Generalized Classification of Satellite Image Time Series with Thermal Positional Encoding

Joachim Nyborg[1,3] Charlotte Pelletier[2] Ira Assent[1]

[1] Department of Computer Science, Aarhus University, Denmark
[2] IRISA UMR 6074, Univ. Bretagne Sud, France
[3] FieldSense A/S, Denmark

jnyborg@cs.au.dk, charlotte.pelletier@univ-ubs.fr, ira@cs.au.dk

Abstract

Large-scale crop type classification is a task at the core of remote sensing efforts with applications of both economic and ecological importance. Current state-of-the-art deep learning methods are based on self-attention and use satellite image time series (SITS) to discriminate crop types based on their unique growth patterns. However, existing methods generalize poorly to regions not seen during training mainly due to not being robust to temporal shifts of the growing season caused by variations in climate. To this end, we propose Thermal Positional Encoding (TPE) for attention-based crop classifiers. Unlike previous positional encoding based on calendar time (e.g. *day-of-year), TPE is based on thermal time, which is obtained by accumulating daily average temperatures over the growing season. Since crop growth is directly related to thermal time, but not calendar time, TPE addresses the temporal shifts between different regions to improve generalization. We propose multiple TPE strategies, including learnable methods, to further improve results compared to the common fixed positional encodings. We demonstrate our approach on a crop classification task across four different European regions, where we obtain state-of-the-art generalization results. Our source code is available at* https://github.com/jnyborg/tpe.

1. Introduction

The increase in openly accessible satellite image time series (SITS) has led to the development of deep learning models using remote sensing data that has significantly improved the state of the art in SITS classification tasks. Among these, crop type classification has numerous applications of economic and ecological importance, such as environmental monitoring, food security, and crop price prediction. Time series data is particularly valuable for crop classification, as it enables models to capture crop *phenology*, *i.e.* the progression of growth over time which characterizes different crop types. Specialized deep learning models for the task thus focus on the temporal aspect of the problem, proposing models based on neural network components that process time, such as temporal convolutions [35, 60], recurrent layers [11, 29, 31, 38], or most recently self-attention [40–42].

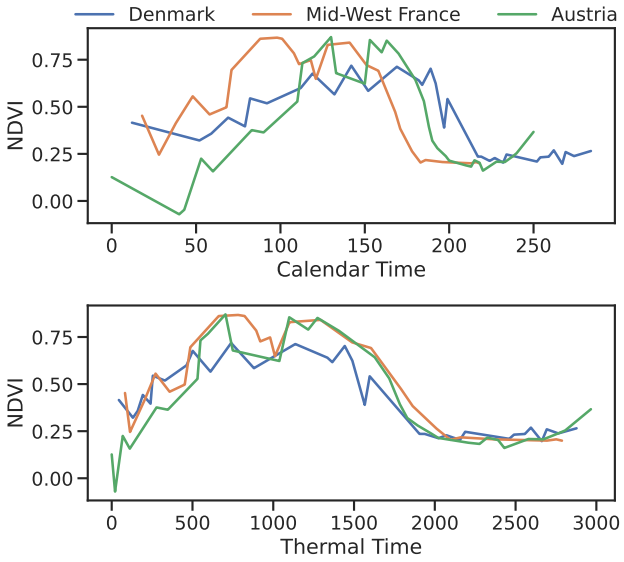


Figure 1. Winter wheat NDVI in different European regions with calendar time and thermal time. With thermal time, temporal shifts of crop growth in different regions are greatly reduced.

Since the growth patterns of crops are similar in different regions of the world [8], it is reasonable to expect that models trained in one region can generalize to another. However, recent works have found that existing models generalize poorly to other regions than those seen during training [24, 33]. Part of the challenge in generalization is the variability in climate which causes different timing of crop growth [5]. For example, in cooler regions, crops reach their growth

stages later than in warmer regions, which models must account for to generalize [33].

To model the progression of time, the predominant approach in existing models is to use calendar time to include temporal context, either during pre-processing to interpolate the data into regular temporal sampling [13, 35, 40, 55] or as an explicit additional input [25, 39]. Notably, state-of-the-art methods based on self-attention input calendar time via *positional encoding* [6, 42]. Since self-attention is position agnostic [52], this provides explicit positional information about the temporal location of images within the growing season. This helps crop classification as the particular timing for the phenological events of a crop type can be an important clue in its classification, *e.g.* to distinguish spring wheat from winter wheat. However, the phenological calendar timing of one region is not generally shared with other regions due to temporal shifts, which causes existing models to generalize poorly [17, 33].

To overcome this challenge, we propose Thermal Positional Encoding (TPE) to improve the generalization of crop classifiers. Our core idea is to use a representation that captures the climatic variation affecting growth rates without relying on calendar time. To this end, we propose positional encoding based on *thermal time* [26, 27] for self-attention models. Thermal time is typically measured for crops by units of *Growing Degree Days* (GDD) [8, 26, 28, 58], computed by accumulating daily average temperatures above a baseline. As crop growth is directly related to the accumulation of heat over the growing season [5, 16], an earlier crop growth corresponds to an earlier increase in GDD and vice versa. This is illustrated in Figure 1 using normalized difference vegetation index (NDVI) to display winter wheat phenology in three different regions. Thermal time improves generalization of models by making SITS from different regions invariant to temporal shifts. At the same time, it provides a temporal location of images which allows thermal time to directly replace calendar time in crop classifiers.

To encode positional information, existing works generally use sinusoidal encoding [52]. However, as this approach is predefined and not learned, it lacks flexibility and may not capture crop-specific positional information. In this paper, we propose multiple TPE methods to encode thermal time in a data-driven way. By learning an encoding function instead of, *e.g.* an embedding vector for each position [4, 36], our approach is inductive. This allows us to handle when the thermal time of test regions differs from that of training, which is common in practice. We evaluate our approach on a crop classification task across four different European regions on the TimeMatch dataset [34], containing Sentinel-2 SITS expanded with daily temperature data, and demonstrate that we obtain state-of-the-art generalization results in new regions. Our main contributions are:

- We propose the use of thermal time in crop classification to increase robustness to temporal shifts and improve generalization.
- We propose TPE methods, which are based on thermal time and can easily be implemented in recent attention-based crop classifiers.
- We demonstrate that TPE greatly improves generalization across four different European regions.

2. Related Work

Satellite Image Time Series Classification. Multiple traditional machine learning approaches, such as random forests or support vector machines, have been applied to crop classification [12, 53, 54, 56]. These approaches require input features to be extracted by hand. For instance, a widely used feature is NDVI, combining the red and near-infrared spectral bands, which relates to the photosynthesis of crops [47]. Other works also include phenological features [15, 50] or meteorological information [59]. Although these handcrafted features are robust and interpretable, deep learning approaches are mostly employed as they enable the automatic extraction of richer features from raw SITS. Deep convolutional networks have been widely applied to process the spatial dimensions of the data [19, 39], while the temporal dimension has been processed by recurrent units [31, 38], 1D convolutions [35, 60], or combinations thereof [13, 39]. Recently, self-attention [52] has led to significant improvements in pixel [40] and parcel classification [41, 42], as well as semantic and panoptic segmentation [6]. Since self-attention is position-agnostic, existing works use sinusoidal positional encoding [52] of calendar time to capture the position of images in the growing season. We propose positional encoding based on thermal time [26, 27] to improve the generalization of the promising self-attention mechanism.

Domain Generalization for SITS. Several prior works have reported that existing crop classification models fail to generalize across space and time due to not being robust to temporal shifts of the growing season [17, 23, 33, 55]. This problem has mainly been tackled by unsupervised domain adaptation (UDA), where models are trained with labeled data from a source region and unlabeled data from a target region [48]. Phenology Alignment Network [55] addresses this problem by learning domain-invariant features obtained with a maximum mean discrepancy loss [49] for the unlabeled target data. TimeMatch [33] obtains further improvements by directly estimating the temporal shift of the target region, and utilizing the shift estimation to train with pseudo-labels for the unlabeled target region. Our setting differs from UDA, as we do not aim to adapt models to particular regions by training with unlabeled data, but to improve the generalization of a crop classifier model trained with labeled data from multiple areas to any new region.

Most similar to our work, Kerner *et al*. [17] improve the generalization of crop classifiers by inputting satellite data at specific time steps which correspond to particular growth stages (greenup, peak, and senescence), computed from the NDVI sequence for each input. By dynamically selecting these time steps, this approach can account for temporal shifts of the growing season, but information is lost since the complete time series is not involved in the prediction. In comparison, we aim to train self-attention models which attend to the most relevant time steps in the complete time series automatically by incorporating thermal time.

Positional Encodings. A vast literature exists in positional encoding for the self-attention mechanism. Absolute positional encoding is most widely used. In the original Transformers [52], vectors are encoded from the absolute position in the sequence by sinusoidal functions, but this approach is less flexible as the vectors are fixed and not learned. To overcome this issue, a common approach is to learn an embedding vector for each position [4, 36] similar to word embeddings, but this approach requires all possible positions to be seen during training to ensure all the embeddings are updated by gradient descent. This is ill-suited for irregularly sampled SITS, which does not guarantee that all possible (calendar or thermal) positions are available for training. Instead, approaches that learn a function that maps positions to vectors [20, 21, 32] do not have this requirement and can thus generalize to unseen positions at test time. We therefore build upon these in this paper.

Another line of work is relative positional encoding [3, 10, 44], which encodes the positional difference between each pair in the input sequence instead of the absolute position of individual elements. While relative positions can be more relevant than absolute in other tasks, in SITS classification, the absolute position is crucial information. For example, a satellite image taken during the winter will not contain the same information about crop growth compared to an image from the spring, which cannot be captured by only the relative positions, *e.g*. the difference in days between the two images. Thus, we focus on absolute positional encoding in this work.

3. Self-Attention for Crop Classification

In crop classification, we are given a satellite image time series $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)}]$, where T is the length of the time series. The goal of the classification task is to associate $\boldsymbol{x}$ with one of K classes. In our setting, each $\boldsymbol{x}^{(t)} \in \mathbb{R}^{T \times N \times C}$ consists of a sequence of N pixels of C spectral bands within a *parcel*, *i.e*., a homogeneous agricultural plot of land. This approach requires parcel shapes to be available in the region for classification, which is widely available in the European Union (EU) [43] or can alternatively be acquired by a segmentation step [6, 45].

Our goal is to improve the generalization of existing crop classifiers by accounting for temporal shifts of the growing season. Owing to its state-of-the-art performance, we build upon the PSE+LTAE model [41]. The network consists of the Pixel-Set Encoder (PSE) and the Lightweight Temporal Attention Encoder (LTAE). Given a randomly sampled pixel-set of size S among the N available pixels of an input $\boldsymbol{x}$, the PSE handles the spatial and spectral context of SITS by processing each time step individually to a sequence of embedding vectors $\boldsymbol{e} = [\boldsymbol{e}^{(1)}, \dots, \boldsymbol{e}^{(T)}] \in \mathbb{R}^{T \times D}$, where D is the embedding dimension. PSE does not process the temporal dimension. We thus focus on handling temporal shifts in the LTAE module. Given $\boldsymbol{e}$, LTAE extracts temporal features using a simplified version of the multi-headed self-attention, as we describe next.

Self-Attention. In the original Transformer model [52], self-attention is computed with a query-key-value triplet $(\boldsymbol{q}^{(t)}, \boldsymbol{k}^{(t)}, \boldsymbol{v}^{(t)})$ for each element in the input sequence using three fully-connected layers. The output is a sequence where each element is a sum of all values $\boldsymbol{v}^{(t)}$ weighted by their attention score. The attention scores for a time step t are computed as the similarity (dot product) between all keys and the query $\boldsymbol{q}^{(t)}$, re-scaled by a softmax layer. The computation of the query-key-value triplets can be performed in parallel, which enables the Transformer model to take full advantage of GPUs for a significant speed increase compared to the sequential computation of recurrent neural networks (RNN). In multi-headed self-attention, the triplets are computed multiple times in parallel with different parameters, or "heads", which further increase efficiency and also the representational capacity as each head can specialize in different parts of the sequence.

Sinusoidal Positional Encoding. As the self-attention mechanism is position-agnostic [52], various positional encodings (PE) have been introduced to capture positional information. This is typically done by mapping scalar positions to a vector, either by learning or by heuristics, and adding each embedding vector with their positional encoding $\boldsymbol{e}^{(t)} + \boldsymbol{p}^{(t)}$ before applying self-attention. The original Transformer model [52] uses a fixed sinusoidal encoding with predefined wavelengths, defined as:

$$\boldsymbol{p}^{(t)} = [\sin(\omega_i t), \cos(\omega_i t)]_{i=1}^{D/2} \tag{1}$$

where $\omega_i = (1/\tau)^{2i/D}$ and $\tau = 10000$.

Lightweight Temporal Attention Encoder. While the original self-attention maps the input embeddings $\boldsymbol{e}$ to an output sequence of embeddings, the goal of SITS classification is to map the entire time series into a single embedding.

To address this, the LTAE module [41] modifies the self-attention mechanisim by replacing the queries $\boldsymbol{q}^{(t)}$ with a single learnable "master" query $\hat{\boldsymbol{q}}$, resulting in a single output embedding instead of a sequence. The computation is also made more lightweight by employing a channel grouping strategy [57], where each attention head operates on its own subset of input channels. The LTAE module uses the sinusoidal PE (Equation 1 with $\tau = 1000$) but encodes the day of the year $\text{day}^{(t)}$ instead of the position index t. This enables the model to account for the inconsistent temporal sampling of SITS, but also introduces problems with handling temporal shift [33].

4. Method

In this work, we observe that the positional encoding used by the LTAE module has two issues. First, since it encodes calendar time, it introduces the temporal shift problem as displayed in Figure 1. While calendar time is useful to identify the crop types in a particular region, it hinders generalization to new regions [17]. For example, while spring and winter crops can be similar in appearance, they are easily separated by the timing of their growth stages as spring crops are planted later in a growing season than winter crops. However, because of temporal shifts, the same time positions of spring crops could represent winter crops in another region. Without any way of accounting for temporal shifts, calendar time positional encodings are unlikely to generalize. Second, since the positional encoding is fixed and not learned, it prevents the model from taking advantage of end-to-end training the encoding function to further improve generalization [20, 21, 36].

A possible remedy to the first issue is to augment the training data with random temporal shifts, such as ShiftAug, a SITS augmentation technique proposed in [33], so that the model does not learn to associate a specific position with the phenological events seen in the training data. While this solution increases the invariance of the model to temporal shifts, the temporal shift is in some cases an important clue to distinguish crop types—such as the spring and winter crops. Instead, we want models that are shift-invariant between different regions, but shift-variant within the same region. That is, we want models which can use class-wise temporal shifts for classification but are unaffected by temporal shifts of the growing season.

To address the second issue, a common alternative to the fixed sinusoidal encodings is to treat each position as a discrete token that can be uniquely represented as a learnable vector [4, 7, 36]. While this approach enables the model to learn the positional encoding from data, it fails to generalize to positions not encountered during training. This is an issue for high-resolution SITS, as we typically do not have an observation for every possible position. For example, the Sentinel-2 satellites acquire images every five days. Moreover, images with high cloud coverage are often filtered, further reducing the positions available. In comparison, sinusoidal positional encoding is more practical for SITS, as an encoding vector is well-defined for every position independent of the training data.

4.1. Thermal Positional Encodings

We argue that successful positional encoding for SITS should meet the following requirements:

(1) Making SITS from *different* regions shift-*invariant* to address the temporal shift problem.

(2) Making SITS from the *same* region shift-*variant* by providing absolute information of where an observation is located in the growing season.

(3) Must be inductive to be able to handle positions not seen during training.

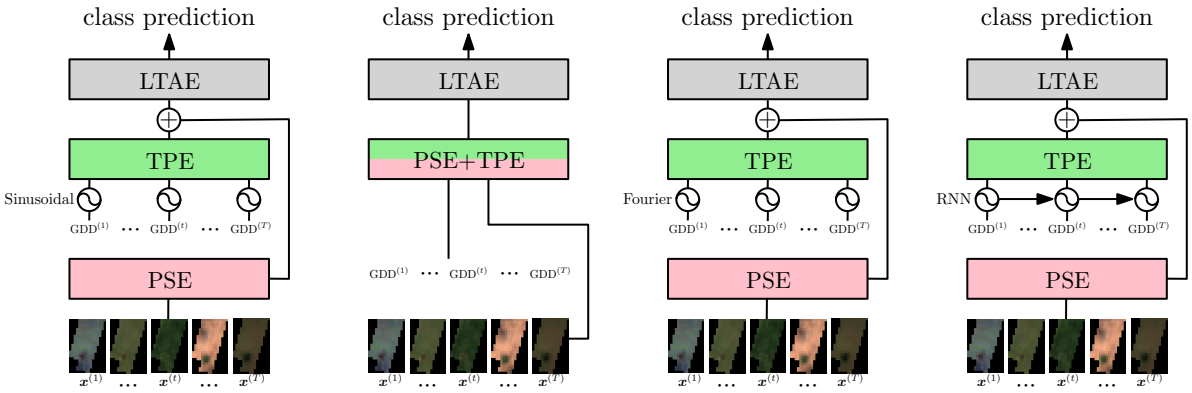
(4) Being data-specific and thus learnable.

While the LTAE sinusoidal positional encoding based on calendar time meets the second and third requirements, it is not invariant to temporal shifts between different regions or trainable which violates the first and fourth requirements. To address this, we replace calendar time with thermal time to meet both the first and second requirements and propose four TPE strategies, including learnable methods to meet the third and fourth requirements.

Thermal time. When studying crop phenology, thermal time is a good proxy for the rate of crop growth [5, 26, 46]. Thermal time is typically measured in units of *growing degree days* (GDD). The GDD measured at a time t is computed by accumulating daily average temperatures above a baseline:

$$\text{GDD}^{(t)} = \sum_{i=1}^{t} \max\left(\frac{T_{min}^{(i)} + T_{max}^{(i)}}{2} - T_{base}, 0\right) \quad (2)$$

where $T_{min}^{(i)}, T_{max}^{(i)}$ is the minimum and maximum temperatures for day i, accumulated for all the previous days $i = 1, 2, \ldots, t$. Temperature values are often clipped to a range $[T_{base}, T_{cap}]$ chosen depending on the crop type. Since we do not know the crop type of the input beforehand, we choose standard values $T_{base} = 0$ and $T_{cap} = 30$ [26, 28] for all crops, since growth typically stagnates below 0°C and does not grow any faster above 30°C. We accumulate from the starting day of the input SITS, in our case January 1. Since GDD is computed by a cumulative sum, it is a monotonically increasing function and thus preserves the order of the input time series. This enables GDD to directly replace day of year for the time positions in the self-attention computation. By replacing calendar time with thermal time, we can



(a) TPE-Sinusoidal (b) TPE-Concat (c) TPE-Fourier (d) TPE-Recurrent

Figure 2. Schematic illustration of our Thermal Positional Encoding (TPE) methods with the PSE+LTAE model [41].

reduce the temporal shift of SITS between different regions while retaining the shift between classes within the same region and thereby satisfy the first and second requirements.

TPE Methods. We propose the following TPE methods to input thermal time to PSE+LTAE [41].

- TPE-Sinusoidal: We replace calendar time with thermal time in the sinusoidal PE, but the encoding is not learned.
- TPE-Concat: We learn SITS and positional input embeddings jointly by concatenating thermal time to an intermediate feature of the PSE module.
- TPE-Fourier: We learn the sinusoidal PE function by the method proposed in [20].
- TPE-Recurrent: We learn a positional encoding function that captures the development in GDD by a recurrent neural network (RNN).

An overview of the TPE methods is shown in Figure 2.

4.2. TPE-Sinusoidal

To use GDD with the sinusoidal PE, we follow Equation 1 but replace t with $\mathrm{GDD}^{(t)}$. The benefit of using the sinusoidal positional encoding for GDD is that an encoding vector is well-defined for every possible GDD value. This ensures that even if we train with only a subset of possible accumulated temperatures, a positional encoding exists for unseen positions at test time. However, as the sinusoidal PE is fixed and not learned, it prevents the model from capturing data-specific positional information for the crop classification task.

4.3. TPE-Concat

While the original Transformer network [52] takes pre-trained word embeddings as inputs, in our case, the embeddings are learned by the PSE module, which is learned simultaneously to the LTAE module. Thus, we propose an alternative to positional encoding where the encoding for the SITS and positions are learned jointly by the PSE. In particular, for each time step t, we concatenate $\mathrm{GDD}^{(t)}$ to the intermediate PSE embedding $\hat{\boldsymbol{e}}^{(t)}$ before the final PSE output layer MLP_2:

$$\boldsymbol{e}^{(t)} = \mathrm{MLP}_2([\hat{\boldsymbol{e}}^{(t)} \,||\, \mathrm{GDD}^{(t)}]), \tag{3}$$

where $[\cdot \,||\, \cdot]$ indicates concatenation. The PSE output layer MLP_2 [42] is a multi-layer perceptron (MLP) consisting of a linear layer, batch normalization [14], and ReLU [30] activation function. We note that this approach is similar to the method of inputting extra parcel geometric features in the original PSE. By concatenating positions to the embedding function, TPE-Concat removes the need for complex positional encoding functions, which may be more beneficial for SITS.

4.4. TPE-Fourier

Li *et al.* [20] propose a learnable PE based on Fourier features [37], which can also be viewed as a generalization of the sinusoidal PE. For a position $t \in \mathbb{R}$, the Fourier PE is computed by:

$$\boldsymbol{r}^{(t)} = \frac{1}{\sqrt{D}}[\cos(\boldsymbol{W}_r t) \,||\, \sin(\boldsymbol{W}_r t)], \tag{4}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{D/2}$ is a trainable vector. To give the representation additional capacity, the encoding is passed through an MLP:

$$\boldsymbol{p}^{(t)} = \mathrm{MLP}(\boldsymbol{r}^{(t)})\boldsymbol{W}_p \tag{5}$$

where MLP consists of a linear layer with GeLU [9] activation function, and $\boldsymbol{W}_p$ are parameters for projecting the representation to the dimension of the input embeddings. The TPE-Fourier reveals whether it is more beneficial to learn the sinusoidal PE compared to the fixed TPE-Sinusoidal.

4.5. TPE-Recurrent

Compared to natural language processing (NLP), where positions typically increase linearly with the sequence length, GDD increases non-linearly over the growing season (see Figure 4), as a result of the higher daily temperatures during the summer than the winter. It may therefore be beneficial not to only encode independent GDD values, but also incorporate previous values to account for different rates of crop growth over the year. To handle this, we propose to use an RNN to learn the positional encoding. RNNs have been successfully used for positional encoding in NLP tasks [21, 32]. We follow the RNN approach of Liu *et al*. [21]. In particular, we use a GRU [1], which computes its output $\boldsymbol{h}^{(t)} \in \mathbb{R}^{H_{out}}$ for each time step t given an input $\boldsymbol{z}^{(t)} \in \mathbb{R}^{H_{in}}$ and the previous hidden state $\boldsymbol{h}^{(t-1)}$ by:

$$\boldsymbol{h}^{(t)} = \mathrm{GRU}(\boldsymbol{z}^{(t)}, \boldsymbol{h}^{(t-1)}). \tag{6}$$

Then, we obtain a positional encoding with target dimension D by a linear projection:

$$\boldsymbol{p}^{(t)} = \boldsymbol{W}_p^{\top} \boldsymbol{h}^{(t)} + \boldsymbol{b}_p, \tag{7}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{H_{out} \times D}$ and $\boldsymbol{b}_p \in \mathbb{R}^{D}$. Instead of scalar values $\mathrm{GDD}^{(t)}$, we use vectorized positions as the inputs $\boldsymbol{z}^{(t)}$, which are obtained by obtained by the sinusoidal positional encoding of $\mathrm{GDD}^{(t)}$ (Equation 1) as done in [21]. TPE-Recurrent learns a positional encoding that captures the temporal development in GDD, but is more computationally expensive due to the sequential computation of an RNN.

5. Experiments

5.1. Setup

Dataset. We evaluate our approach on the TimeMatch dataset [34] with Sentinel-2 L1C SITS from four different tiles: 33UVP (Austria), 32VNH (Denmark), 30TXT (midwest France), and 31TCJ (southern France). We refer to these regions by AT1, DK1, FR1, and FR2, respectively. We display the locations of these tiles in Figure 3. The dataset contains all available observations of these tiles between January 1, 2017, and December 31, 2017, with cloud cover $\leq 80\%$ and coverage $\geq 50\%$. The atmospheric bands (1, 9, and 10) are left out, keeping the remaining 10 spectral bands. The 20m bands are bilinearly interpolated to 10m.

The dataset is prepared for parcel classification by cutting the pixels of each parcel from the SITS using geo-referenced parcel shapes available from the Land Parcel Identification

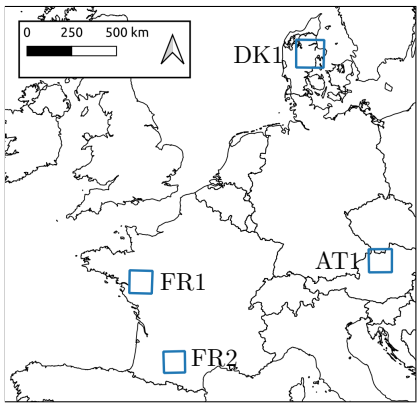


Figure 3. The geographical locations in Europe of the four Sentinel-2 tiles in the dataset [34]. Figure adapted from [33].

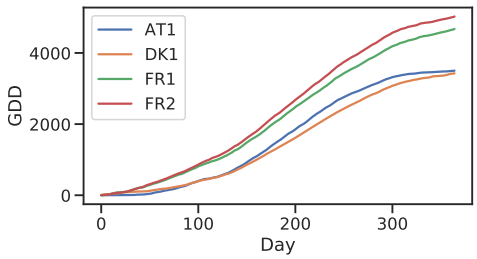


Figure 4. The development of GDD on average in the four Sentinel-2 tiles from January 1 to December 31, 2017.

System (LPIS) in each country. The total amount of parcels is 280K with 15 crop classes. The frequency of these classes varies greatly between tiles, for example, sunflowers are only frequent in the two France tiles. To ensure all tiles have enough samples of each class to learn their classification, we select the 9 crop types with at least 200 samples in all tiles: corn, horsebeans, meadow, spring barley, winter barley, winter rapeseed, winter triticale, winter wheat, and unknown. Here, the unknown class contains all parcels with crop type not of the other 8 classes. Each tile has its own train/validation/test sets, created by assigning all parcels in a tile at random to these sets by a 70%/10%/20% ratio.

We expand the TimeMatch dataset with weather information from the Europe-wide E-OBS dataset [2]. We use the daily minimum and maximum temperature from the $0.1°$ regular grid of 2017 to compute GDD for each parcel, georeferenced by the parcel centroid. Figure 4 displays the average GDD computed for the four regions, showing the southern France tile FR2 is the warmest and the Danish tile DK1 the coldest.

Implementation details. We follow the original implementation of PSE+LTAE [41]. All models are trained for 100 epochs with a batch size of 128 on a single GTX 1080Ti GPU

Method	AT1		DK1		FR1		FR2		Avg.	
	F1	OA	F1	OA	F1	OA	F1	OA	F1	OA
PSE+LTAE [41]	68.3	90.5	55.4	62.6	74.6	90.9	73.5	87.5	68.0	82.9
+ w/o PE	84.1	94.4	66.3	76.2	79.3	91.9	74.0	86.4	75.9	87.2
+ w/ ShiftAug [33]	84.2	94.1	71.6	78.5	83.9	93.3	79.8	89.4	79.9	88.8
+ TPE-Sinusoidal	85.6	94.7	78.7	84.8	83.0	92.6	81.1	**90.4**	82.1	90.6
+ TPE-Concat	85.7	94.7	78.6	83.1	85.1	93.3	**81.4**	89.6	82.7	90.2
+ TPE-Fourier	84.7	94.4	79.0	**86.0**	77.3	91.5	80.0	89.4	80.3	90.3
+ TPE-Recurrent	**86.5**	**95.0**	**80.3**	85.4	**86.0**	**93.8**	80.5	89.8	**83.3**	**91.0**
Upper-bound	94.6	97.5	92.0	94.0	93.1	96.4	87.4	93.9	91.8	95.4

Table 1. Leave-one-region-out spatial generalization results in macro F1 score (F1) and overall accuracy (OA) (both in %). Each column shows the classification results in a new region after training on the others.

with Adam optimizer [18]. The learning rate is initialized to $1e{-}3$ and decayed each epoch by cosine annealing [22]. We use weight decay of $1e{-}4$. The 16-bit input pixels are normalized to $[0, 1]$ by dividing by $2^{16} - 1$. Our code is available at https://github.com/jnyborg/tpe.

Experimental setup. To evaluate whether our proposed thermal positional encoding improves generalization to new regions, we adopt a leave-one-region-out setup where we hold one Sentinel-2 tile out for testing and train on the remaining. In contrast to the domain adaptation setup of TimeMatch [33], where data is only available from one tile for training, our setup contains multiple regions for training. In practice, we typically have many tiles available for training [43], so this setup allows us to evaluate against the naive approach of improving generalization by adding more training data.

Model variants. In comparison to TPE, we consider the following model variants:

- *PSE+LTAE* [41]. This is the baseline model which encodes calendar time (day of the year) with the sinusoidal positional encoding [52].
- *w/o PE*. This is PSE+LTAE where self-attention is computed without any positional information.
- *w/ ShiftAug* [33]. PSE+LTAE trained with calendar time augmented with random temporal shifts.
- *Upper-bound*. We train the best performing TPE method (TPE-Recurrent) with all four available regions to obtain the results of a fully-supervised upper bound.

5.2. Parcel Classification Results

In Table 1, we detail the performance obtained for the leave-one-region-out spatial generalization experiments. We report the class-averaged F1 score (F1) and the overall accuracy (OA). Compared to calendar time models (top), all our TPE models (bottom) have much better generalization results with the use of thermal time. TPE-Recurrent shows the best performance by being learnable and capturing the temporal development in GDD, increasing F1 on average by $+15.3\%$ over the default PSE+LTAE [41] model and $+3.4\%$ over the ShiftAug [33] augmented model. Our TPE greatly improves generalization, but there is still a gap to the upper-bound performance. TPE addresses the temporal shifts between regions but does not account for changes in the spectral signature of crops, which can be caused by differences in *e.g.* the topography, soil, or varieties of the cultivated crop type. We leave this direction to future work.

Analysis of results. We observe that the default PSE+LTAE with calendar time generalizes worst, obtaining an F1 score of 68.0% on average. Interestingly, simply removing the positional encoding outperforms the baseline significantly, leading to an average performance increase of $+7.9\%$. Since this model variant is given no information about the order of images in the SITS, it is also invariant to temporal shifts, which explains the performance increase. However, without positional information, the model should not be able to model the class-wise timing differences, which should degrade performance. But the performance increase indicates the model is able to do so. We argue that this is because the model is able to extract some positional information from the SITS. For example, satellite images taken during the winter differ from those during the summer, enabling the model to extract some degree of temporal order. However, in the case that two images at different times appear similar, the extracted positions can be ambiguous, which is avoided by providing explicit positional information. This is also indicated by the result of ShiftAug [33], where calendar time is augmented with random temporal shifts, which fur-

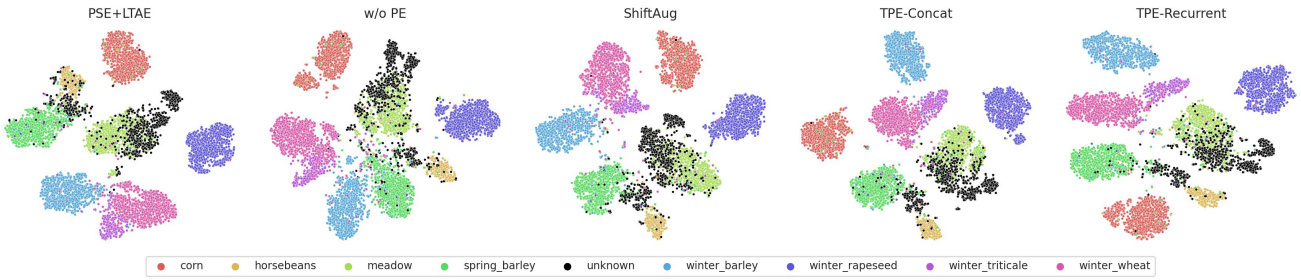


Figure 5. LTAE features of different methods embedded with t-SNE [51] for DK1 after training with the remaining regions.

Method	Training time (s/epoch)
TPE-Sinusoidal	16.1
TPE-Concat	**15.5**
TPE-Fourier	16.4
TPE-Recurrent	17.2

Table 2. The training time of TPE in seconds per training epoch.

ther increases the F1 results by $+11.9\%$ on average over the baseline, outperforming no positional encoding by $+4.0\%$. This indicates that direct positional information is indeed important to the crop classification task to avoid ambiguous order information from images only.

In comparison, our TPE models outperform all calendar time models. This highlights the benefits of using thermal time for reducing the temporal shift between different regions without introducing any augmentations, while also providing explicit positional information for modelling the class-wise timing differences. The TPE-Sinusoidal model is the default PSE+LTAE model but where calendar time positions are replaced with thermal time. This simple change significantly improves the F1 generalization results by $+14.1\%$ on average. Learning a sinusoidal PE with TPE-Fourier, however, is not beneficial, resulting in a decrease in F1 compared to TPE-Sinusoidal by -1.8%. TPE-Concat learns embedding and positional representations jointly in the PSE module, and obtains comparable results to TPE-Sinusoidal, with higher F1 ($+0.6\%$) but lower OA (-0.4%). But as TPE-Sinusoidal introduces extra computation because of the sinusoidal encoding function, TPE-Concat is computationally more efficient as shown in Table 2. This indicates that the approach of adding positional encodings to input embeddings common in natural language processing may be unnecessary for SITS classification. TPE-Recurrent learns a positional encoding that captures the development in GDD, leading to an increase in F1 of $+1.2\%$ over TPE-Sinusoidal. TPE-Recurrent thus shows the best performance but also introduces sequential computation which increases computation requirements as shown in Table 2. We suggest the choice of TPE method is a trade-off between performance and efficiency. Practitioners can easily implement TPE-Concat by concatenating thermal time in PSE [41], and enjoy improved generalization and efficiency. If more computation can be afforded, TPE-Recurrent offers the best results.

5.3. Visual Analysis

To better understand how TPE obtains improvements, we visualize in Figure 5 t-SNE [51] embeddings of features output by the LTAE. For TPE methods, we observe denser and better separated clusters, indicating better class separation by accounting for temporal shifts. For the baseline PSE+LTAE model [41], we observe some classes are well clustered despite the temporal shift, such as corn and winter rapeseed, indicating these classes are less impacted by temporal shifts. Others are mixed, such as spring barley/horsebeans and winter wheat/winter triticale. We observe that removing the PE results in less dense clusters. Particularly, the clusters for spring barley and winter barley overlaps. This could indicate difficulties in resolving class-wise temporal shifts, since these are better separated with ShiftAug [33].

6. Conclusion

In this work, we propose Thermal Positional Encodings (TPE) to address the temporal shift issue of SITS classifiers and improve generalization. While existing work uses calendar time, our TPE uses thermal time, which enables models to account for the varying rates of crop growth in different climates and thereby address the temporal shift issue. We propose multiple positional encoding methods for thermal time, including fixed and learned approaches. On a parcel classification dataset with SITS from four different European regions, we demonstrate that TPE significantly improves generalization compared to existing methods.

Acknowledgements. Joachim Nyborg is funded by the *Innovation Fund Denmark* under reference *8053-00240*. We acknowledge the E-OBS dataset from the EU-FP6 project UERRA (https://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu).

References

[1] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1724–1734. ACL, 2014. 6

[2] Richard C Cornes, Gerard van der Schrier, Else JM van den Besselaar, and Philip D Jones. An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018. 6

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 2978–2988. Association for Computational Linguistics, 2019. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019. 2, 3, 4

[5] B. Franch, E.F. Vermote, I. Becker-Reshef, M. Claverie, J. Huang, J. Zhang, C. Justice, and J.A. Sobrino. Improving the timeliness of winter wheat production forecast in the United States of America, Ukraine and China using MODIS data and NCAR Growing Degree Day information. *Remote Sensing of Environment*, 161:131–148, 2015. 1, 2, 4

[6] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021. 2, 3

[7] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017. 4

[8] Pengyu Hao, Liping Di, Chen Zhang, and Liying Guo. Transfer learning for crop classification with cropland data layer data (CDL) as training samples. *Science of The Total Environment*, 733:138869, 2020. 1, 2

[9] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint*, abs/1606.08415, 2016. 6

[10] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, and Douglas Eck. An improved relative self-attention mechanism for transformer with application to music generation. *arXiv preprint*, abs/1809.04281, 2018. 3

[11] Dino Ienco, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1685–1689, 2017. 1

[12] Jordi Inglada, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, Guadalupe Sepulcre, Sophie Bontemps, Pierre Defourny, et al. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9):12356–12379, 2015. 2

[13] Roberto Interdonato, Dino Ienco, Raffaele Gaetano, and Kenji Ose. DuPLO: A DUal view Point deep Learning architecture for time series classificatiOn. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149:91–104, 2019. 2

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015. 5

[15] Kun Jia, Shunlin Liang, Xiangqin Wei, Yunjun Yao, Yingru Su, Bo Jiang, and Xiaoxia Wang. Land cover classification of Landsat data with phenological features extracted from time series MODIS NDVI data. *Remote Sensing*, 6(11):11518–11532, 2014. 2

[16] Hamlyn G Jones. *Plants and microclimate: a quantitative approach to environmental plant physiology*. Cambridge University Press, 2013. 2

[17] Hannah Kerner, Ritvik Sahajpal, Sergii Skakun, Inbal Becker-Reshef, Brian Barker, Mehdi Hosseini, Estefania Puricelli, and Patrick Gray. Resilient in-season crop type classification in multispectral satellite observations using growth stage normalization. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining Workshops*, 2020. 2, 3, 4

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

[19] Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017. 2

[20] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 4, 5

[21] Xuanqing Liu, Hsiang-Fu Yu, Inderjit Dhillon, and Cho-Jui Hsieh. Learning to encode position for transformer with continuous dynamical model. In *International Conference on Machine Learning*, pages 6327–6335. PMLR, 2020. 3, 4, 6

[22] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 7

[23] Benjamin Lucas, Charlotte Pelletier, Daniel Schmidt, Geoffrey I Webb, and François Petitjean. Unsupervised domain adaptation techniques for classification of satellite image time series. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1074–1077. IEEE, 2020. 2

[24] Benjamin Lucas, Charlotte Pelletier, Daniel Schmidt, Geoffrey I Webb, and François Petitjean. A bayesian-inspired,

deep learning-based, semi-supervised domain adaptation technique for land cover mapping. *Machine Learning*, pages 1–33, 2021. 1

[25] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 75–82, 2019. 2

[26] Gregory S. McMaster and W.W. Wilhelm. Growing degree-days: one equation, two interpretations. *Agricultural and Forest Meteorology*, 87(4):291–300, 1997. 2, 4

[27] HJ Mederski, ME Miller, and CR Weaver. Accumulated heat units for classifying corn hybrid maturity 1. *Agronomy Journal*, 65(5):743–747, 1973. 2

[28] Perry Miller, Will Lanier, and Stu Brandt. Using growing degree days to predict plant stages. *Ag/Extension Communications Coordinator, Communications Services, Montana State University-Bozeman, Bozeman, MO*, 59717(406):994–2721, 2001. 2, 4

[29] Dinh Ho Tong Minh, Dino Ienco, Raffaele Gaetano, Nathalie Lalande, Emile Ndikumana, Faycal Osman, and Pierre Maurel. Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1. *IEEE Geoscience and Remote Sensing Letters*, 15(3):464–468, 2018. 1

[30] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814. Omnipress, 2010. 5

[31] Emile Ndikumana, Dinh Ho Tong Minh, Nicolas Baghdadi, Dominique Courault, and Laure Hossard. Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sensing*, 10(8):1217, 2018. 1, 2

[32] Masato Neishi and Naoki Yoshinaga. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 328–338, 2019. 3, 6

[33] Joachim Nyborg, Charlotte Pelletier, Sébastien Lefèvre, and Ira Assent. TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation. *arXiv preprint*, abs/2111.02682, 2021. 1, 2, 4, 6, 7, 8

[34] Joachim Nyborg, Charlotte Pelletier, Sébastien Lefèvre, and Ira Assent. The TimeMatch Dataset, 2021. 2, 6

[35] Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5):523, 2019. 1, 2

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2, 3, 4

[37] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007. 5

[38] Marc Rußwurm and Marco Korner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017. 1, 2

[39] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018. 2

[40] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:421–435, 2020. 1, 2

[41] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 171–181. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[42] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12325–12334, 2020. 1, 2, 5

[43] Maja Schneider, Amelie Broszeit, and Marco Körner. EuroCrops: A pan-european dataset for time series crop type classification. *arXiv preprint*, abs/2106.08151, 2021. 3, 7

[44] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 464–468. Association for Computational Linguistics, 2018. 3

[45] Andrei Stoian, Vincent Poulain, Jordi Inglada, Victor Poughon, and Dawa Derksen. Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems. *Remote Sensing*, 11(17):1986, 2019. 3

[46] David Trudgill, Alois Honek, Daiqin Li, and Nico M. van Straalen. Thermal time - concepts and utility. *Annals of Applied Biology*, 146:1–14, 01 2005. 4

[47] Compton J Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979. 2

[48] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 2016. 2

[49] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint*, abs/1412.3474, 2014. 2

[50] Silvia Valero, David Morin, Jordi Inglada, Guadalupe Sepulcre, Marcela Arias, Olivier Hagolle, Gérard Dedieu, Sophie Bontemps, Pierre Defourny, and Benjamin Koetz. Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sensing*, 8(1):55, 2016. 2

[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 8

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2, 3, 5, 7

[53] Francesco Vuolo, Martin Neuwirth, Markus Immitzer, Clement Atzberger, and Wai-Tim Ng. How much does multi-temporal Sentinel-2 data improve crop type classification? *International Journal of Applied Earth Observation and Geoinformation*, 72:122–130, 2018. 2

[54] Sherrie Wang, George Azzari, and David B. Lobell. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sensing of Environment*, 222:303–317, 2019. 2

[55] Ziqiao Wang, Hongyan Zhang, Wei He, and Liangpei Zhang. Phenology alignment network: A novel framework for cross-regional time series crop classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2949, 2021. 2

[56] Brian D Wardlow and Stephen L Egbert. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains. *Remote Sensing of Environment*, 112(3):1096–1116, 2008. 2

[57] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[58] Senshan Yang, Joanne Logan, and David L Coffey. Mathematical formulae for calculating the base temperature for growing degree days. *Agricultural and Forest Meteorology*, 74(1-2):61–74, 1995. 2

[59] Liheng Zhong, Peng Gong, and Gregory S Biging. Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using landsat imagery. *Remote Sensing of Environment*, 140:1–13, 2014. 2

[60] Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, 221:430–443, 2019. 1, 2