# Spatial Context-Aware Networks for Mining Temporal Discriminative Period in Land Cover Detection

Xiaowei Jia [*], Sheng Li[†], Ankush Khandelwal[*], Guruprasad Nayak[*], Anuj Karpatne[‡], Vipin Kumar[*]

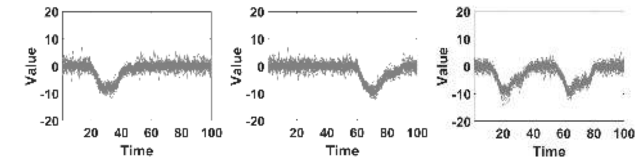
## Abstract

Detecting land use and land cover changes is critical to monitor natural resources and analyze global environmental changes. In this paper, we investigate the land cover detection using the remote sensing data from earth-observing satellites. Due to the natural disturbances, e.g., clouds and aerosoles, and the data acquisition errors by devices, remote sensing data frequently contain much noise. Also, many land covers cannot be easily identified in most dates of a year. Instead, they show distinctive temporal patterns only during certain period of a year, which is also referred to as the discriminative period. To address these challenges, we propose a novel framework which combines the spatial context knowledge with the LSTM-based temporal modeling for land cover detection. Specifically, the framework learns the spatial context knowledge selectively from its neighboring locations. Then we propose two approaches for discriminative period detection based on multi-instance learning and local attention mechanism, respectively. Our evaluations in two real-world applications demonstrate the effectiveness of the proposed method in identifying land covers and detecting discriminative periods.

## 1 Introduction

The monitoring of Land Use and Land Cover (LULC) changes has drawn great attention from governments and companies for many years. Accurate accounting of LULC changes can provide promising insights needed for management of natural resources and help understand the impact of human actions and climate changes on the environment [10, 24]. For example, deforestation and forest fires in tropical regions lead to massive carbon emission. Monitoring crop varieties and planting area can help analyze the consumption of water and energy in tillage, irrigation and harvesting.

Effective monitoring of LULC changes requires the ability to precisely identify land covers over large regions and over long periods. Recent advances in storing and processing remote sensing data from satellites provide tremendous potential for land cover detection. These remote sensing datasets, such as MODIS and Landsat, contain multi-spectral features collected at a global scale under regular time interval, which makes it possible to detect large-scale land covers over long periods.



(a) (b) (c)

Figure 1: Example vegetation sequences of three burned locations. Fires occur in different time and frequency.

Most existing land cover products are manually created through visual interpretation, which takes advantage of human expertise in the labeling process [21, 23]. However, the limitations of this approach are manifold. First, manual labeling may result in both false positives and false negatives due to observational mistakes. Second, this approach usually requires multiple observers to delineate land covers. Their own subjective biases can result in inconsistent results. Most importantly, the required substantial human resources make it infeasible for large regions or for long periods.

In this paper, we investigate automated land cover detection using a data-driven approach. Compared with traditional classification problem, land cover detection is challenging for several reasons. First, the remote sensing data contain much noise, which is caused by either natural disturbances (e.g., clouds and aerosoles) or data acquisition errors by devices. Second, we cannot distinguish between many land covers in most single dates of a year. Instead, the successful detection requires the discovery of distinctive temporal patterns from a sequence of collected data. Third, these land covers show their distinctive temporal patterns only during certain period of a year, which is also referred to as the discriminative period. For example, croplands can be identified in certain part of growing season, but they look similar to barren land after they are harvested.

Moreover, these discriminative periods can shift across different locations and across different years. For example, we show the time series of vegetation index (i.e., higher value indicates higher greenness) for three burned locations in Fig. 1. We can observe similar patterns (a vegetation decrease followed by a recovery process) in different periods and even with different number of occurrences. Similarly, in cropland monitoring,

[*] University of Minnesota. {jiaxx221,khand035,nayak013, kumar001}@umn.edu

[†] University of Georgia. sheng.li@uga.edu

[‡] Virginia Tech. karpatne@vt.edu

farmers plant and harvest crops in different time across years due to weather conditions. Such heterogeneity can severely degrade the performance of traditional sequential models when they are tested at different locations or in different years [15]. Moreover, since we cannot easily identify crop types from crop residues, the periods before planting and after harvesting are less relevant to the classification and bring noise to the learning process.

To address these challenges in land cover detection, we propose a spatio-temporal framework, **C**ontext-aware **A**nalysis for **L**and covers with **D**iscriminative period (CALD). We utilize a Long-Short Term Memory (LSTM) model to capture long-term temporal dependencies in sequential remote sensing data and embed the raw data into a more representative feature space at each time step. After we embed the raw sequential input using LSTM, we combine each location with its neighboring locations to enhance the learning process. Since most land covers are contiguous over space, the incorporation of spatial context assists in mitigating the noise factors at individual locations. However, the locations in the spatial neighborhood may belong to different land covers (when located along the boundary), or contain much noise. To this end, we design an attention mechanism which selectively decreases/increases the weight on certain neighboring locations and then aggregates the information over the spatial neighborhood.

Next, we jointly utilize the information at each location and the obtained spatial context knowledge to detect discriminative period and conduct classification. For most land covers, the discriminative period usually persists for several consecutive time steps. For example, it takes several weeks for the land to recover from forest fires. The modeling of such persistence not only conforms to the underlying land cover patterns, but also reduces the impact from noise and outliers at certain time steps. To effectively capture the persistence and detect discriminative period, we propose two approaches, based on multi-instance learning (MIL) and local attention mechanism [20], respectively. While both approaches can be used to capture discriminative periods, they have different characteristics and can adapt to different application scenarios.

The detection of discriminative periods provides promising insights to explain classification results, which is of great interest to domain researchers. On the other hand, most existing land cover detection methods [4, 11] make classifications only after collecting data from the entire year. In contrast, our proposed framework can potentially identify land covers at an early stage after it detects the discriminative period and classify land covers with sufficient confidence.

We extensively evaluate the proposed method in cropland mapping and burned area detection. Cropland mapping is challenging for agricultural domain because different crop types look similar in most dates and are only distinguishable in certain periods of a year. It becomes more challenging to apply the learned model to different years due to the shift in planting and harvesting time. For the burned area detection, the training samples collected in the same year have different burning dates, which results in the heterogeneity within the training set. The results confirm that our proposed method outperforms multiple baselines in both tasks. In addition, we demonstrate that the proposed method can successfully detect the discriminative period and achieve reasonable accuracy in early-stage detection.

## 2 Problem Definition

In this work, we are provided with the data points at $N$ locations, $X = \{x_1, x_2, ..., x_N\}$. Each data point $x_i$ is a sequence of multi-variate spectral features with $T$ time steps, $x_i = \{x_i^1, ..., x_i^T\}$, where $x_i^t \in \mathbb{R}^M$. Also, we are provided with the labels of these sequential data, $Y = \{y_1, y_2, ..., y_N\}$. The label of each location belongs to one of $K$ land cover classes, i.e., $y_i \in \{1, 2, ..., K\}$.

Our objective is to train a classification model using the provided sequential data and labels. The learned model can then be applied to classify any test sequence into one of $K$ classes. Besides the classification, we also wish to find the most discriminative time period for each sample, which provides interpretability to the classification result.

## 3 Method

In this section, we first describe the LSTM model used to capture temporal dependencies and embed input features. Then we discuss the spatial context learning through an attention mechanism. Finally, we propose two approaches to detect discriminative periods from sequential data. We show the entire flow in Fig. 2.

**3.1 Sequential modeling by LSTM** Many land covers (e.g., croplands, plantations, forests, etc.) show temporal/seasonal changes over long periods. Also, the remote sensing data are influenced by long-term weather conditions. Therefore, we model the temporal relationships among different time steps using Long-Short Term Memory (LSTM). Compared with traditional RNN, LSTM can better memorize the temporal dependencies over a long period of time.

The input features $x^t$ at each time step $t$ contain multi-variate spectral features, which are collected from different spectral bandwidths. Recent studies show that combinations of multiple bandwidths can produce features that better distinguish between different land covers [29]. However, these combinations are manually designed using only a few bandwidths and do not take full advantage of multi-spectrum. On the other hand, the incorporation of temporal information can also help detect land covers. Consider the cropland mapping as an example. The identification of temporal patterns in crop growing seasons has turned out to be more effective for classification than using data from single dates [29].

Here we aim to automatically discover meaningful combinations of bandwidths from the multi-spectrum while also incorporating the temporal information. Hence, we utilize LSTM to embed $x^t$ at each time step into a hidden representation $h^t \in \mathbb{R}^H$, which encodes representative features for classification. In LSTM model, $h^t$ is generated through an LSTM cell with the input of multi-spectral features $x^t$ and the information from previous time steps.

Each LSTM cell contains a cell state $c_t$, which serves as a memory and allows reserving information from the past. Specifically, the LSTM first generates a candidate cell state $\tilde{c}_t$ by combining $x_t$ and $h_{t-1}$, as:

$$\tilde{c}^t = \tanh(W_h^c h^{t-1} + W_x^c x^t), \tag{3.1}$$

where $W_h^c \in \mathbb{R}^{H\times H}$ and $W_x^c \in \mathbb{R}^{H\times M}$ denote the weight parameters used to generate candidate cell state. Hereinafter we omit the bias terms as they can be absorbed into weight matrices. Then we generate a forget gate layer $f^t \in \mathbb{R}^H$, an input gate layer $g^t \in \mathbb{R}^H$ and an output gate layer $o^t$ using the sigmoid function:

$$\begin{aligned} f^t &= \sigma(W_h^f h^{t-1} + W_x^f x^t), \\ g^t &= \sigma(W_h^g h^{t-1} + W_x^g x^t), \\ o^t &= \sigma(W_h^o h^{t-1} + W_x^o x^t), \end{aligned} \tag{3.2}$$

where $\{W_h^f \in \mathbb{R}^{H\times H}, W_x^f \in \mathbb{R}^{H\times M}\}$ and $\{W_h^g \in \mathbb{R}^{H\times H}, W_x^g \in \mathbb{R}^{H\times M}\}$ denote two sets of weight parameters for generating forget gate layer $f^t$ and input gate layer $g^t$, respectively. The forget gate layer is used to filter the information inherited from $c^{t-1}$, and the input gate layer is used to filter the candidate cell state at time $t$. In this way we obtain the new cell state $c^t$ and the hidden representation as follows:

$$\begin{aligned} c^t &= f^t \otimes c^{t-1} + g^t \otimes \tilde{c}^t, \\ h^t &= o^t \otimes \tanh(c^t), \end{aligned} \tag{3.3}$$

where $\otimes$ denotes entry-wise product.

### 3.2 Spatial context learning

Since most land covers are contiguous over space, it is possible to learn spatial context knowledge from the spatial neighborhood to facilitate the detection process. This can also greatly help mitigate the noise at individual locations.

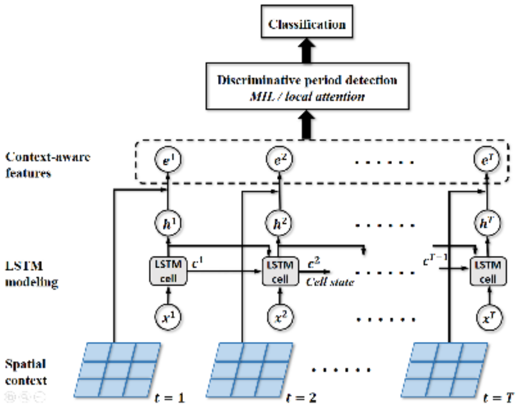


Figure 2: The flow chart of the CALD framework.

However, locations in the spatial neighborhood can misguide the detection in some situations. Considering a location along the boundary of a corn cropland, some of its neighboring locations may fall into other croplands (i.e., other classes) and therefore can mislead the classification. Also, if some neighboring locations are disturbed by natural factors or data acquisition errors, their multi-spectral features should not be trusted.

For these reasons, we propose to learn the spatial context knowledge selectively from neighboring locations through an attention mechanism. For each location, the attention mechanism measures the relevance between this location and each of its neighboring location. Then based on the relevance scores, the attention model aggregates the information from the entire spatial neighborhood. More formally, for each location $i$, we compute a relevance vector $\alpha_i^t$ at each time step $t$. Each entry $\alpha_{i,j}^t$ denotes the relevance score between location $i$ and its neighboring location $j$, and is computed as the similarity between the hidden representation of location $i$ and location $j$, as follows:

$$\alpha_{i,j}^t = \mathrm{softmax}({h_j^t}^{\mathrm{T}} h_i^t), \tag{3.4}$$

where the softmax function enforces that the relevance scores for all the neighboring locations sum to 1.

After obtaining the relevance vector $\alpha^t$ at time $t$, we compute a spatial context vector $s_i^t$ for location $i$ as the weighted summation of $h_j^t$ over the spatial neighborhood, as follows:

$$s_i^t = \sum_{j\in N(i)} \alpha_{i,j}^t h_j^t, \tag{3.5}$$

where $N(i)$ denotes the spatial neighborhood of location $i$. The range of spatial neighborhood can be adjusted for different applications. In this work, we set it to be a 1500m-by-1500m squared region centered at location $i$.

Then by combining the context vector $s_i^t$ and the hidden representation $h_i^t$, we generate a context-aware features $e_i^t \in \mathbb{R}^E$ for location $i$ at time $t$, as follows:

$$e_i^t = \tanh(W_h^e h_i^t + W_s^e s_i^t), \tag{3.6}$$

where $w_h^e \in \mathbb{R}^{E\times H}$ and $W_s^e \in \mathbb{R}^{E\times H}$ denote the weight parameters. Later we will utilize the sequence of $\{e_i^t\}_{t=1:T}$ as input for discriminative period detection and classification. Hereinafter we focus on the sequential modeling of each location independently. Hence, we omit subscript $i$ when it causes no ambiguity.

### 3.3 Discriminative Period Detection

We now investigate the detection of discriminative period using the obtained context-aware features $\{e^1, ..., e^T\}$. As the discriminative period in land covers usually persists over time, we propose two methods to model such temporal persistence and capture the discriminative period. In the first approach, we design a multi-instance learning (MIL) structure to aggregate multiple time steps into the final class label while also modeling the contribution of each time step. As an alternative approach, local temporal attention is utilized to detect discriminative period and conduct classification. Later we will discuss the difference between these two approaches.

#### 3.3.1 MIL aggregation

The multi-instance learning aims to map a bag of instances to a class label of the entire bag. In our work, we treat each time step as an instance in the bag (sequence). With the LSTM model described in Section 3.1, we establish the temporal dependencies between different instances. Then we wish to capture the persistence of discriminative period and aggregate all the time steps for classification.

Having obtained the context-aware features $e^t$ for $t = 1$ to $T$, we first generate a latent output $p^t$ at each time step as follows:

$$p^t = \sigma(W_p e^t), \tag{3.7}$$

where $W_p \in \mathbb{R}^{K\times E}$ denotes the weight parameters to transform $e^t$ into a $K$-dimensional output $p^t$. Each entry $p_k^t$ represents the intermediate classification measure for $k^{th}$ class using the collected information by time $t$.

To incorporate the persistence of discriminative periods, we consider $p^t$ values for several consecutive time steps to make classification decision. This can also greatly mitigate the ubiquitous noise and outliers (caused by natural factors or device errors) that occur at certain time steps.

In particular, if a location shows high $p_k^t$ values for consecutive $D$ time steps, then it is highly likely to belong to class $k$. Later we will describe the selection of $D$ value. For each class $k \in \{1, ..., K\}$, we take the maximum of the average $p_k^t$ value over consecutive $D$ time steps, as follows:

$$\tilde{y}_k = \max_t avg(p_k^t, p_k^{t+1}, ..., p_k^{t+D-1}), k = 1, ..., K, \tag{3.8}$$

where the larger $\tilde{y}_k$ requires the higher average value of $p_k^t$ for consecutive $D$ time steps corresponding to the discriminative period.

Then based on the collected $\tilde{y}_k$ for $k \in \{1, ..., K\}$, we adopt a softmax function to generate posterior probability for each class $k$:

$$P(\hat{y} = k|x) = \frac{exp(\tilde{y}_k)}{\sum_{k'} exp(\tilde{y}_{k'})}, \tag{3.9}$$

where we utilize $\hat{y}$ to distinguish the predicted label with the provided true label $y$.

With the predicted labels, we express the cost function by an entropy-based function, as follows:

$$\mathcal{J} = \sum_i \sum_k -\mathbb{1}(y_i = k) \log P(\hat{y}_i = k|x), \tag{3.10}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

The gradient of cost function with respect to $p_k^t$ can be computed as follows:

$$\frac{\partial \mathcal{J}}{\partial p_k^t} = \begin{cases} P(\hat{y} = k|x) - \mathbb{1}(y = k), & t \in [t_k^*, t_k^* + D - 1] \\ 0, & \text{otherwise}, \end{cases} \tag{3.11}$$
$$\text{where } t_k^* = \underset{t}{\operatorname{argmax}}\, avg(p_k^t, p_k^{t+1}, ..., p_k^{t+D-1}).$$

The gradient with respect to model parameters can be further derived by standard back-propagation algorithm. The time complexity is $O(NT(K+\eta))$, where the number of classes $K$ is a constant factor in our problem, and $\eta$ is a constant factor determined by the dimensionality of input features, hidden representation, context-aware features and the size of spatial neighborhood.

#### 3.3.2 Local temporal attention

An alternative solution is to utilize the local attention mechanism [20] to determine the discriminative period. Given the sequence of context-aware features $\{e^1, ..., e^T\}$, we aim to enforce the model to only attend to a period $[l, l+D-1]$, which indicates the discriminative period.

Specifically, the model first computes the starting time step through a sigmoid function, as:

$$\tilde{l} = (T - D + 1) \cdot \sigma(u^{\mathrm{T}} \tanh(W_l v)), \tag{3.12}$$

where $v \in \mathbb{R}^E$ represents an embedding of the entire sequence, which has the same dimensionality with context-aware features, and is jointly learned during the training process [17]. In the simplest case, we can embed $x^{1:T}$ into $v$ using another LSTM. $W_l \in \mathbb{R}^{U\times E}$ and $u \in \mathbb{R}^U$ are the weight parameters. The resulted $\tilde{l} \in (0, T - D + 1)$, and we select $l = \lceil \tilde{l} \rceil$.

We then measure the relevance score of each time step $t$ within the period $[l, l+D-1]$ according to the similarity between its context-aware features $e^t$ and the sequence embedding $v$. More formally, the relevance score of time step $t$ is computed via a softmax function over the time steps within $[l, l+D-1]$, as:

$$\beta^t = \text{softmax}(v^{\mathrm{T}} e^t). \tag{3.13}$$

Then we aggregate $e^t$ from all the time steps within the discriminative period $[l,\ l+D-1]$ based on the generated relevance $\beta$, and apply a softmax function for final classification:

$$\hat{y} = \text{softmax}(W_y \sum_{t \in [l, l+D-1]} \beta^t e^t), \tag{3.14}$$

where $W_y \in \mathbb{R}^{K \times E}$ denotes the parameters to transform context-aware features to the classification output. This approach has the same level of complexity with MIL-based approach (but depends on the constant $D$).

**3.3.3 Model discussion and Implementation details** Now we compare these two approaches. In Algorithm 1, the MIL-based approach detects the discriminative period using the latent outputs at each time step. On the other hand, the local attention-based approach selects the discriminative period using the sequence embedding $v$, which encodes the global information over the entire sequence. Such global information makes it more robust to noise and outliers than the MIL-based approach. Moreover, the local attention-based approach not only detects the discriminative period, but also models the contribution of each time step within this period by $\{\beta^t\}_{t=l:l+D-1}$.

However, if we are given the streaming remote sensing data, at a new time step $t$ we need to re-calculate the relevance $\beta^t$ in local attention model. In contrast, the MIL-based approach can be incrementally updated, which potentially leads to a real-time detection. Besides, the MIL approach captures the discriminative period for each class separately, which can better adapt to the scenarios with a diversity of land cover classes.

In this work, we utilize a validation set with $N_v$ locations to adjust the value of $D$. We train Artificial Neural Networks (ANN) separately at each time step, and measure the classification posterior probability $\{q_i^t = P(\hat{y}_i^t = y_i)\}_{t=1:T}$ for each location $i$ in the validation set. We select $D$ to be sufficiently large to cover the consecutive time windows with stronger discriminative signals than the remaining periods. More formally, we represent the distribution of $\{q_i^t\}_{t=1:T, i=1:N_v}$ as $Q$. Then we compute the average $q^{t=1:T}$ over all the locations, represented by $\bar{q}^{t=1:T}$. We set $D$ to be the maximum number of consecutive time windows, s.t. $\exists t'$, for $t = t'$ to $t'+D$, $\bar{q}_k^t$ is larger than 80 percentile of $Q$.

**3.4 Early-stage detection** By explicitly modeling the discriminative period, we wish to not only improve the detection, but also detect land covers at an early stage. Given the streaming remote sensing data for a location until time $t$, if the proposed framework has already captured the true discriminative period and classified the location with high confidence, we can make decision before collecting more data. Otherwise, the framework keeps collecting more remote sensing data until it detects the dicriminative period that leads to a confident classification. In this work, we utilize the posterior probability in Eqs. 3.9 and 3.14 to measure the classification confidence (more results in Section 4).

## 4 Experiments

In this section, we evaluate the proposed framework CALD in two real-world applications - cropland mapping and burned area detection.

To populate the input sequential features, we utilize MODIS MOD09A1 multi-spectral data product [2], collected by MODIS instruments onboard NASA's Terra satellites. This dataset provides global data for every 8 days at 500m spatial resolution. At each date, MODIS dataset provides reflectance values on 7 spectral bands for every location. To better learn short-term temporal patterns, we concatenate spectral features in every 32-days window as a time step and slide the window by 8 days. Totally we have 43 time steps in a year.

We implement two versions of our proposed framework - CALD$^{mil}$ and CALD$^{att}$ with MIL and local attention, respectively. We compare them to a diversity of baselines. The baselines include static approaches - Artificial Neural Networks (ANN) and Random Forest (RF) that are applied on the concatenation of sequential data, as well as widely used advanced sequential models - SVM$^{hmm}$ [5], 1-NN$^{dtw}$ (The nearest neighbor classifier with dynamic time warping distance), S2V [18, 27], and standard LSTM. We also compare to variants of CALD:

CALD with RNN (CALD$^{rnn}$): Rather than using LSTM, in this baseline we implement CALD$^{mil}$ with traditional RNN.

CALD without context (CALD$^{wc}$): Without using the spatial context knowledge, we directly take hidden representation $\{h^t\}_{t=1:T}$ as input to the MIL approach for discriminative period detection.

CALD without context (CALD$^{smt}$): In this baseline, we first implement CALD$^{wc}$ and then conduct an average smoothing over the spatial neighborhood.

### 4.1 Description of learning tasks and datasets

**4.1.1 Cropland Mapping** We aim to distinguish between corn and soybean in southwestern Minnesota,

Table 1: Performance(±standard deviation) of each method in cropland mapping (same-year/cross-year tests) and burned area detection in terms of AUC and F-1 score.

| | *Same-year cropland mapping* | | *Cross-year cropland mapping* | | *Burned area detection* | |
|---|---|---|---|---|---|---|
| **Method** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** |
| ANN | 0.717(±0.018) | 0.711(±0.011) | 0.660(±0.021) | 0.678(±0.015) | 0.805(±0.021) | 0.781(±0.020) |
| RF | 0.746(±0.013) | 0.733(±0.010) | 0.655(±0.014) | 0.678(±0.012) | 0.816(±0.014) | 0.793(±0.011) |
| SVM$^{hmm}$ | 0.753(±0.015) | 0.737(±0.011) | 0.721(±0.015) | 0.697(±0.013) | 0.833(±0.022) | 0.798(±0.016) |
| 1-NN$^{dtw}$ | 0.756(±0.029) | 0.750(±0.020) | 0.693(±0.031) | 0.695(±0.021) | 0.683(±0.025) | 0.670(±0.021) |
| S2V | 0.789(±0.038) | 0.758(±0.018) | 0.746(±0.038) | 0.712(±0.020) | 0.812(±0.028) | 0.815(±0.022) |
| LSTM | 0.804(±0.019) | 0.766(±0.009) | 0.760(±0.019) | 0.701(±0.010) | 0.928(±0.020) | 0.902(±0.019) |
| CALD$^{rnn}$ | 0.808(±0.030) | 0.763(±0.022) | 0.782(±0.031) | 0.749(±0.022) | 0.911(±0.021) | 0.907(±0.018) |
| CALD$^{wc}$ | 0.842(±0.024) | 0.790(±0.014) | 0.798(±0.023) | 0.740(±0.014) | 0.942(±0.024) | 0.932(±0.016) |
| CALD$^{smt}$ | 0.890(±0.023) | 0.835(±0.015) | 0.822(±0.023) | 0.753(±0.014) | 0.950(±0.020) | 0.935(±0.013) |
| CALD$^{mil}$ | 0.914(±0.019) | 0.855(±0.013) | 0.850(±0.017) | 0.801(±0.015) | 0.964(±0.018) | 0.960(±0.013) |
| CALD$^{att}$ | **0.931**(±0.012) | **0.866**(±0.016) | **0.856**(±0.013) | **0.809**(±0.009) | **0.975**(±0.018) | **0.977**(±0.014) |

US. In this test, we take 5,000 locations for each of corn class and soybean class in 2014 and 2016. The ground-truth information on these two classes is provided by USDA Crop Data Layer product [4]. This task is challenging in agricultural domain mainly for two reasons: 1) corn and soybean frequently look similar with each other in most dates of a year, and 2) each MODIS location is in 500 m spatial resolution and may contain multiple crop patches when located along the boundary, likely introducing noisy features.

We randomly select 40% locations and utilize their sequential features in 2016 as training data, and take another 10% as validation set. Then we conduct two groups of tests: 1) We test on the remaining locations in 2016 (same-year test), and 2) We conduct a cross-year test on the data acquired from 2014 using the learned models from 2016 (cross-year test). It is noteworthy that the planting time differs between these two years because of the weather conditions in Minnesota. In this test, the selected $D$ value is 5.

**4.1.2 Burned area detection** In this application, we randomly select 4,000 burned locations and another 4,000 unburned locations from Montana, US in 2007. We divide the data in the same proportion with our test for cropland mapping. We obtained fire validation data from government agencies responsible for monitoring and managing forests and wildfires [3]. In this application, the selected $D$ value is 8.

**4.2 Classification performance** We repeat the experiment with random initialization and random selection of training set. In Table 1, we report the performance for cropland mapping and burned area detection. It can be seen that the proposed framework outperforms all the baseline methods by a considerable margin. Among the baseline methods, we observe that the sequential baselines (SVM$^{hmm}$, 1-NN$^{dtw}$, S2V and LSTM) result in a better detection than static baselines (ANN and RF) because temporal patterns are helpful for identifying land covers. The exception is that 1-NN$^{dtw}$ performs poorly for burned area detection because dynamic time warping aligns sequences in an unsupervised way and can be easily disturbed by long irrelevant periods (before/after fires).

These sequential baselines do not perform as well as CALD since they do not take account of the discriminative periods. Consequently, they are negatively impacted by the irrelevant period in the year (e.g., after the harvesting/before fires) and the data heterogeneity across locations. For example, S2V focuses on mining frequent patterns, which can be noisy fluctuation or common patterns for both classes.

The improvement from CALD$^{rnn}$ to CALD$^{mil}$ shows that long-term dependencies are important for mining land cover patterns from multi-spectral sequence. By comparing CALD$^{wc}$, CALD$^{smt}$ and CALD$^{mil}$, we conclude that spatial context knowledge learned from the proposed attention mechanism can greatly improve the detection. In contrast, by equally weighting neighboring locations, the detection of CALD$^{smt}$ can be disturbed by neighboring locations along the boundary or with much noise.

**Cross-year performance in cropland mapping:** In Table 1, we also observe that the performance in 2016 (same-year test) is better than the performance in 2014 (cross-year test). This is mainly due to two reasons. First, the planting time of 2016 is in ahead of 2014, and thus a successful classification requires the method to automatically detect such a shift for discriminative period. Second, the collected multi-spectral features vary across years due to environmental variables, such as precipitation, sunlight, etc. Nevertheless, it can be seen that CALD$^{mil}$ and CALD$^{att}$ still produce a reasonable cross-year detection, which stems from their capacity in capturing discriminative periods.

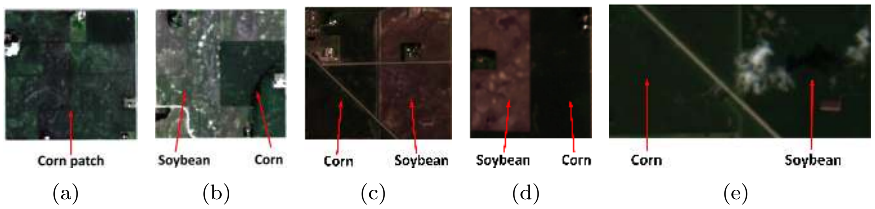


Figure 3: Sentinel-2 satellite images in RGB at 10m spatial resolution [1]. (a)-(d) Cropland patches with corn and soybean on Jun 23, 2016. Corn shows higher greenness level than soybean on this date. (e) Another cropland patch captured on Aug 06, 2016, where corn and soybean cannot be easity distinguished.

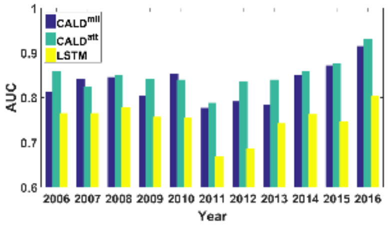


Figure 4: The performance of CALD$^{mil}$, CALD$^{att}$ and LSTM on cropland mapping from 2006 to 2016 using the training set in 2016.

Due to the importance of cross-year performance in agricultural domain, we report the cross-year performance on every year from 2006-2016 using the learned models (CALD$^{mil}$, CALD$^{att}$, and LSTM) from 2016 (Fig. 4). Compared with LSTM, CALD produces better prediction when applied to different years because they can automatically detect the shift of discriminative periods caused by environmental changes. This test demonstrates their robustness when tested on sequential data with variations to training data.

**4.3 Discriminative period detection** We now verify that CALD can indeed detect discriminative periods. **Cropland mapping:** In 2016, both CALD$^{mil}$ and CALD$^{att}$ detect the discriminative period for cropland mapping is from Jun 9 to July 11. To verify this, we show high-resolution Sentinel-2 images at 10m resolution. Fig. 3 (a)-(d) show some corn and soybean patches in four example regions using Sentinel-2 images on Jun 23, which show that corn patches turns into green faster than nearby soybean patches. Therefore they can be easily distinguished in this period.

The method CALD$^{mil}$ also detects the discriminative period (with second highest average $p^t$ values) from Jul 19 to Aug 20. During this period, both corn and soybean samples show very high greenness level and therefore it is difficult to distinguish between them by human (e.g., the Aug 06 Sentinel-2 image shown in Fig. 3 (e)). Here to verify that this period is indeed a discriminative period, we only use the multi-spectral features from Jul 19 to Aug 20 to train and test a simple ANN model, which produces AUC and F1-score of 0.829 and 0.778, respectively. It is noteworthy this is better than the ANN baseline that is trained using full-year sequences (Table 1). This improvement demonstrates that our framework has potential to detect the discriminative period from the full multi-spectrum, which can not even be observed directly by human experts.

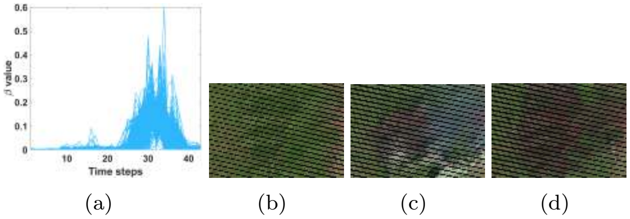


Figure 5: (a) The relevance scores $\beta$ for all the detected burned locations, and the Landsat satellite imagery (30m resolution) for a large burned region in Montana captured on (b) Jul 31, (c) Aug 09 and (d) Aug 25 in 2007. Stripes in imagery are caused by device errors.

**Burned area detection:** we also present the result for the discriminatve period detection using CALD$^{att}$. Fig. 5 (a) shows the relevance scores (i.e. $\beta$ value) of each time step within $[l, l + D - 1]$ for all the detected burned locations. According to the results, the discriminative period for most burned locations starts at 28$^{th}$ time step (∼Aug 5, 2007). In our test, CALD$^{mil}$ and CALD$^{att}$ result in the same discriminative period.

We validate this using Landsat images at 30m resolution (since Sentinel data are not available in 2007). In Fig. 5 (b), we show the largest burned area in Montana with the Landsat image taken on Jul 31. Here we can observe that fires have not occurred by Jul 31. In Fig. 5 (c), we show an image taken on Aug 09 for the same region, where we notice the burning and the smoke caused by fires. This conforms very well with our detection. Then in Fig. 5 (d) captured on Aug 25, we can see that fires left a burning scar in the region.

**4.4 Validation for spatial attention** Next, we demonstrate the effectiveness of the spatial attention mechanism by two examples, as shown in Fig. 6. The centered location in Fig. 6 (a) is along the boundary of a corn cropland patch. From Fig. 6 (b), it can be seen that the attention mechanism automatically reduces the weights for the neighboring locations that do not belong to the same corn patch. Besides, we show a different case in Fig. 6 (c) where the centered location has its neighboring locations also in the same cropland patch. However, as we scrutinize the original ground-truth data created at higher resolution (30m), we find these neighboring locations (in red box) are mixed pixels which contain around 40% area from other cropland patches. From Fig. 6 (d), we can observe that the attention mechanism can detect such mixed pixels and consequently reduce the impact of these noisy data.

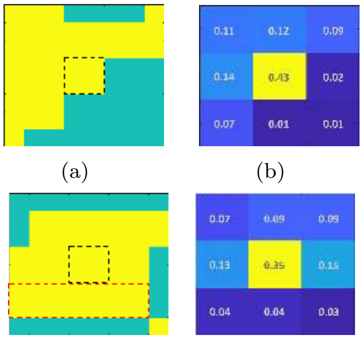
**4.5 Early detection** In Fig. 7 (a), we show the classification confidence (i.e., the posterior probability) by CALD$^{mil}$ and CALD$^{att}$ for the detected corn samples/burned area over time. In cropland mapping, we can observe that the confidence increases slowly during first 18 time steps. This is because crops have not been planted by this time and crop residues contain limited discriminative information. By around 27$^{th}$ time step, we observe that the learning model has gained enough confidence. Hence we can conduct early-stage detection by this time without waiting for the incoming data. In our test, if we make classification using the data by 27$^{th}$ time step, CALD$^{mil}$ and CALD$^{att}$ can reach AUC of 0.887 and 0.896 respectively, which are better than the performance of most baseline methods using complete sequences throughout the year.

For the burned area detection (Fig. 7 (b)), we observe that the confidence sharply increases around 27$^{th}$-28$^{th}$ time steps, which corresponds to the fires. We can see that the learning model gains high confidence (>0.92) at around 30$^{th}$ time step. If we stop collecting data by this time, we can identify burned area with AUC of 0.962 and 0.968 by CALD$^{mil}$ and CALD$^{att}$.

## 5 Related work

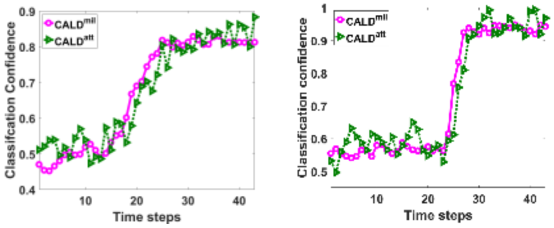
Most LULC detection methods focus on static satellite imagery at individual locations while ignoring their spatial context and temporal patterns [23, 11, 12]. However, many land covers cannot be captured using single-date snapshots, e.g., tree plantations and croplands.

With the development of deep learning, RNN modelhas been ubiquitously implemented. Due to the vanishing gradient problem [7], researchers have applied LSTM and Gated Recurrent Unit (GRU) to memorize long-term dependencies. These methods have shown great promise to land cover problems [9]. However, without explicitly modeling the discriminativ period, these methods suffer from the shift of discriminative periods and noise brought by irrelevant periods when tested at different locations or in different years.



(a) (b)
(c) (d)

Figure 6: (a)(c) Two example regions. Yellow color denotes corn patches and green color denotes soybean patches. The black box indicates the centered location and the red box indicates the mixed pixels. (b)(d) The obtained relevance scores using spatial attention mechanism for the centered locations in (a) and (c).



(a) (b)

Figure 7: The classification confidence of (a) corn vs. soybean over time in 2016, and (b) burned area vs. unburned area over time in 2007.

There are also existing efforts to explore important periods or patterns for sequence classification [26, 18, 16, 6]. Rather than detecting discriminative periods, these works aim to detect frequent patterns, which are likely to reflect common non-discriminative patterns shared among classes and also vulnerable to noise.

On the other hand, there exist many works on handling spatio-temporal data [19, 28, 14, 13]. These works utilize spatial information as constraints [8] or leverage spatial texture to facilitate the classification [25]. However, since they do not model the relevance of different neighboring locations, these approaches perform poorly for regions along the boundary or with strong noise. Fully Convolutional Networks (FCN) is also widely used for classifying each individual pixel/location [22]. However, these methods mostly focus on classifying static images. Moreover, the training process of FCN is extremely inefficient. In contrast, the CALD framework produces spatial context-aware features for each location separately, and thus avoids the heavy training process on the entire image. This is extremely helpful when we apply the method to a country level or global level.

## 6 Acknowledgement

This work was funded by the NSF Awards 1029711 and DTC seed grant. Access to computing facilities was provided by Minnesota Supercomputing Institute.

## 7 Conclusion

In this paper, we propose a framework CALD that combines spatial context and discriminative temporal information in land cover detection. The experimental results on two real-world applications demonstrate that the spatial context knowledge and the modeling of discriminative period can greatly improve the detection and also provide the interpretability to predictions. With the ability to identify land covers at an early stage, CALD can provide timely information for governments and companies to manage natural resources.

Given advances in remote sensing technology, CALD can contribute to many land cover problems. Also, the method can be applied to other important applications, such as heath-care data analysis where we can utilize the similarity relation among patients and the progression patterns of diseases to make prediction.

## References

[1] Land viewer - eos data analytics. https://eos.com/landviewer/.

[2] Modis product table. https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod09a1.

[3] Monitoring trends in burn severity. https://www.mtbs.gov/.

[4] Usda national agricultural statistics service cropland data layer. https://nassgeodata.gmu.edu/CropScape/.

[5] Yasemin Altun et al. Hidden markov support vector machines. In *ICML*, 2003.

[6] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. In *SIGKDD*. ACM, 2012.

[7] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *ITNN*, 1994.

[8] Yi Chen, Nasser M Nasrabadi, and Trac D Tran. Hyperspectral image classification using dictionary-based sparse representation. *TGRS*, 2011.

[9] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *J-STARS*, 2014.

[10] Nancy B Grimm et al. Global change and the ecology of cities. *science*, 2008.

[11] Matthew C Hansen et al. High-resolution global maps of 21st-century forest cover change. *Science*, 2013.

[12] Xiaowei Jia, Ankush Khandelwal, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Learning large-scale plantation mapping from imperfect annotators. In *IEEE BigData*, 2016.

[13] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Incremental dual-memory lstm in land cover prediction. In *SIGKDD*. ACM, 2017.

[14] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. Predict land covers with transition modeling and incremental learning. In *SDM*, 2017.

[15] Anuj Karpatne et al. Monitoring land-cover changes: A machine-learning perspective. *IEEE Geoscience and Remote Sensing Magazine*, 2016.

[16] Jae-Gil Lee, Jiawei Han, Xiaolei Li, and Hong Cheng. Mining discriminative patterns for classifying trajectories on road networks. *TKDE*, 2011.

[17] Huayu Li, Martin Renqiang Min, Yong Ge, and Asim Kadav. A context-aware attention network for interactive question answering. In *SIGKDD*. ACM, 2017.

[18] Xiaoyi Li, Xiaowei Jia, Guangxu Xun, and Aidong Zhang. Improving eeg feature learning via synchronized facial video. In *IEEE Big Data*. IEEE, 2015.

[19] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI*, 2016.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[21] Jukka Miettinen, Chenghua Shi, Wee Juan Tan, and Soo Chin Liew. 2010 land cover map of insular southeast asia in 250-m spatial resolution. *RSL*, 2012.

[22] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[23] RACHAEL Petersen et al. Mapping tree plantations with multispectral imagery: preliminary results for seven tropical countries. *WRI*, 2016.

[24] Roger A Pielke. Land use and climate change. In *Science*, 2005.

[25] Grant J Scott et al. Training deep convolutional neural networks for land–cover classification of high-resolution imagery. *IEEE GRSL*, 2017.

[26] Jianyong Wang and George Karypis. Harmony: Efficiently mining the best rules for classification. In *SDM*, 2005.

[27] Guangxu Xun, Xiaowei Jia, and Aidong Zhang. Detecting epileptic seizures with electroencephalogram via a context-learning model. *BMC MIDM*, 2016.

[28] Huaxiu Yao et al. Deep multi-view spatial-temporal network for taxi demand prediction. *arXiv preprint:1802.08714*, 2018.

[29] Linglin Zeng et al. A hybrid approach for detecting corn and soybean phenology with time-series modis data. *RSE*, 2016.