction as a classification problem following Andree et al. (2020).

An image encoder  $\phi$  (cf. Eq. (1)), possibly pre-trained as described in the previous section, projects a tile onto a tile embedding  $\phi(I)$ . A multilayer perceptron (MLP) with 1 hidden layer then projects the tile embedding to a probability distribution over the IPC scores:  $\hat{s}^{\text{tile}} \sim \text{MLP}(\phi(I))$ .

We train the classification MLP and potentially the image encoder  $\phi$  with a cross-entropy loss to assign the highest probability to the IPC score of the AU to which the location of a tile belongs to, on the date of the satellite image. Unless otherwise stated, we *predict* the IPC score gathered by FEWS NET *for the same date* as the satellite image was taken. In one experiment, we will *forecast* future IPC scores, gathered for dates *up to 12 months after* the date of the satellite image that was used as input.

##### 4.1. Score aggregation

Since IPC scores are defined in AUs, and since one AU contains many different locations and hence satellite image tiles, we need a way to aggregate our network’s predictions per tile into one prediction per



AU. If  $M$  tiles  $\{I_i\}_{i=1,\dots,M} \in \text{AU}_k$ , where  $\text{AU}_k$  is an AU, we need a single predicted IPC score  $\hat{s}_k^{\text{AU}}$  for the whole unit, based on the predicted IPC score  $\hat{s}_i^{\text{tile}}$  per tile  $I_i$ :

$$\hat{s}_k^{\text{AU}} = \text{Agg}\left(\{\hat{s}_i^{\text{tile}}\}_{i=1,\dots,M}\right), \quad (7)$$

where  $\hat{s}_i^{\text{tile}} = \text{argmax}(\text{MLP}(\phi(I_i)))$ ,

and where  $\text{Agg}: \{\hat{s}_1^{\text{tile}}, \dots, \hat{s}_M^{\text{tile}}\} \mapsto \hat{s}^{\text{AU}}$  is an aggregation function. We consider three aggregation methods:

1. Majority voting: the predicted score for the AU is the score that has been predicted most often for tiles within that AU.
2. Maximum voting: the predicted score for the AU is the maximum of the predicted tile scores.
3. Individual tiles: predicting and evaluating the IPC scores on a per-tile basis, which is arguably harder since a tile's IPC score may be determined by another location in the AU.

## 5. Experiments

### 5.1. Data

#### 5.1.1. Pre-training

We make use of publicly available imagery from the LANDSAT-8 satellite<sup>3</sup> (Roy et al., 2014). LANDSAT is the longest-running satellite photography program and is a collaboration between the US Geological Service (USGS) and the National Aeronautics and Space Administration (NASA). The satellite captures landscapes from all over the world with a spatial resolution of 30 m per pixel and a temporal resolution of 16 days. To reduce the impact of clouds on the satellite images, we use Google Earth Engine<sup>4</sup> (GEE; Gorelick et al., 2017) to generate composite images comprised of individual images spanning 3–4 months, matching the temporal frequency of the downstream food insecurity samples. We use all seven available surface reflectance spectral bands: one ultra-blue, three visible (RGB), one near-infrared, and two short-wave infrared. We use images of the entire surface area of Somalia (640K km<sup>2</sup>), which were captured between May 2013 (earliest LANDSAT-8 data availability in GEE) and March 2020 (latest available IPC score), resulting in 10 three-month and 13 four-month composites. We divide the images into tiles of 145×145 pixels so they can be processed by a CNN. Figure 2 shows the visible RGB bands of three such tiles. One tile corresponds to almost 19 km<sup>2</sup>. We end up with 800K tiles, consisting of 35K locations across 23 moments in time.

#### 5.1.2. Food insecurity prediction

We use the data on food insecurity in 21 developing countries made available by Andree et al. (2020) to finetune our results. FEWS NET<sup>5</sup>, an information provider that monitors and publishes data on food insecurity events, defines the target variable: the IPC score. The IPC score has five possible values: (1) minimal, (2) stressed, (3) crisis, (4) emergency, and (5) famine. The scores are measured using the IPC system, which is an analytical framework to qualitatively assess the severity of a food crisis and consecutively recommend policies to mitigate and avoid crises (Hillbruner and Moloney, 2012). IPC scores are given per AU, of which the boundaries are set by the UN Food and Agriculture Organization. Figure 3 shows the IPC score distribution per country and for Somalia per year. The classes are heavily imbalanced: the relative frequencies of IPC scores 1, 2, 3, and 4 are 14%, 71%, 13%, and 1.6%, respectively.

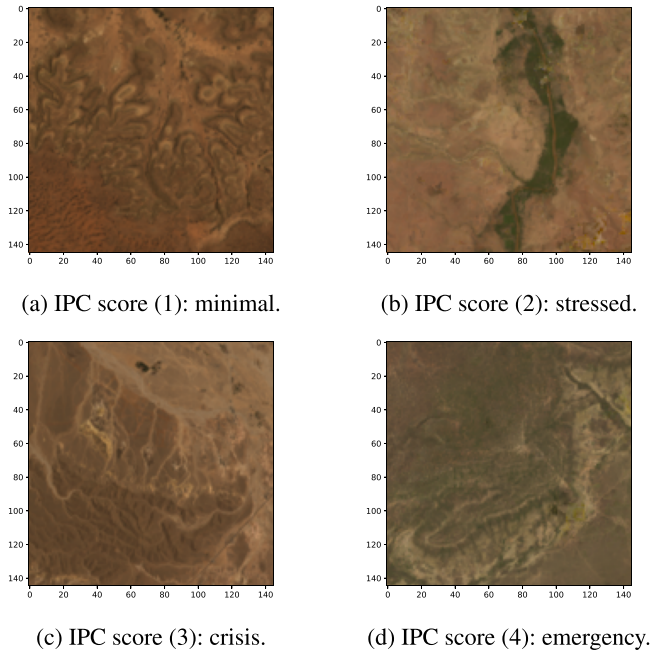
To limit resource usage, we chose to focus on IPC score prediction for Somalia, since four out of five possible IPC scores occur in Somalia between 2013 and 2020, and because food insecurity in Somalia is

<sup>3</sup> <https://www.usgs.gov/core-science-systems/nli/landsat/Landsat-8>.

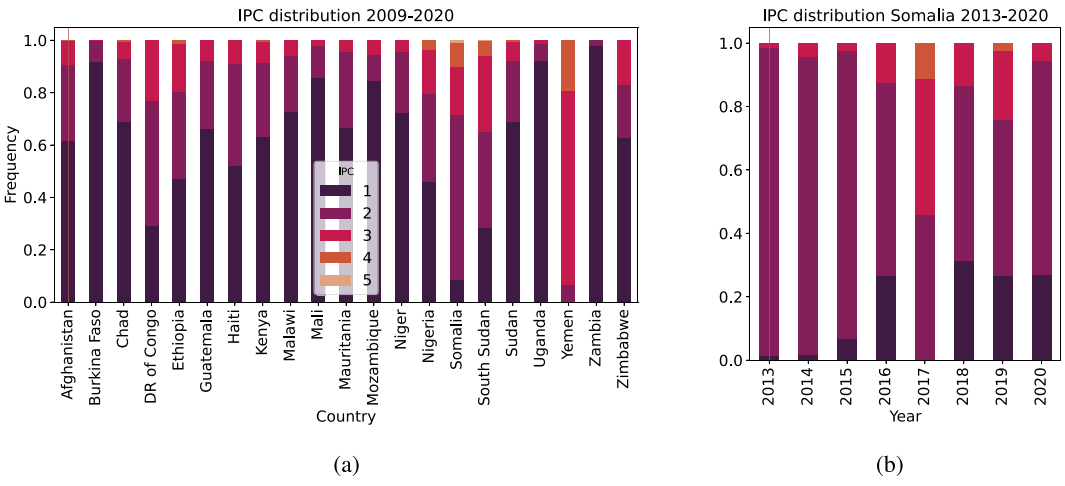
<sup>4</sup> <https://earthengine.google.com/>.

<sup>5</sup> <https://fews.net/IPC>.



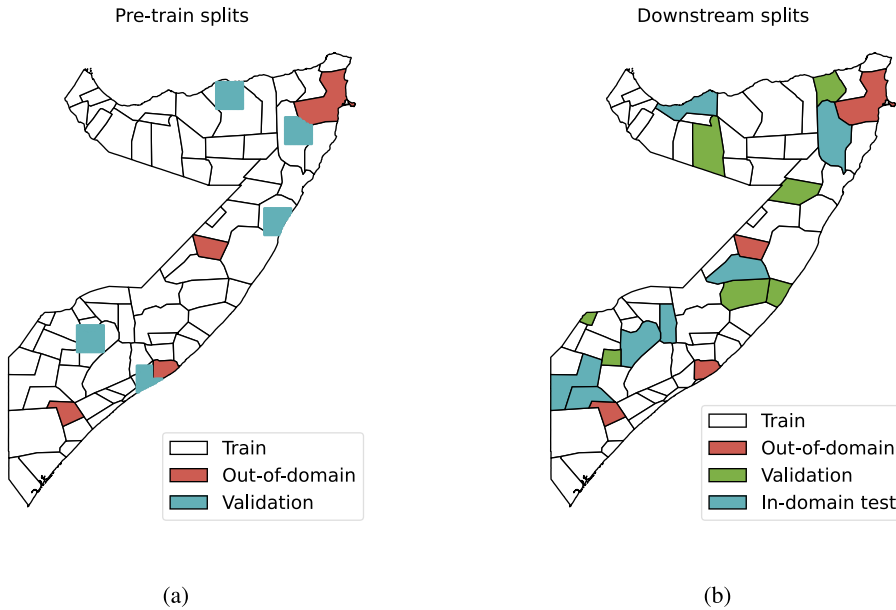


**Figure 2.** Examples of  $145 \times 145$  pixel tiles taken from composite LANDSAT-8 images of Somalia, exported from GEE (only RGB bands visualized), with corresponding IPC scores. Note that the difference between images with different IPC scores is not easily discernible.



**Figure 3.** IPC score distribution (a) for each country in the dataset from 2009 to 2020 and (b) for Somalia per year from 2013 until 2020. Note that IPC score 5 does only occur in 2011 in Somalia.

mainly caused by agricultural and rainfall factors (Andree et al., 2020). We also experimented with predicting IPC scores for South Sudan, but results were far worse. This can be explained by the fact that food insecurity in South Sudan in 2013–20 was caused by non-environmental factors such as markets and conflicts (Andree et al., 2020). The timeframe of the IPC score extends from August 2009 to February 2020. The score is reported at quarterly frequency from 2009 to 2016, and three times per year from 2016 to 2020. We limit our timeframe to August 2013 until March 2020 as for the pre-training images. The 3- or 4-month satellite image composite start and end dates (e.g., May 2013 to August 2013) are chosen so that



**Figure 4.** (a) Geography of pre-train data splits: train data are used for SSSL pre-training, validation data are used to select the best checkpoint after pre-training, and out-of-domain data are set aside. (b) Geography of downstream IPC score prediction data splits: train data are used for IPC score classification, validation data are used for early stopping and selecting the best checkpoint, out-of-domain and in-domain test data are used for evaluation.

the end date always corresponds to a date for which an IPC score is available. A composite (and the  $145 \times 145$  tiles it is split into) is matched to the IPC score gathered by FEWS NET in the month of the composite end date: for instance, tiles generated out of a composite of satellite images taken between May 2013 and August 2013 are matched to the IPC score gathered in August 2013.

### 5.1.3. Data splits

We take 4 of the 74 AUs as out of domain: all 44K tiles belonging to 1.9K locations within these AUs, together with these regions' 92 IPC scores (one per region per date), form the out-of-domain test set  $\mathcal{D}_{\text{ood}}^{\text{ipc}}$ . These tiles are not included in pre-training data.

For *pre-training*, we divide all locations in the 70 remaining AUs over two data splits: the training set  $\mathcal{D}_{\text{train}}^{\text{pre}}$  (31K locations, 712K tiles) and the validation set  $\mathcal{D}_{\text{val}}^{\text{pre}}$  (1.9K locations, 43K tiles, or little more than 5%). All tiles (timestamps) belonging to one location are always in the same split, but the train-val split does not necessarily respect AU boundaries. The validation split consists of a number of contiguous square areas randomly spread over Somalia in order to make sure that every location has a sufficiently large spatial neighborhood to sample positives from (for contrastive learning). Figure 4a shows the spatial division of the pre-training data.

For the *downstream task* (IPC score prediction) of training and evaluation, we take 7 out of the 70 AUs (74 minus 4 for the out-of-domain split) for the validation set  $\mathcal{D}_{\text{val}}^{\text{ipc}}$  (3.1K locations, 72K tiles, 161 IPC scores) and another 7 for the in-domain test set  $\mathcal{D}_{\text{test}}^{\text{pre}}$  (4.6K locations, 105K paths, 161 IPC scores). The remaining 56 AUs make up the training set  $\mathcal{D}_{\text{train}}^{\text{ipc}}$  (25K locations, 578K tiles, 1.3K IPC scores). Figure 4b shows the geography of the downstream task splits. To test performance when the amount of available labeled data for supervised downstream task training decreases, we also construct training sets with a decreasing number of AUs: 70%, 50%, 20%, 5%, and 1% of the full training set  $\mathcal{D}_{\text{train}}^{\text{ipc}}$ .

**Table 1.** Comparison of (pre-training) dataset sizes in related work

Study	Dataset name	Number of images	Pixels per image	Total number of pixels
This study	Total	799K	$145 \times 145$	17B
	$\mathcal{D}_{\text{train}}^{\text{pre}}$	712K	$145 \times 145$	15B
Relational reasoning (Patacchiola and Storkey, 2020)	CIFAR-10 and CIFAR 100	60K	$32 \times 32$	61M
	Tiny-ImageNet	100K	$64 \times 64$	409M
	NAIP			12B
Tile2Vec (Jean et al., 2019)	LANDSAT-8: American cities			60M
	LANDSAT-7: Uganda	16K	$145 \times 145$	344M
	DigitalGlobe			2.9B
Geography-aware SSL (Ayush et al., 2021a)	Functional map of the world	417K	$224 \times 224$	21B

Table 1 compares the total number of pixels of our data splits with those used by other self-supervised pre-training for (satellite) image studies and shows that we match the order of magnitude of the most large-scale study of Ayush et al. (2021a).

## 5.2. Experimental setup and methodology

### 5.2.1. SSSL pre-training

To define positive and negative pairs of patches for SSSL pre-training, we explore spatial resolutions  $D_g$  of  $0.15^\circ$ ,  $0.4^\circ$ , and entire AUs, and temporal resolutions  $D_t$  of 1 (meaning only tiles from the same date are considered similar), 4, 12, 36, or 84 months (the length of our entire timeframe, which means spatially nearby tiles are considered similar regardless of their date). When  $D_t$  equals 84 months and  $D_g$  is small enough, our positive and negative pairs are similar to those used by Ayush et al. (2021a) (their spatial threshold is actually so small that only the exact same location is considered similar, while our smallest spatial threshold of  $0.15^\circ$  still considers for nearby but not identical locations to be similar). Jean et al. (2019) used positive pairs determined by spatial locality (with a small spatial threshold), which resemble our pairs when  $D_t$  equals 1 month and  $D_g$  is small but large enough to include more than a single location.

*Baselines.* We compare SSSL with the following pre-training baselines. The best pre-training checkpoints are chosen based on IPC score prediction performance from the frozen checkpoint weights, but we perform further evaluations involving both frozen and finetuned pre-trained weights.

1. The relational reasoning method of Patacchiola and Storkey (2020), which uses image augmentations of the anchor images like random flips, random crops, etc., to define positive instead of spatial and temporal thresholds.
2. The Tile2Vec contrastive pre-training method for satellite imagery, which uses a triplet loss to pull an anchor tile's representation closer to a nearby positive tile's representation in feature space while pushing it away from a far-away negative tile (Jean et al., 2019). We adjust the algorithm to work with a configurable spatial threshold instead of a fixed one. We add a configurable temporal threshold, so we can directly compare this baseline to our SSSL pre-training for different spatial and temporal thresholds  $D_g$  and  $D_t$ .

We chose Tile2Vec as baseline that uses contrastive pre-training specifically designed for satellite imagery since it is more easily extendable to configurable spatial and temporal thresholds and to limit the resources required for our study. The methods proposed by Ayush et al. (2021a) and Manas et al. (2021) explicitly

rely on fixed thresholds that only consider tiles of the exact same location across time to be similar to arrive at temporally invariant and spatially variant representations, so they cannot be as naturally extended to our flexible threshold setting.

*Hyperparameters and settings.* We use a total number of positive (and negative) pairs  $K = 8$  for SSSL, and batch size  $N = 50$  for both SSSL and Tile2Vec pre-training. Minibatches are constructed by sampling  $N$  anchor samples from the dataset, and adding  $K - 1 = 7$  for SSSL and 1 for Tile2Vec positives per anchor to the batch. For each anchor, random other anchors or other anchors' positives from the same minibatch are used as negatives.

We pre-train all CNN backbones on  $\mathcal{D}_{\text{train}}^{\text{pre}}$  for a fixed number of epochs and save all intermediate checkpoints for later evaluation (one per epoch). We stopped SSSL pre-training after 10 epochs, and Tile2Vec pre-training after 20 (40 would have been in some sense more fair since Tile2Vec only sees one positive and one negative per sampled anchor tile, while SSSL sees  $K - 1 = 7$  positives and negatives per anchor tile, so SSSL batches are  $4 \times$  larger than Tile2Vec batches, but we noticed that downstream task performance steadily decreased after 10 epochs, while the needed training time for 20 epochs of Tile2Vec training was already significantly more than 10 epochs of SSSL pre-training, due to increased overhead).

We use the ResNet-18 architecture as CNN backbone for all experiments to balance performance with resource usage (He et al., 2016). We also tested the Conv4 network that Patacchiola and Storkey (2020) used, but results were much worse. We use the Adam optimizer (Kingma and Ba, 2015) with learning rate  $1e-4$  and  $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 1e-6)$ . We set  $\gamma = 2.0$  in the focal factor (Eq. (3)) and use a weight decay factor of  $1e-4$  for SSSL. For Tile2Vec, we set the margin for the triplet loss to 1.0 and the L2 regularization weight to 0.01.

We used a 16 GB NVIDIA Tesla P100 GPU for all pre-training and downstream task runs. SSSL pre-training on  $\mathcal{D}_{\text{train}}^{\text{pre}}$  took on average 36 h. Tile2Vec pre-training took on average 51 h. Neural network training and evaluation was implemented with PyTorch (Paszke et al., 2019).<sup>6</sup>

### 5.2.2. Downstream task: IPC score prediction

After pre-training, we train a single-layer MLP on  $\mathcal{D}_{\text{train}}^{\text{ipc}}$  to predict IPC scores from the frozen image features of the CNN backbone of each pre-training checkpoint until macro F1 on the validation set  $\mathcal{D}_{\text{val}}^{\text{ipc}}$  converges. We use these validation scores to choose the best pre-training checkpoint for every pre-training run and to choose the best performing spatial and temporal thresholds and the best performing score aggregation method. We then use the best checkpoints (best validation performance) of the pre-training runs with the best spatial and temporal thresholds for further evaluations.

We report macro F1 scores since higher IPC scores (indicating a higher degree of food insecurity) occur much less frequently than lower scores, but are at least as important (if not more important) to detect. The macro F1 weighs all IPC scores equally, as opposed to micro averaged metrics that would give more weight to more frequent IPC scores.

*Baselines.* In addition to the pre-training baselines, for which we used the checkpoints to initialize an IPC score predictor as described in the beginning of this section, we consider the following IPC prediction baselines that don't need manual pre-training.

1. A randomly initialized CNN backbone, without any pre-training.
2. A CNN pre-trained on ImageNet classification (Deng et al., 2009; He et al., 2016). Since ImageNet images consist of three RGB channels instead of seven like our LANDSAT-8 images, we copy the convolution weights of the RGB channels from the pre-trained checkpoint but add randomly initialized weights for four additional channels.
3. A random forest, like the one proposed by Andree et al. (2020).

<sup>6</sup> Upon acceptance, we will publish all training, evaluation and data preprocessing code, as well as the scripts used to export satellite images from GEE, and trained checkpoints of our models.

*Random forest.* We compare our neural network’s food crisis predictions to those of a random forest classifier, as used by Andree et al. (2020). Andree et al. (2020) merged the five IPC score categories into two—food crisis or not—and trained a binary classifier. They used the following input variables for 20 developing countries from 2009 until 2020:<sup>7</sup>

- the coordinates of the central points;
- the district size;
- the population;
- the terrain ruggedness;
- the cropland and pastures area shares;
- the NDVI—a measure of the “greenness,” relative density, and health of vegetation of the earth’s surface;
- the rainfall;
- the evapo-transpiration (ET);
- conflict events;
- food prices.

For a fair comparison, we only use the data for Somalia and the 2013–20 timeframe. We perform food insecurity prediction under two setups: binary classification, following Andree et al. (2020), and multi-class classification with the five possible IPC scores, as described thus far. Our random forests consist of 50 decision trees. A leaf node needs to contain at least three samples to be considered during training, and it needs to contain at least 10 samples to be split into new leaf nodes. Out of the 11 features a sample has, three are considered per split point. Trees have a maximum depth of six nodes. The class weights for random forest training are inversely proportional to their frequency.

After comparing the learned image encoder features with handcrafted features, we also combine both by adding the neural network predictions as additional input features to assess their complementarity.

*Hyperparameters and settings.* Again, we use the Adam optimizer, now with the weight decay factor set to 0.01. If the pre-trained CNN backbone is not frozen during downstream task training, but finetuned, its weights are updated with a lower learning rate of  $1e-5$  than the classification MLP (which is updated with learning rate  $1e-4$ ). We use early stopping on the validation macro F1 score of predictions that use majority voting as aggregation method and reduce the learning rate when the validation macro F1 reaches a plateau. To counteract class imbalance, we weigh the IPC classes in the cross-entropy loss inversely proportional to their frequency in the training data.

Training for the downstream task  $\mathcal{D}_{\text{train}}^{\text{ipc}}$  took approximately 4 h when freezing the pre-trained CNN backbone, and 8 h when finetuning it. We used Scikit-learn to implement the random forest (Pedregosa et al., 2011).

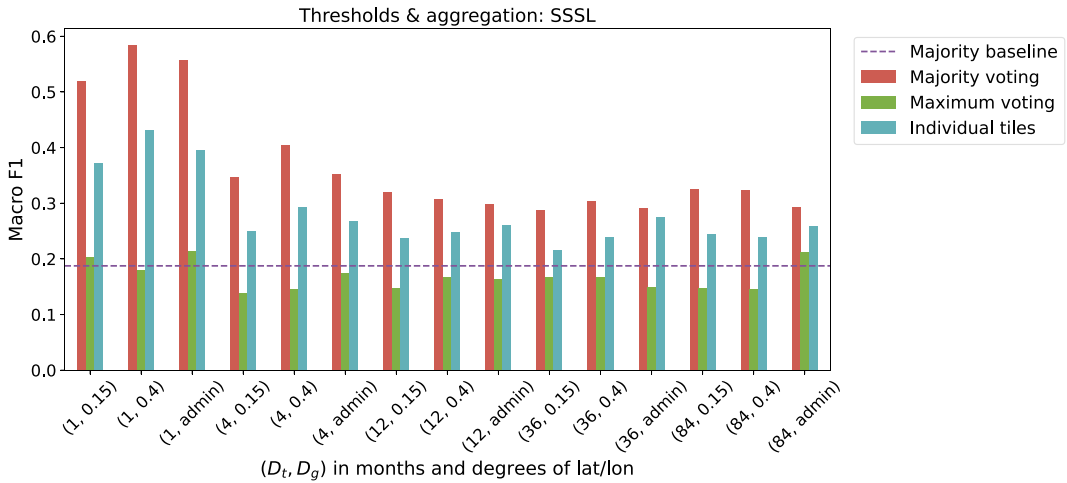
## 6. Results

### 6.1. Spatial and temporal thresholds

Figure 5 shows the validation macro F1 on the downstream task for different combinations of spatial and temporal threshold values for positive pair selection, as well as for different aggregation methods, after pre-training with SSSL on images in  $\mathcal{D}_{\text{train}}^{\text{pre}}$ .

It is clear that the best performing configurations use a small temporal threshold, with by far the best performance when using  $D_t = 1$  month (so only spatially nearby tiles of the same 3- or 4-month composite are considered similar). This makes the representations time-variant by minimizing mutual information between image representations of the same location at different times. Since our downstream task is time-

<sup>7</sup> We do not use time data like the month and the year of an IPC measurement as input for the random forest, since the test and validation sets are spatially but not temporally separated, and since the neural networks also do not have access to this information.



**Figure 5.** Macro F1 on validation set  $D_{val}^{ipc}$  using different configurations of positive and negative pairs (determined by temporal threshold  $D_t$  and spatial threshold  $D_g$ ) for SSSL pre-training, with  $D_t$  and  $D_g$  denoted on the x-axis. The baseline in this plot always predicts the majority class. “admin” means using administrative units instead of longitude/latitude to define spatial positive pairs.

dependent as well (regions might be food-insecure during certain time periods but not during others), this is not surprising.

Using a fixed spatial threshold  $D_g$  of either  $0.15^\circ$  or  $0.4^\circ$  usually gave better results than defining spatial positive pairs based on AUs. This means it is desirable to maximize mutual information between image representations of locations that share a medium-sized vicinity, but not when this vicinity’s size increases or decreases too much. This is somewhat surprising because the granularity of one AU corresponds exactly to the granularity of the IPC scores, and one might thus expect maximizing information between image representations of locations that share an IPC score to work best. But while patches in one AU share one IPC score, AUs can be quite large ( $>10K \text{ km}^2$ ), and patches might thus be quite different. If the patches are too different, or if they do not share the properties informative to IPC score prediction, the network might thus be forced to ignore important properties.

## 6.2. Score aggregation

“Individual tiles” in Figure 5 means predicting an entire AU’s IPC score from a single patch, which is inherently difficult since the IPC score might be determined by much more information than a single patch contains.

Majority voting almost always performed best. Maximum voting performed much worse, which could be caused by FEWS NET not giving an AU the worst IPC score of its subregions, and by the fact that a single incorrect patch prediction is more likely to change the entire AU prediction with maximum voting than with majority voting.

The rest of the experiments use SSSL with one configuration of positive pairs: a temporal threshold  $D_t$  of 1 month and a spatial threshold  $D_g$  of  $0.4^\circ$ , with majority voting as aggregation method. Conclusions for different spatial and temporal thresholds for Tile2Vec pre-training are largely similar, with the best performing setting  $D_t = 1$  and  $D_g = 0.15^\circ$  (see Figure A1 in Appendix A).

## 6.3. SSSL vs. baselines

Table 2 compares the macro F1 on the in-domain and out-of-domain test sets of the best pre-training settings for SSSL and Tile2Vec to the original pre-training proposed by Patacchiola and Storkey (2020)



**Table 2.** Macro F1 on the in-domain and out-of-domain test set of the SSSL model with spatial and temporal positive pairs vs. baselines: Tile2Vec (also with spatial and temporal pairs), the data augmentation-based model of Patacchiola and Storkey (2020), ImageNet pre-training, random initialization, and the random forest (RF) of Andree et al. (2020). The best result per column is marked in bold.

Pairs	$\mathcal{D}_{\text{test}}^{\text{ipc}}$		$\mathcal{D}_{\text{ood}}^{\text{ipc}}$	
	Frozen	Unfrozen	Frozen	Unfrozen
Maj. baseline	0.201		0.205	
RF	0.398		0.390	
Random init.	0.278	0.422	0.356	0.369
ImageNet	0.444	0.483	0.347	0.456
Tile2Vec	0.356	0.484	0.393	0.407
Data aug.	0.392	0.478	0.358	0.387
SSSL	<b>0.543</b>	<b>0.654</b>	<b>0.477</b>	<b>0.542</b>

Note: Results of CNN backbones with both frozen backbone weights and unfrozen backbone weights during supervised training are reported. Maj. baseline corresponds to always predicting the majority class.

that uses data augmentations (color jitter, resized crop, etc.) instead of our spatiotemporal model. It also shows the results of a model that does not integrate pre-training (i.e., starting from randomly initialized weights and training these only during the supervised training of food insecurity prediction), of pre-training on ImageNet (Deng et al., 2009; He et al., 2016) (i.e., starting the supervised training with the convolutional weights initialized to publicly available weights that were trained on the ImageNet classification dataset), and of a random forest that uses handcrafted features (Andree et al., 2020). We consider both freezing and not freezing the CNN backbone's weights in the supervised stage.

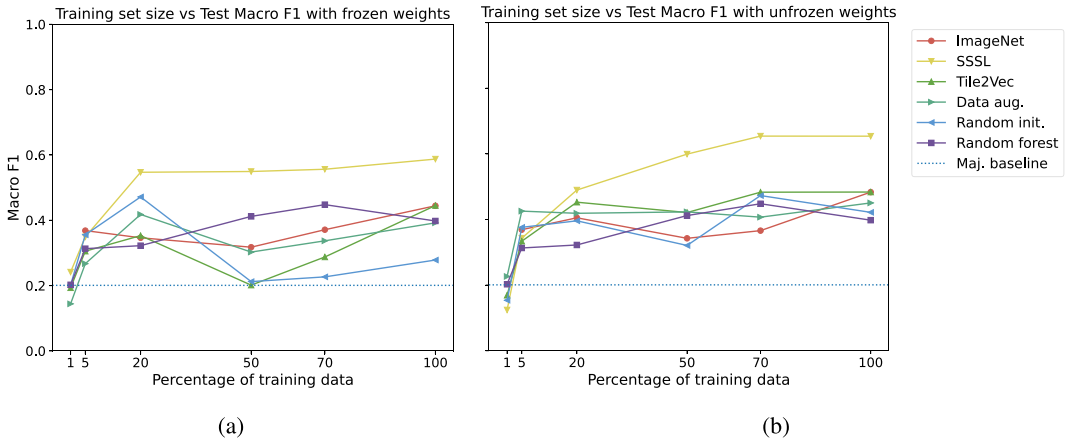
SSSL significantly outperforms all baselines in all settings, with 21–39% relative improvement over the second best model. All neural network-based models (bottom five rows) scored better than the randomly initialized neural network baseline across all settings, although in some cases only marginally, especially on the out-of-domain test set. Tile2Vec and data augmentations showed comparable performance, and only outperformed the random forest baseline when their CNN weights were finetuned. Surprisingly, ImageNet outperformed Tile2Vec and data augmentations with frozen backbone weights on the in-domain test set  $\mathcal{D}_{\text{test}}^{\text{ipc}}$ , even though the CNN backbone had only seen images of daily scenes like cats and dogs during ImageNet pre-training, while the latter two baselines were pre-trained on satellite images in the  $\mathcal{D}_{\text{train}}^{\text{pre}}$  dataset. Finetuning the CNN weights improved performance compared to freezing them, often significantly.

#### 6.4. Transferability

Performance generally drops on the out-of-domain test set, but stays well above the random and majority baselines. This shows that it is harder but still feasible to make good IPC score predictions for locations for which no imagery was included in the pre-training data. Note that none of the images and IPC scores in  $\mathcal{D}_{\text{test}}^{\text{ipc}}$  or  $\mathcal{D}_{\text{ood}}^{\text{ipc}}$  were included in the downstream task training set  $\mathcal{D}_{\text{train}}^{\text{ipc}}$ , but images in  $\mathcal{D}_{\text{test}}^{\text{ipc}}$  might have been included in the pre-training data  $\mathcal{D}_{\text{train}}^{\text{pre}}$ , while images in  $\mathcal{D}_{\text{ood}}^{\text{ipc}}$  were definitely not. Also note that this only makes a difference for SSSL, Tile2Vec, and data augmentations, since the other models were not pre-trained on  $\mathcal{D}_{\text{train}}^{\text{pre}}$  anyway.

Therefore, our model could be used not only to predict food insecurity for locations for which no labeled data are available, but also for locations on which it has not been pre-trained (although it is preferable to pre-train on all locations for which IPC predictions need to be made). The out-of-domain





**Figure 6.** Test macro F1 on  $\mathcal{D}_{test}^{ipc}$  with frozen (a) and unfrozen (b) CNN backbone weights for models with different weight initializations using increasing amounts of labeled training data.

locations in  $\mathcal{D}_{ood}^{ipc}$  are in separate AUs, but of course still in the same country as and adjacent to AUs in pre-training and downstream training data. Some degree of similarity can thus still be expected. It would be interesting to test how performance degrades when distance or dissimilarity between out-of-domain test data and training data increases, for example, on locations in different countries or climates.

### 6.5. Decreasing labeled dataset size

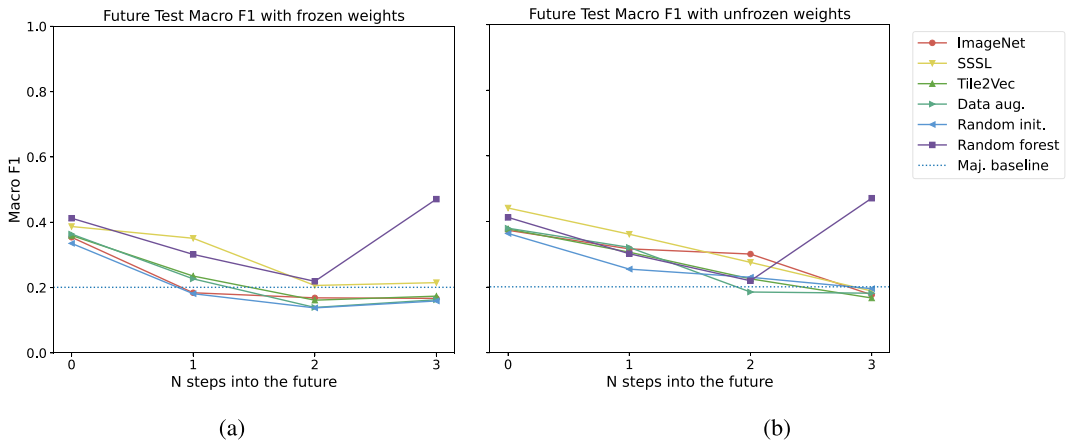
Figure 6 shows the macro F1 on the in-domain test set  $\mathcal{D}_{test}^{ipc}$  for models with different weight initializations for different amounts of labeled data used to train for the downstream task. As expected, macro F1 decreases with decreasing training set sizes, but it does so gradually, not disproportionately. This is the case both for SSSL and most baselines, except for Tile2Vec when freezing its CNN’s weights, for which performance drops rapidly to the majority baseline. Performance starts falling sharply when decreasing the training set size further than 20% of its original size, but up until a decrease to 5% of available data, all models perform better than the majority baseline. SSSL pre-training outperforms the baselines for training set sizes above 5%, both with finetuned and frozen weights.

The random forest performed better than neural baselines (but not SSSL) with frozen weights and performed equally well or better than neural baselines (but not SSSL) with unfrozen weights, for 50–70% of training data, meaning that it is more robust to slight decreases in training set size. Surprisingly, its performance when trained on all data drops compared to when trained on 50–70%, which might be explained by some samples being excluded from training data that are “harder” or more dissimilar to test data.

Although overall performance with unfrozen CNN weights is better than with frozen weights, the latter is more robust to decreasing training set size. This could be explained by the fact that not updating the representations reduces the number of trainable parameters vastly, and hence the risk to overfit a small labeled training set. We noticed some training instability on the smaller training sets when freezing the CNN weights (shown, e.g., by the unexpected bump in Random init. performance for 20% of the training data).

Figure B1 in Appendix B shows the same plots but for the out-of-domain test set  $\mathcal{D}_{ood}^{ipc}$  instead of the in-domain test set (of which the satellite image tiles were not included in pre-training data). The gap between SSSL and baseline is smaller with frozen weights. For unfrozen weights, SSSL performance drops below baselines when trained on 50% or less of labeled data.

We can conclude from these experiments that (1) contrastive SSSL pre-training and (2) defining rules for positive/negative pair selection that are tailored to satellite images, by making use of their spatial and temporal dimensions instead of data augmentations, will improve results for varying amounts of labeled



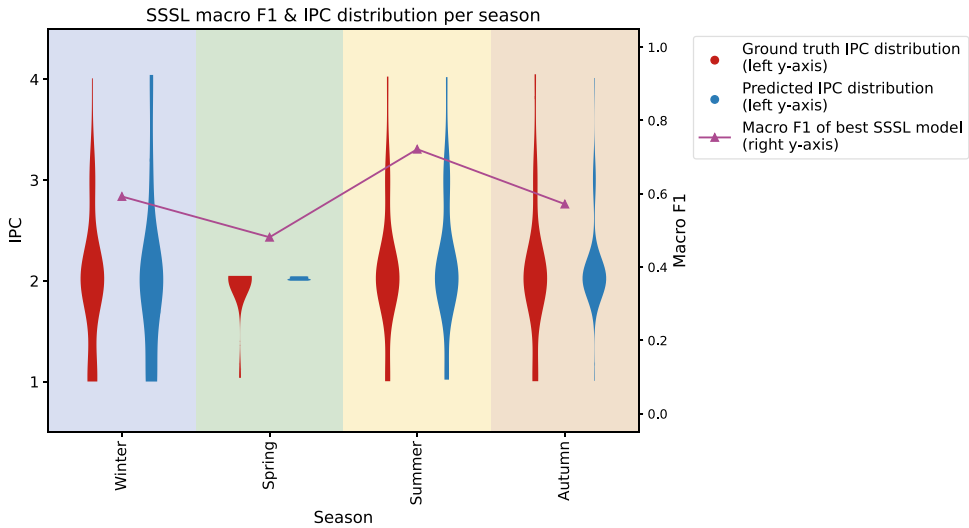
**Figure 7.** Test macro F1 on  $\mathcal{D}_{\text{test}}^{\text{ipc-temp}}$  with frozen (a) and unfrozen (b) CNN backbone weights for neural networks with different weight initializations and a random forest when predicting an increasing number of time steps into the future (one step corresponds to 3–4 months, two to 6–8 months, and three to 9–12 months).

training data. Little labeled data are needed for finetuning to a downstream task. Performance with decreasing training set size is better retained when the model has seen the locations during its pre-training stage.

### 6.6. Forecasting food insecurity in the future

Figure 7 shows the macro F1 on a different, temporally separated test set  $\mathcal{D}_{\text{test}}^{\text{ipc-temp}}$  for different models (with frozen (7a) and unfrozen (7b) CNN weights), when forecasting food insecurity in the future. Here the future means a later relative point in time than the date the input satellite image was acquired. To allow time for preventive political measures or timely humanitarian action, a system that warns about food insecurity more than 3–4 months before it actually occurs would be useful. Hence we train and evaluate models for predicting the next gathered IPC score for every location ( $N = 1$ , which corresponds to 3–4 months later), the one after that ( $N = 2$ , 6–8 months later), and the one after that ( $N = 3$ , 9–12 months later). While pre-training remains the same, we no longer use the geographically separated  $\mathcal{D}_{\text{train}}^{\text{ipc}}$ ,  $\mathcal{D}_{\text{val}}^{\text{ipc}}$ ,  $\mathcal{D}_{\text{test}}^{\text{ipc}}$  for finetuning. Instead we separate all the in-domain data *temporally* (exclusively for the experiments in this section): we use the IPC scores from March 2020 for validation (i.e., the last available IPC scores at the time of the start of this study, corresponding to the date of the last LANDSAT-8 tiles that were included in pre-training). We train on the IPC scores up to November 2019 (up until one step before the validation scores). The first IPC scores used for training are chosen so that all of the training sets for this experiment (i.e., corresponding to a different number of steps into the future) are of equal size. The test IPC scores are from June 2020 and have been published by FEWS NET since the start of this study. This means that no satellite imagery corresponding to the time of the test IPC scores has been used for pre-training. Note that for the experiments in this section, predicting 0 steps into the future ( $N = 0$ ) still uses the temporally instead of geographically separated splits, and is therefore not identical to previously discussed runs. The same temporal splits are used to obtain the train, validation, and test set for the random forest.

Figure 7 shows that forecasting into the future is difficult for any of the considered methods, and that generally performance decreases when forecasting further into the future. We consider the sudden increase in macro F1 when forecasting three steps into the future with the random forest an anomaly: since only one IPC score and corresponding covariates have been collected per AU per timestep, the data sets to train and evaluate the random forest are relatively small. With frozen weights, SSSL pre-training performs comparable to the random forest as used by Andree et al. (2020), better than the majority



**Figure 8.** Test macro F1 on  $\mathcal{D}_{\text{test}}^{\text{ipc}}$  of the SSSL model with unfrozen CNN weights (magenta line, right vertical axis), and ground truth (red) and predicted (blue) IPC score distributions (violin plots, left vertical axis), both versus the season of the IPC score measurement (x-axis). Note that only four IPC scores are depicted, since only four out of five possible IPC scores occur in Somalia between 2013 and 2020.

baseline (although barely for  $N \geq 2$ ) and better than the other methods (which drop below the majority baseline for more than one or two timesteps into the future). With unfrozen weights, SSSL performs best when forecasting  $N = 1$  steps into the future, and slightly worse than ImageNet for  $N = 2$ .

To verify to what extent models are able to predict future IPC scores that do not change over time, we compute the macro F1 on the subset of AUs whose IPC scores actually changed since the acquisition of the image. Performance dropped significantly: for  $N = 1$ , SSSL performance dropped from 0.351/0.361 to 0.217/0.262 with frozen/unfrozen weights, but stayed above the baselines' performance (e.g., the random forest scored 0.075 on this subset).

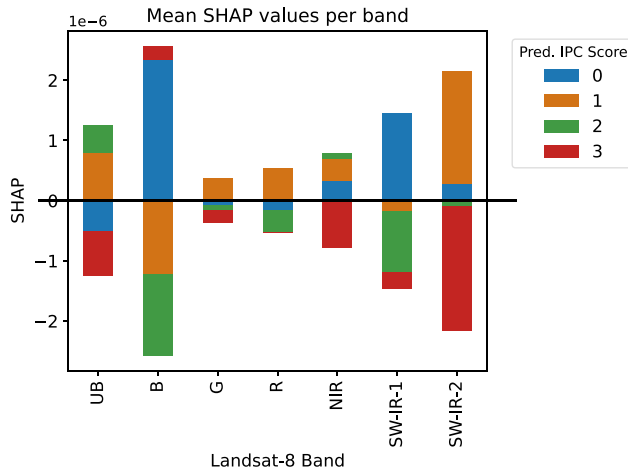
### 6.7. Seasons

To rule out the possibility that IPC scores correlate heavily with seasons, and that the model relies on this correlation to predict IPC scores by predicting which season an image was taken. Figure 8 shows the macro F1 of the best SSSL model per season, as well as the distribution of both ground truth and predicted IPC scores in the geographically separated test set  $\mathcal{D}_{\text{test}}^{\text{ipc}}$  during that season.

Note that there are far fewer available IPC labels during spring, since these were only collected every 3 months between 2013 and 2015, and after that every 4 months, hence skipping spring. The figure shows that different IPC labels occur in different seasons, so that making accurate IPC predictions cannot be reduced to predicting a satellite tile's season. It also shows that the model does predict different IPC scores during different seasons, hence the model does not attempt to shortcut IPC score prediction by predicting a tile's season.

### 6.8. Feature importance

We compute the importance of input features with the SHAP framework's version of DeepLIFT (Lundberg and Lee, 2017; Shrikumar et al., 2017), a method that attributes the output of a neural network to its individual input features by backpropagating the activations of neurons to the input, and comparing each neuron's activation to a reference activation for that neuron. The reference activations are computed



**Figure 9.** Mean SHAP values per LANDSAT-8 band for 100 tiles per IPC score. A positive mean SHAP value for one band and one predicted IPC score means that strong activations for features in this band make the prediction of this IPC score more likely.

from 700 randomly sampled image tiles per IPC score, and the importance values (SHAP values) are computed for 100 tiles per score against those reference tiles.

Figure 9 shows the importance values per LANDSAT-8 band, where the SHAP values are averaged across the pixels and across all 400 ( $100 \times 4$ ) tiles. It shows that the neural network learns physically sensible patterns: activated infrared bands (NIR and SW-IR wavelengths are reflected by healthy vegetation) contribute positively to lower predicted IPC scores and negatively to higher predicted IPC scores. It further shows that the network does not only look at vegetation greenness: for example, the blue and ultra-blue bands have high SHAP values. Figure C1 in Appendix C shows examples of image tiles and the magnitude and direction of each pixel's contribution toward an IPC score prediction. It shows that pixels portraying vegetation or a river contribute positively toward lower IPC scores.

### 6.9. SSSL vs. random forest food insecurity predictor

Table 3 reports the performance of SSSL and ImageNet pre-training versus the random forest models based on Andree et al. (2020), both on multiclass IPC prediction (with five IPC scores) and on binary IPC prediction (where five IPC scores are mapped into two classes: risk or no risk). The first row represents the random forest using only the handcrafted input features as input. As shown already, the neural networks outperformed the random forest significantly, the unfrozen SSSL model giving relative improvements of 64% (multiclass) and 46% (binary) in macro F1. This is a striking result: in absolute percentage points 25% more accurate results can be obtained by only analyzing widely available raw satellite images, instead of a set of handcrafted features from different sources.

The last two rows represent the same random forest, now using the majority voted IPC prediction per AU per date by the frozen or unfrozen SSSL model as extra input feature. This combination improves the random forest's performance up to more or less the level of the neural network, but not more. The handcrafted features from Andree et al. (2020) and the LANDSAT-8 tiles do not appear to be complementary in this setting. We noticed that when using IPC prediction by the unfrozen SSSL pre-trained neural network as a feature for the random forest, the random forest often copies the neural network prediction, resulting in very similar scores.

Note that the random forests get the pre-computed NDVI feature as input, and yet are significantly outperformed by the neural networks, which means the neural networks manage to extract more useful information than simply an NDVI proxy from the seven satellite image bands.

**Table 3.** Random forest performance for binary and multiclass predictions compared to pre-trained neural networks. The best result per column is marked in bold.

Model	Multiclass F1	Binary F1
Random forest with handcrafted features only (Andree et al., 2020)	0.398	0.562
ImageNet frozen	0.444	0.575
ImageNet unfrozen	0.483	0.720
SSSL frozen	0.587	0.733
SSSL unfrozen	0.654	<b>0.823</b>
Random forest with handcrafted features and predictions of frozen SSSL model	0.645	0.787
Random forest with handcrafted features and predictions of unfrozen SSSL model	<b>0.679</b>	0.811

## 7. Conclusion

Several conclusions can be drawn from this study. We showed that the remote-sensing data and neural networks improve predictions vastly compared to using handcrafted input variables. One important remark is that this conclusion is only valid for regions in which food insecurity is actually linked to phenomena that are observable from satellite images.

Next, we showed that compared to not pre-training or using different weight initialization or pre-training paradigms, the relational reasoning framework of Patacchiola and Storkey (2020) for contrastive pre-training improves predictions significantly, especially when using the spatial and temporal dimensions that are inherent parts of satellite imagery. Self-supervised pre-training fits the domain of satellite imagery especially well, as vast amounts of unlabeled data are (publicly) available. We showed that using spatial and temporal thresholds is preferred over using data augmentations as in Patacchiola and Storkey (2020). The study also found that, unlike Ayush et al. (2021a), using a non-zero spatial threshold and a small temporal threshold would work best on food insecurity prediction. These conclusions remain valid for varying amounts of available labeled data, and in fact, we found the required amount of labels to be low. Our model generalizes to locations that it has not seen during pre-training and/or finetuning, but performance was better and fewer labeled data were needed for locations the model was pre-trained on.

We found that forecasting future food insecurity is difficult, but our proposed model is competitive with baselines. We analyzed whether ground truth and predicted IPC score distributions follow seasons and conclude that food insecurity prediction cannot be reduced to detecting the season. We also analyzed the importance of the satellite's bands and found that the model does not only look at vegetation greenness. We hope this work paves the way for further research into using satellite images to predict food insecurity (and potentially other socioeconomic indicators).

### 7.1. Future work

The first way in which future work could be built upon our work is by using more data. Because of computational resource limitations, we focused our study on satellite images and IPC scores only of Somalia during 7 years. However, LANDSAT-8 images are available for the entire world and continuously since 2013, meaning pre-training data comparable to ours is available in greater quantities by several orders of magnitude. Besides, satellites other than LANDSAT-8 also provide publicly available images. The World Bank also made available much more IPC score data: for 21 developing countries since 2009, meaning much more data are available to test finetuning strategies and food insecurity prediction. These data would be interesting not only to potentially improve the model's performance but also to pinpoint the countries in which satellite images help food insecurity prediction. Although the number of pixels we use exceeds that of Patacchiola and Storkey (2020), as with all deep-learning

applications, and even more so with self-supervised pre-training, it can be expected that the more data are used, the better the performance of the model.

A methodological limitation of this study is that we did not have the resources to do multiple runs for each configuration in each experiment, which would have canceled out some of the inherent stochasticity of training deep-learning models. It would be valuable to test other image encoders, like larger CNNs or different architectures, and more contrastive pre-training baselines, like the methods proposed by Manas et al. (2021) and Ayush et al. (2021a).

The satellite images we used were derived from the LANDSAT-8 satellite and have a resolution of 30 m per pixel. More modern satellites provide images with much higher resolutions of < 1 m per pixel, albeit often commercial and not producing publicly available data.<sup>8</sup> Whereas our experiments showed that it is possible to detect food insecurity from relatively low-resolution satellite images, if it is caused by agricultural factors (like in Somalia), presumably since these factors have a detectable effect on the images, experiments also showed that it was not possible to detect food insecurity in regions where it is caused by political or economical factors (like South Sudan), presumably because these factors do not have a detectable effect on the low-resolution images. However, it seems plausible that some effects of political or economic instability could be detectable from higher-resolution images, like the presence of military vehicles, large civil protests, abandoned factories, and so forth. Hence, future work could test whether food insecurity driven by other factors than agricultural or weather-related ones could be predicted from higher-resolution images.

It would also be interesting to further test the generalization capabilities of the method, like testing how different degrees of distance or dissimilarity to training regions impact performance. Future work could also apply SSSL to different downstream tasks. It would be particularly worthwhile to evaluate SSSL for tasks that require different degrees of temporal and spatial variance by matching the spatiotemporal thresholds accordingly.

**Acknowledgments.** The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation–Flanders (FWO) and the Flemish government.

**Author contribution.** Conceptualization: T.F., E.C., R.C., D.S., M.-F.M.; data curation: R.C.; data visualization: R.C.; methodology: R.C., D.S., T.F., E.C.; writing original draft: R.C., T.F., E.C., D.S., M.-F.M. All authors approved the final submitted draft.

**Competing interest.** The authors declare no competing interests exist.

**Data availability statement.** All LANDSAT-8 images used in this study are publicly available via a number of platforms, for instance, through Google Earth Engine. We make available scripts to export the images used for this study at <https://github.com/rubencart/SSSL-food-security>. FEWS NET data are available from <https://fews.net/data/> (Korpi-Salmela et al., 2012). The handcrafted input features are also publicly available. Andree et al. (2020) state their sources in the appendix and make their preprocessed data available at <https://microdata.worldbank.org/index.php/catalog/3811/>.

**Ethics statement.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Funding statement.** This work is part of the CALCULUS project, funded by the ERC Advanced Grant H2020-ERC-2017 ADG 788506.<sup>9</sup> It also received funding from the Research Foundation–Flanders (FWO) under Grant Agreement No. G078618N.

## References

- Andree B, Chamorro A, Kraay A, Spencer P and Wang D (2020) Predicting food crises. *World Bank Working Paper*, 9412. <https://doi.org/10.1596/1813-9450-9412>
- Ayush K, Uz Kent B, Meng C, Tanmay K, Burke M, Lobell DB and Ermon S (2021a) Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, pp. 10181–10190, October 2021. IEEE.
- Ayush K, Uz Kent B, Tanmay K, Burke M, Lobell DB and Ermon S (2021b) Efficient poverty mapping from high resolution remote sensing images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on*

<sup>8</sup> For example, the SkySat constellation owned by Planet Labs: <https://earth.esa.int/eogateway/missions/skysat>.

<sup>9</sup> <https://calculus-project.eu/>.



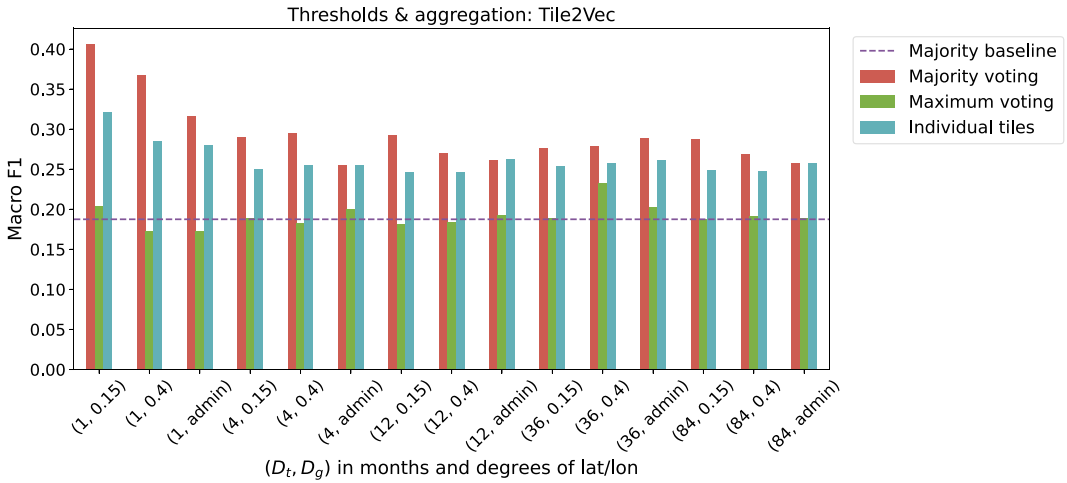
- Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event*, pp. 12–20. AAAI Press.
- Bansal C, Jain A, Barwaria P, Choudhary A, Singh A, Gupta A and Seth A** (2020) Temporal prediction of socio-economic indicators using satellite imagery. In Roy RS (ed.), *CoDS-COMAD 2020: 7th ACM IKDD CoDS and 25th COMAD*, Hyderabad. ACM, pp. 73–81.
- Bengio Y, Courville A and Vincent P** (2013) Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8), 1798–1828.
- Breiman L** (2001) Random forests. *Machine Learning* 45(1), 5–32.
- Burke M, Driscoll A, Lobell DB and Ermon S** (2021) Using satellite imagery to understand and promote sustainable development. *Science* 371(6535), eabe8628.
- Chen T, Kornblith S, Norouzi M and Hinton GE** (2020) A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event, Volume 119 of Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR.
- Cheng G, Han J and Lu X** (2017) Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE* 105(10), 1865–1883.
- Chopra S, Hadsell R and LeCun Y** (2005) Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, pp. 539–546. IEEE Computer Society.
- Chu Y, Cao G and Hayat H** (2016) Change detection of remote sensing image based on deep neural networks. In *Proceedings of the 2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016)*, Beijing, pp. 262–267. Atlantis Press.
- de Jong KL and Bosman AS** (2019) Unsupervised change detection in satellite images using convolutional neural networks. In *International Joint Conference on Neural Networks, IJCNN 2019*, Budapest, pp. 1–8. IEEE.
- Deng J, Dong W, Socher R, Li L, Li K and Fei-Fei L** (2009) Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE Computer Society.
- Devlin J, Chang M, Lee K and Toutanova K** (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics.
- Dosovitskiy A, Fischer P, Springenberg JT, Riedmiller MA and Brox T** (2016) Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(9), 1734–1747.
- Florensa C, Degraeve J, Heess N, Springenberg JT and Riedmiller M** (2019) Self-supervised learning of image embedding for continuous control. *CoRR*, abs/1901.00943.
- Goldblatt R, Heilmann K and Vaizman Y** (2019) Can medium-resolution satellite imagery measure economic activity at small geographies? Evidence from landsat in Vietnam. *The World Bank Economic Review* 34, 635–653. <https://doi.org/10.1093/wber/lhz001>
- Gong M, Zhao J, Liu J, Miao Q and Jiao L** (2016) Change detection in synthetic aperture radar images based on deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 27(1), 125–138.
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R** (2017) Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202, 18–27.
- Grill J, Strub F, Alché F, Tallec C, Richemond PH, Buchatskaya E, Doersch C, Pires BÁ, Guo Z, Azar MG, Piot B, Kavukcuoglu K, Munos R and Valko M** (2020) Bootstrap your own latent – A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual*.
- Han T, Xie W and Zisserman A** (2020) Self-supervised co-training for video representation learning.
- He K, Fan H, Wu Y, Xie S and Girshick RB** (2020) Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, pp. 9726–9735. Computer Vision Foundation/IEEE.
- He K, Zhang X, Ren S and Sun J** (2016) Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, pp. 770–778. IEEE Computer Society.
- Hillbruner C and Moloney G** (2012) When early warning is not enough—Lessons learned from the 2011 Somalia famine. *Global Food Security* 1(1), 20–28.
- Hu W, Patel JH, Robert Z, Novosad P, Asher S, Tang Z, Burke M, Lobell DB and Ermon S** (2019) Mapping missing population in rural India: A deep learning approach with satellite imagery. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019*, Honolulu, HI, pp. 353–359. ACM.
- Jaiswal A, Babu AR, Zadeh MZ, Banerjee D and Makedon F** (2021) A survey on contrastive self-supervised learning. *Technologies* 9(1), 2.
- Jean N, Burke M, Xie M, Davis M and Ermon S** (2016) Combining satellite imagery and machine learning to predict poverty. *Science* 353(6301), 790–794.



- Jean N, Wang S, Samar A, Azzari G, Lobell DB and Ermon S (2019) Tile2vec: Unsupervised representation learning for spatially distributed data. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, HI, pp. 3967–3974. AAAI Press.
- Kang J, Fernández-Beltrán R, Duan P, Liu S and Plaza AJ (2021) Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing* 59(3), 2598–2610.
- Kingma DP and Ba J (2015) Adam: A method for stochastic optimization. In Bengio Y and LeCun Y (eds.), *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA.
- Korpi-Salmela K, Negre T and Nkuzimana T (2012) *Integrated Food Security Phase Classification (IPC) Technical Manual Version 2.0*. Rome: Food and Agriculture Organization of the United Nations.
- Kotkar SR and Jadhav B (2015) Analysis of various change detection techniques using satellite images. In *2015 International Conference on Information Processing (ICIP)*, Québec City, QC, pp. 664–668. IEEE.
- Kussul N, Lavreniuk M, Skakun S and Shelestov A (2017) Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters* 14(5), 778–782.
- Le-Khac PH, Healy G and Smeaton AF (2020) Contrastive representation learning: A framework and review. *IEEE Access* 8, 193907–193934.
- Lentz E, Michelson H, Baylis K and Zhou Y (2019) A data-driven approach improves food insecurity crisis prediction. *World Development* 122, 399–409.
- Li K, Wan G, Cheng G, Meng L and Han J (2020) Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* 159, 296–307.
- Li W, Fu H, Yu L and Cracknell A (2017) Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing* 9(1), 22.
- Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J and Tang J (2023) Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge & Data Engineering*, 35(1), 857–876.
- Lundberg SM and Lee S (2017) A unified approach to interpreting model predictions. In Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, December 4–9, 2017, Long Beach, CA, pp. 4765–4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- Mañosa O, Lacoste A, Giró-i Nieto X, Vázquez D and Rodríguez P (2021) Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2021*, Montreal, QC, pp. 9414–9423. IEEE.
- Mikolov T, Chen K, Corrado G and Dean J (2013) Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR, ICLR 2013*, Scottsdale, AZ, Workshop Track Proceedings.
- Misra I and van der Maaten L (2020) Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, pp. 6706–6716. Computer Vision Foundation/IEEE.
- Mohanty SP, Czakov J, Kaczmarek KA, Pyskir A, Tarasiewicz P, Kunwar S, Rohrbach J, Luo D, Prasad M, Fleer S, Göpfert JP, Tandon A, Mollard G, Rayaprolu N, Salathe M and Schilling M (2020) Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence* 3, 534696. <https://doi.org/10.3389/frai.2020.534696>
- Neuvauori P, Narra N and Lipping T (2019) Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture* 163, 104859. <https://doi.org/10.1016/j.compag.2019.104859>
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang EZ, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J and Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, Vancouver, BC, pp. 8024–8035.
- Patacchiola M and Storkey AJ (2020) Self-supervised relational reasoning for representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual*.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, VanderPlas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12, 2825–2830.
- Qian R, Meng T, Gong B, Yang M, Wang H, Belongie SJ and Cui Y (2021) Spatiotemporal contrastive video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual*, pp. 6964–6974. Computer Vision Foundation/IEEE.
- Roser M and itchie H (2019) Hunger and undernourishment. *Our World in Data*. <https://ourworldindata.org/hunger-andundernourishment>
- Rouditchenko A, Zhao H, Gan C, McDermott J and Torralba A (2019) Self-supervised audio-visual co-segmentation. In *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2357–2361. <https://doi.org/10.1109/ICASSP.2019.8682467>
- Roy DP, Wulder MA, Loveland TR, Woodcock CE, Allen RG, Anderson MC, Helder D, Irons JR, Johnson DM, Kennedy R, Scambos T, Schaaf C, Schott J, Sheng Y, Vermote E, Belward A, Bindschadler R, Cohen W, Gao F, Hipple J, Hostert P,

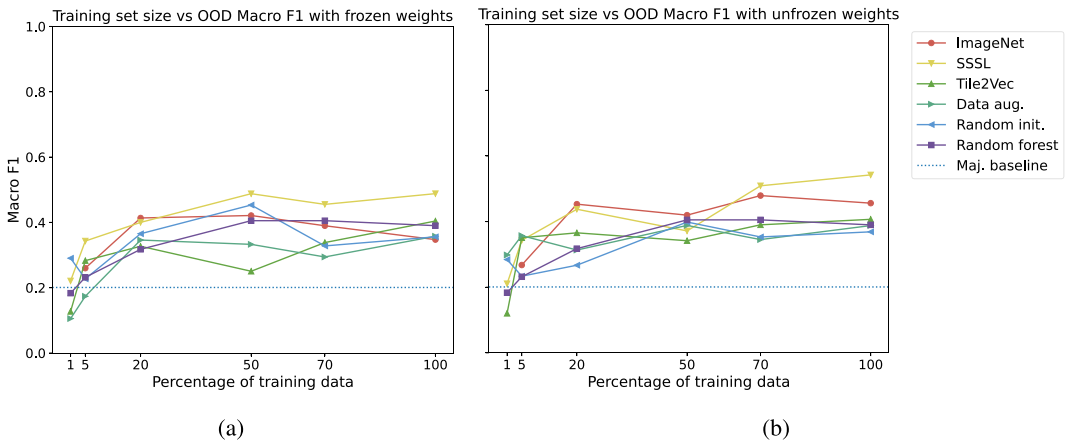
- Huntington J, Justice C, Kilic A, Kovalskyy V, Lee Z, Lymburner L, Masek J, McCorkel J, Shuai Y, Trezza R, Vogelmann J, Wynne R and Zhu Z (2014) Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment* 145, 154–172.
- Rustowicz RM, Cheong R, Wang L, Ermon S, Burke M and Lobell DB (2019) Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019*, Long Beach, CA, pp. 75–82. Computer Vision Foundation/IEEE.
- Schmarje L, Santarossa M, Schröder S-M and Koch R (2021) A survey on semi-, self- and unsupervised learning for image classification. *IEEE Access* 9, 82146–82168. <https://doi.org/10.1109/ACCESS.2021.3084358>
- Sheehan E, Meng C, Tan M, Uz kent B, Jean N, Burke M, Lobell DB and Ermon S (2019) Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*, Anchorage, AK, pp. 2698–2706. ACM.
- Shrikumar A, Greenside P and Kundaje A (2017) Learning important features through propagating activation differences. In Precup D and Teh YW (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, Volume 70 of Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR. <http://proceedings.mlr.press/v70/shrikumar17a.html>
- The European Space Agency (2021) Sentinel Online – ESA – Sentinel. Available at <https://sentinels.copernicus.eu/web/sentinel/home> (accessed 21 May 2021).
- Townsend AC and Bruce DA (2010) The use of night-time lights satellite imagery as a measure of Australia’s regional electricity consumption and population distribution. *International Journal of Remote Sensing* 31(16), 4459–4480.
- Uz kent B, Sheehan E, Meng C, Tang Z, Burke M, Lobell DB and Ermon S (2019) Learning to interpret satellite images using wikipedia. In Kraus S (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, Macao*, pp. 3620–3626. ijcai.org.
- Vali A, Comai S and Matteucci M (2020) Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sensing* 12(15), 2495.
- van den Oord A, Li Y and Vinyals O (2018) Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.
- Wang AX, Tran C, Desai N, Lobell DB and Ermon S (2018) Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*, Menlo Park and San Jose, CA, pp. 50:1–50:5. ACM.
- Wang D, Andree BPJ, Chamorro AF and Spencer PG (2020a) Stochastic modeling of food insecurity. *Policy Research Working Papers*. Washington, DC: World Bank.
- Wang Z, Li H and Rajagopal R (2020b) Urban2vec: Incorporating street view imagery and POIs for multi-modal urban neighborhood embedding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, pp. 1013–1020. AAAI Press.
- Williams DL, Goward S and Arvidson T (2006) Landsat. *Photogrammetric Engineering & Remote Sensing* 72(10), 1171–1178.
- Wu Z, Xiong Y, Yu S and Lin D (2018) Unsupervised feature learning via non-parametric instance-level discrimination. *CoRR*, abs/1805.01978.
- Yeh C, Perez A, Driscoll A, Azzari G, Tang Z, Lobell D, Ermon S and Burke M (2020) Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* 11(1), 1–11.
- Zhang R, Isola P and Efros AA (2017) Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, pp. 645–654. IEEE Computer Society.

**Appendix A: Thresholds and score aggregation—Extra figures.**



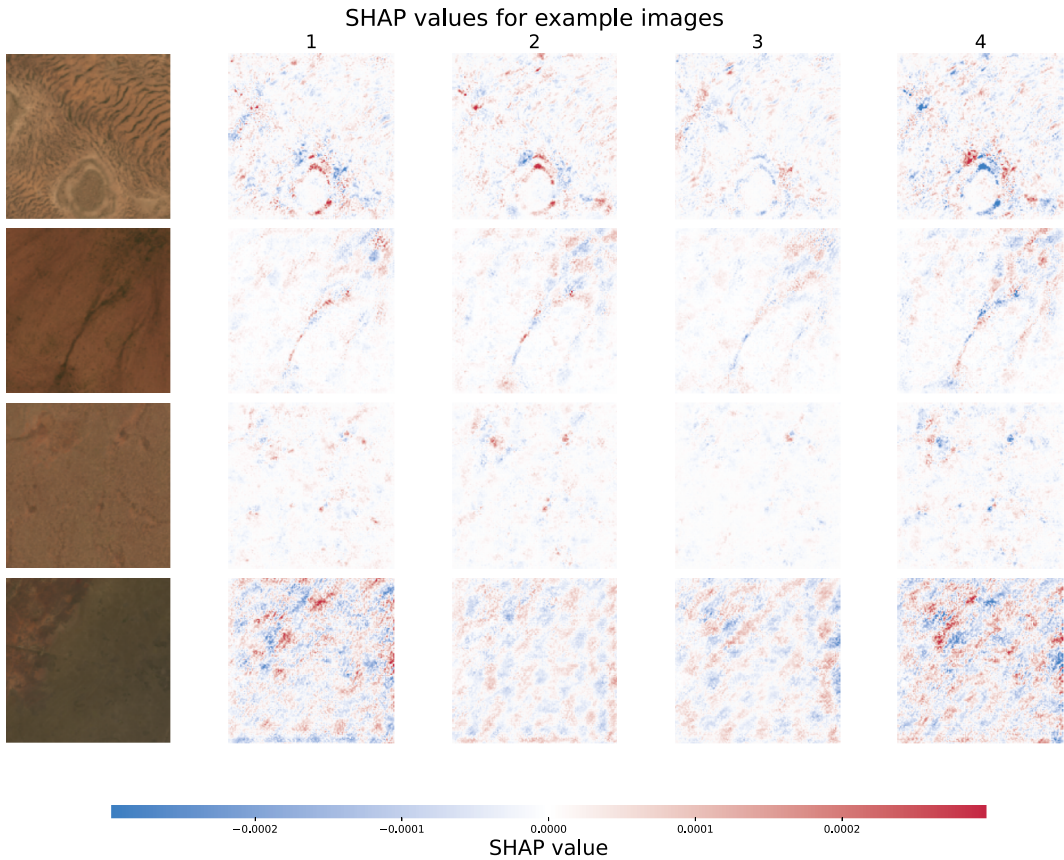
**Figure A1.** Macro F1 on validation set  $\mathcal{D}_{val}^{ipc}$  using different configurations of positive and negative pairs for Tile2Vec pre-training, with  $D_g$  and  $D_t$  denoted on the x-axis. The baseline in this plot always predicts the majority class. “admin” means using administrative units instead of longitude/latitude to define spatial positive pairs.

**Appendix B: Performance on out-of-domain test set with decreasing training set size**



**Figure B1.** Test macro F1 on out-of-domain test set  $\mathcal{D}_{ood}^{ipc}$  with frozen (a) and unfrozen (b) CNN backbone weights for models with different weight initializations using increasing amounts of labeled training data.

## Appendix C: Importance of input features: Geographical



**Figure C1.** Rows show example images with ground truth IPC scores in ascending order (first row shows an image with IPC score 1, etc.), and the last four columns show the SHAP values for the red, near-infrared, and the first shortwave infrared input bands for an output IPC score prediction of 1–4. The pixel contributions follow image features like vegetation, and one pixel contributes in opposite direction to different IPC scores.