Context-aware Deep Representation Learning for Geo-spatiotemporal Analysis

Hanzi Mao*§, Xi Liu†, Nick Duffield†, Hao Yuan†, Shuiwang Ji†, Binayak Mohanty†
*Facebook AI
†Texas A&M University, College Station, TX
hannamao@fb.com, xiliu.tamu@gmail.com, {duffieldng, hao.yuan, sji, bmohanty}@tamu.edu

***Abstract*—The emergence of remote sensing technologies coupled with local monitoring workstations enables us the unprecedented ability to monitor the environment in large scale. Information mining from multi-channel geo-spatiotemporal data however poses great challenges to many computational sustainability applications. Most existing approaches adopt various dimensionality reduction techniques without fully taking advantage of the spatiotemporal nature of the data. In addition, the lack of labeled training data raises another challenge for modeling such data. In this work, we propose a novel semi-supervised attention-based deep representation model that learns context-aware spatiotemporal representations for prediction tasks. A combination of convolutional neural networks with a hybrid attention mechanism is adopted to extract spatial and temporal variations in the geo-spatiotemporal data. Recognizing the importance of capturing more complete temporal dependencies, we propose the hybrid attention mechanism which integrates a learnable global query into the classic self-attention mechanism. To overcome the data scarcity issue, sampled spatial and temporal context that naturally reside in the largely-available unlabeled geo-spatiotemporal data are exploited to aid meaningful representation learning. We conduct experiments on a large-scale real-world crop yield prediction task. The results show that our methods significantly outperforms existing state-of-the-art yield prediction methods, especially under the stress of training data scarcity.**

***Index Terms*—Spatiotemporal Prediction, Semi-supervised Learning, Attention, Crop Yield Prediction**

I. Introduction

Recent years have seen a proliferation of studies concerning computational approaches that exploit geo-spatiotemporal environmental monitoring data for sustainability applications. These applications include crop yield prediction [1], soil moisture downscaling [2], land cover classification [3], wildfire prediction [4], and climate modeling [5]. The input data for these applications are either multi-channel data from a single source (e.g. multispectral remote sensing images) or conflated geodata fused from various sources. Combining multi-feature spatial data in time series results in a high-dimensional dataset. It comprises the data features associated with each represented point in the four-dimensional product space of temporal and spatial coordinates. As an example, Figure 1 shows the available input of a midwestern county in the U.S. for the county-level crop-yield prediction task, which consists of monthly plant growth estimates from remote sensing satellites and climate data from various sources collected during the growing season.

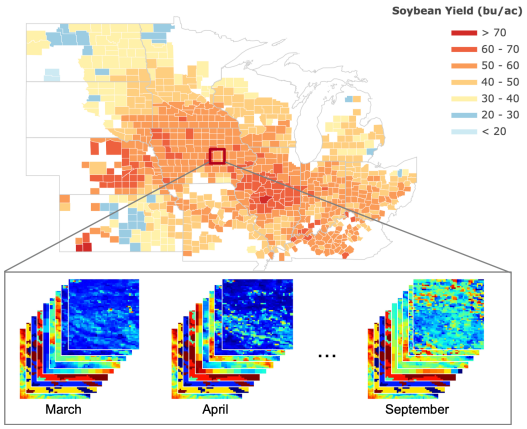

Fig. 1. Up: County-level soybean yield of mid-western U.S. in 2017. Down: At county Tama, Iowa, each channel of the monthly spatial images represents one of the nine features. Features include monthly plant growth estimates (including vegetation indices) and climate data (including land surface temperature, precipitation, soil properties and elevation).

Extracting information from the high-dimensional data remains a challenge. Most previous studies have avoided dealing with the full spatiotemporal data directly through various dimensionality reduction techniques. The most common approach is to spatially average the input resulting in time series scalar features [6]. In [1], spatial images are converted to histograms with the assumption that spatial distributions of the features do not contribute to the accuracy of the predictions. These approaches ignore the spatially explicit heterogeneity and interactions existing among the fused environmental features. The ignored subgrid heterogeneous dynamics, however, play an important role in many sustainability applications, such as crop yield prediction [7] and soil moisture downscaling [2]. In addition, there are studies that generate predictions from a one-time snapshot of the geo-spatiotemporal data [8], [9] without taking advantage of the temporal information.

Another challenge that commonly exists in the applications of spatiotemporal geographical data is the lack of labels [3]. Many computational sustainability tasks rely on label collection procedures that are either expensive or cumbersome [10]. The scarcity of labeled training data becomes an even severe issue for deep learning models that deal with high-

§This work was done while the author was at Texas A&M University.

dimensional data [11].

To fully taking advantage of the spatiotemporal variations contained in the data, we propose in this paper a novel semi-supervised attention-based model that *jointly* learns a prediction function and a spatiotemporal representation function. A hierarchical framework combining the convolutional neural networks with a proposed hybrid attention mechanism is adopted to handle the high-dimensional data. At each time step, spatial information contained in the multi-channel three-dimensional images is first extracted to a latent space as representations through convolutional neural networks. A hybrid attention model is then applied to temporally aggregate the spatial representations at all time steps to generate spatiotemporal representations, from which final predictions are produced. The hybrid attention mechanism is proposed based on the observation that many computational sustainability tasks show task/data-specific temporal variation patterns. It is thus of great importance to capture more complete temporal dependencies among the sequential spatial representations. To this end, we propose the hybrid attention mechanism, where a learnable global query capturing temporal dependencies for all training examples [12] is introduced to the self-attention mechanism [13], resulting in more effective knowledge transferring among different time steps [14].

To overcome the scarcity of training data, we further introduce the spatial and temporal coherence signals that naturally reside in the largely-available unlabeled geo-spatiotemporal data through a novel sampling procedure. Similar to the proximity-based word embedding models in natural language, we make an assumption that images that are spatially or temporally close should have similar representation, and conversely. Under our semi-supervised training framework, this newly introduced spatiotemporal context can aid meaningful representation learning that adapts to the supervised prediction task at the same time.

We evaluate our approach through a large-scale real-world crop yield prediction task. Experimental results show that our semi-supervised hybrid attention model outperforms existing state-of-the-art crop yield prediction methods and its counterparts, the supervised-only hybrid attention model and semi-supervised self-attention model, especially when there are less labeled training data. The contributions of this work are summarized as follows:

- We propose a novel semi-supervised hybrid attention model which learns spatiotemporal representations for prediction tasks. This model takes full advantage of the multi-channel geo-spatiotemporal data with both spatial and temporal variations captured.
- We propose a hybrid attention mechanism where a trainable global query is introduced to the classic self-attention mechanism. The hybrid attention mechanism captures more complete temporal dependencies that adapt to specific learning task.
- We aid the representation learning with spatial and temporal context sampled from largely-available unlabeled geo-spatiotemporal data. Under the semi-supervised training framework, context-aware representations that adapt to the supervised prediction task are learned.
- We conduct extensive experiments on a challenging large-scale real-world crop yield prediction task. Experimental results demonstrate the effectiveness of our semi-supervised hybrid attention model over existing state-of-the-art yield prediction methods, especially under the stress of scarce labeled data.

II. Related Work

With the emergence of environmental monitoring in the past decade, much progress has been made in many computational sustainability tasks through the use of geo-spatiotemporal data [15]. The most common approach to prediction in this area, either as regression or classification, is to incorporate multiple features from various sources with the consideration of domain knowledge. Specifically, much research [2], [6], [16] take time series scalar features as input to generate predictions for ground-truth labels within a predefined spatial unit. Input features with higher resolutions than the labels are averaged without considering the subgrid heterogeneous dynamics. However, it has been studied and well acknowledged in many environmental sustainability domains that the spatially explicit dynamics and interactions among environmental factors play a great role determining the characteristics of environmental factors at larger spatial scales [7], [17]. Ignoring these subgrid spatial heterogeneity results in unnecessary information loss. In [1], multispectral remote sensing images with spatial distributions are converted to histograms with the assumption of permutation invariance. Their assumption is less appropriate when conflated geodata with spatially explicit dynamics and interactions are used as inputs. In addition, several studies only incorporates static spatially distributed data without considering the temporal patterns in the geo-spatiotemporal data, such as land cover prediction [18], poverty mapping [3]. There is however valuable information contained in the temporal variations that can aid the predictions [7], [16].

The temporal attention-based representation learning model proposed in this paper, instead, adopts a hierarchical framework where spatial representations at each time step are learned first, followed by a temporal aggregation through a hybrid attention model. It extracts both spatial and temporal variations in the data with spatiotemporal representations learned for the prediction task.

Recurrent convolutional networks (CNN-LSTM) [19] shares a similar architecture with our proposed model, but processes latent output at each time step in a temporally explicit order. While the self-attention model has demonstrated its superiority in capturing global information and handling sequential data at longer lengths in many fields [20], [21], there are few studies to demonstrate its ability to mine spatiotemporal patterns from geographical data. Geo-spatiotemporal data normally show distinct but relatively consistent temporal patterns [7]. The knowledge transferring among different time steps enabled by the self-attention model can aid the generation of more meaningful aggregated representations. The hybrid attention

model we proposed in this paper introduces a global query to the self-attention model resulting in more effective temporal pattern extractions. The 3d convolutional neural networks (C3D) [22] also aims at spatiotemporal feature learning. They apply 3d convolutions across both the spatial and temporal dimensions.

Another remaining challenge for learning spatiotemporal environmental monitoring data is the lack of labels. In addition to dimensionality reduction [1], data augmentation [18] and transfer learning [23] that have been proposed to alleviate this issue, there have been attempts to introduce spatial context to aid model learning [3], [24]. In [24], weighted representations of all spatial neighbors within a predefined region are considered. However, their settings are not practical when dealing with high-dimensional spatiotemporal data. An existing study [3] introduces the spatial context through sampling from the neighboring region to aid their unsupervised representation learning, and learn representations independently from downstream tasks. In this research, we sample spatial and temporal context from the largely-available unlabeled data. Information from the sampled context is learned together with a supervised prediction task under the semi-supervised training framework. Context-aware representations that adapt to the prediction task at the same time can thus be learned.

III. Method

A. Problem Definition

In this section, we present the formulation for the geo-spatiotemporal unit-wise prediction task. Specifically, given the multi-channel geo-spatiotemporal data as input, the goal is to predict labels for each geographical unit. Given a unit i, there are T different multi-channel spatial images centered at the unit. Different images represent the geo-spatiotemporal data at different time steps. We denote the time series spatial images as $\mathbf{A}_i = (\mathbf{A}_i^1, ..., \mathbf{A}_i^T)$ where $\mathbf{A}_i^t \in \mathbb{R}^{h \times w \times d}$ and $\mathbf{A}_i \in \mathbb{R}^{T \times h \times w \times d}$. Note that h and w denote the height and width of the image, and d is the number of features incorporated from either a single dataset or a conflated dataset fused from various sources. To utilize the spatial and temporal context information, for each unit i, we obtain three context-aware images, known as the spatial neighbor $\mathbf{SN}_i$, spatial distant $\mathbf{SD}_i$, and temporal neighbor $\mathbf{TN}_i$ through a novel sampling procedure, which is introduced in Section III-C. These context-aware images share the same dimensions $T \times h \times w \times d$ as $\mathbf{A}_i$.

Our objective is to learn a predictive model to predict the label for unit i from the time series spatial image quadruplets. Formally, it can be written as

$$y_i = \mathcal{F}\big(\{\mathbf{A}_i^1, \mathbf{SN}_i^1, \mathbf{SD}_i^1, \mathbf{TN}_i^1\}, \cdots, \{\mathbf{A}_i^T, \mathbf{SN}_i^T, \mathbf{SD}_i^T, \mathbf{TN}_i^T\}\big), \tag{1}$$

where y_i denotes the prediction for the geographical unit i. Hereinafter we omit the subscript index i when it causes no ambiguity.

B. Framework Overview

The architecture of the proposed framework is presented in Figure 2. The framework consists of three parts. First, at each time step $t \in (1, ..., T)$, we employ convolutional neural networks (CNNs) to extract its spatial representations from the image quadruplets $\{\mathbf{A}^t, \mathbf{SN}^t, \mathbf{SD}^t, \mathbf{TN}^t\}$. Since CNNs have demonstrated its superior ability to extract spatial information in computer vision field [25], we adopt CNNs to extract spatially explicit dynamics and interactions existing among the fused environmental features. Specifically, we use a modified ResNet-18 architecture [26] where the final classification layer is removed. Formally, for each image in the quadruplets, the CNNs map it to an m-dimensional embedding, denoted as $g(\cdot) \in \mathbb{R}^m$. Then we denote the resulted context-aware spatial representations at time step t as $\{g(\mathbf{A}^t), g(\mathbf{SN}^t), g(\mathbf{SD}^t), g(\mathbf{TN}^t)\}$. Note that for all different time steps and images, the same CNNs are shared. Second, a hybrid attention model is applied to temporally aggregate the spatial representations $(g(\mathbf{A}^1), ..., g(\mathbf{A}^T))$ at the labeled regions. Spatiotemporal representations that capture both the spatial and temporal variations are produced from the hybrid attention model. Finally, a prediction layer composed of a fully-connected neural network is deployed to generate predictions from the extracted spatiotemporal representations.

The objective function of our model is composed of four parts:

$$L := L_S + \alpha(L_{US} + \beta L_{UT} + \gamma L_R), \tag{2}$$

where L_S denotes the supervised loss for the prediction task. L_{US} and L_{UT} denote unsupervised losses for spatial and temporal context respectively. L_R is added here to regularize the learned representations to a meaningful hypersphere. α, β, γ are trade-off weights. We first explain how we produce spatial and temporal context-aware representations through unsupervised learning and a novel sampling procedure in Section III-C. Descriptions of the hybrid attention model for generating spatiotemporal representations are presented in Section III-D.

C. Context-aware Representation Learning

The widely explored proximity-based word embedding models [27] assume that "a word is characterized by the company it keeps"; thus words that appear in similar contexts should have similar representations. Extending this idea to our spatiotemporal distributed geographical data, we make an assumption that spatial images that are close spatially or temporally should have similar representations than those that are far apart.

To constrain the closeness in spatiotemporal space, we introduce the concepts of *spatial neighborhood* and *temporal neighborhood*. First, we denote the regions where labels are obtained and predictions ought to be produced as anchor regions. The spatial neighborhood is then defined as a larger spatial region that is within a predefined spatial distance of an anchor region and appears at the same time step. An example of a spatial neighborhood region is shown in Figure 3. The

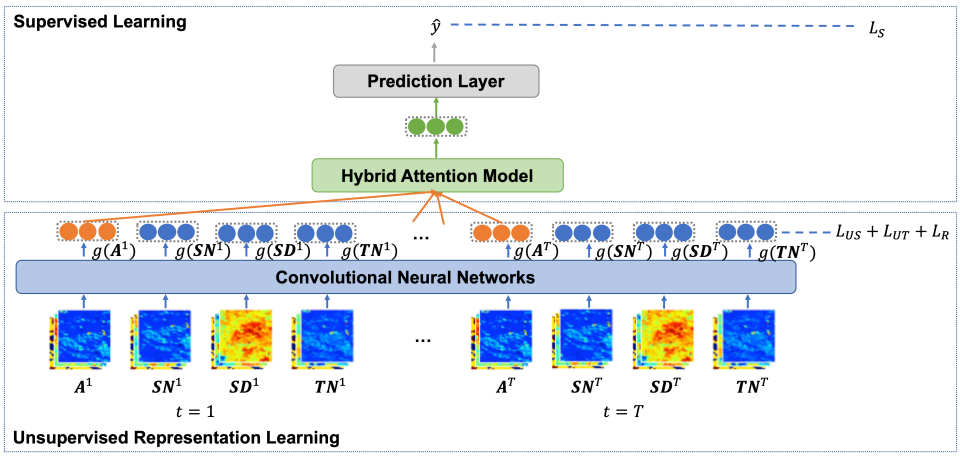

Fig. 2. Architecture overview of the semi-supervised context-aware attentive representation learning model.

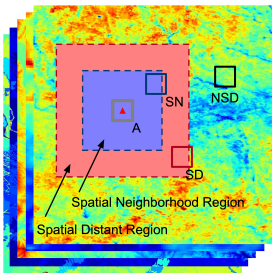

Fig. 3. An example of spatial neighborhood region and spatial distant region corresponding to an anchor image $\mathbf{A}$. Spatial neighbor $\mathbf{SN}$ is sampled from the spatial neighborhood region. Spatial distant $\mathbf{SD}$ is sampled from the spatial distant region. A conflated geodata that fuse various environmental features, such as vegetation indices, temperature, soil properties, etc., are plotted in background.

temporal neighborhood, instead, is defined as the same spatial region of the anchor region but appears at a time step within a predefined temporal distance. Spatial neighbor image $\mathbf{SN}$, spatial distant image $\mathbf{SD}$, and temporal neighbor image $\mathbf{TN}$ can be obtained in the context of the spatial neighborhood and temporal neighborhood through a sampling procedure, which will be introduced in Section III-C2.

1) Learning Objective: Now suppose that we have time series spatial representations $\{g(\mathbf{A}^t), g(\mathbf{SN}^t), g(\mathbf{SD}^t), g(\mathbf{TN}^t)\}$ generated from the CNNs with $\{\mathbf{A}^t, \mathbf{SN}^t, \mathbf{SD}^t, \mathbf{TN}^t\}$ as input. At each time step $t \in (1 \cdots T)$, we seek to minimize the Euclidean distance between the representation vectors of the anchor image and spatial/temporal neighbor image, while maximize the distance between the representation vectors of the anchor image and spatial distant image. The unsupervised loss for the spatial and temporal context can be calculated by

$$L_{US}^t = max(0, ||g(\mathbf{A}^t) - g(\mathbf{SN}^t)||^2 - ||g(\mathbf{A}^t) - g(\mathbf{SD}^t)||^2 + p) \tag{3}$$

and

$$L_{UT}^t = ||g(\mathbf{A}^t) - g(\mathbf{TN}^t)||^2, \tag{4}$$

respectively. Following existing work [3], a rectifier with margin p is introduced here to control the extent how the representations of the spatial distant image are pushed away compared to the representation of the spatial neighbor image.

To constrain the learned embeddings within a hypersphere where better representations with meaningful relative distance can be learned, we further introduce a L2 regularization with loss L_R^t:

$$L_R^t = ||g(\mathbf{A}^t)||^2 + ||g(\mathbf{SN}^t)||^2 + ||g(\mathbf{SD}^t)||^2 + ||g(\mathbf{TN}^t)||^2. \tag{5}$$

Finally, for given a dataset of N geographical regions with T time steps, the unsupervised loss is given as

$$\min_{\theta} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} (L_{US(i)}^t + \beta L_{UT(i)}^t + \gamma \frac{L_{R(i)}^t}{\sqrt{m}}) \tag{6}$$

where m is the dimension of the embedding vectors and θ is the parameters of the CNNs. This unsupervised loss is jointly trained with the supervised prediction task to generate context-aware representations that are suitable for the prediction task.

2) Quadruplet Sampling: We adopt the following procedure to generate times-series image quadruplets with spatial and temporal context. First, anchor images $\mathbf{A}$ are collected as images that cover the area of interest with fixed image size. Second, spatial neighbor $\mathbf{SN}$ and spatial distant $\mathbf{SD}$ images are sampled with respect to the anchor images based on spatial distance. We adopt a similar sampling procedure as in [3] with the spatial neighborhood introduced to constrain the sampling of spatial neighbor images. Specifically, the center of the spatial neighbor images must be within a predefined number of pixels of the anchor image center both vertically and horizontally. Different from the pure-unsupervised learning in [3] where the size of spatial neighborhood is relatively more flexible, we find in our setting that the size of spatial neighborhood should adapt to the supervised task. In practice, we find that choosing the spatial neighborhood similar to the size of the anchor images produces better supervised predictions.

As computational sustainability tasks normally conduct experiments in large-scale, e.g. the county-level crop yield prediction across the United States [1], we further constrain the distant region from which the spatial distant images can be sampled. As shown in Figure 3, in addition to the spatial neighborhood region colored in blue, there is a spatial distant region colored in red around the anchor image $\mathbf{A}$. The image $\mathbf{NSD}$ in the black rectangle, while is also far away from $\mathbf{A}$ as the sampled distant image $\mathbf{SD}$, will not be considered as spatial distant in this setting. This is similar to the "hard negative" idea introduced to the temporal embedding learning in [28]. We find in practice that this newly added constraint help us learn better representations with the prediction accuracy improved. Finally, the temporal neighbor images are sampled from a fixed temporal window as temporal context for each anchor image. The sampling procedure is applied at all time steps separately to enrich the spatial and temporal proximity signals the model sees.

D. Hybrid Attention Model

We present here a hybrid attention model used for the generation of spatiotemporal representations. As output from the CNNs, we have the time series context-aware spatial representations $(g(\mathbf{A}^1), g(\mathbf{A}^2), \cdots, g(\mathbf{A}^T))$ for the labeled regions. As these representations are generated independently at each time step, there is no order information learned to aid networks in later stage understanding the relative distance of these embeddings. To incorporate order of sequence information, we add "positional embedding" [13] to the spatial representations through position-wise summation. Specifically, we use the positional encoding in a sinusoidal form. The encoding added to a spatial representation at time step t is given as:

$$\begin{aligned} PE(t, 2i) &= sin(t/10000^{2i/m}) \\ PE(t, 2i+1) &= cos(t/10000^{2i/m}) \end{aligned} \tag{7}$$

where m is the length of the representation embedding and $i \in [0, m/2]$. The added parameter-free position-dependent embeddings can help followed networks incorporate the temporal order of the spatial representations without incurring extra computation burden.

After adding the positional encoding, we first reformulate the time series position-dependent spatial representations in a matrix form $\mathbf{Z} \in \mathbb{R}^{T \times m}$. Following the self-attention model [13], this representation matrix is then linearly projected to three matrices, queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$), independently. The three projected matrices share the same dimensions as $\mathbf{Z}$. Note that as $\mathbf{Z}$ contains the time-series spatial representations for each training example separately, the linearly projected query matrix $\mathbf{Q}$ also serves as high level representations of queries "how other time steps should attend to this time step" for time steps of each specific training example.

Recognizing the importance of capturing more complete temporal dependencies in analyzing geo-spatiotemporal data, we propose to introduce a global trainable query $\mathbf{G}$ with the full hybrid query matrix being expressed as

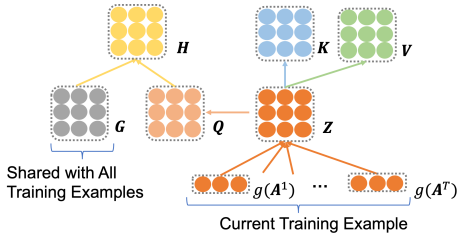

Fig. 4. Conceptual graph showing the generations of hybrid query matrix $\mathbf{H}$, key matrix $\mathbf{K}$ and value matrix $\mathbf{V}$ for the hybrid attention model. For simplicity, positional encoding is omitted here.

$$\mathbf{H} = \mathbf{Q} + \mathbf{G}, \tag{8}$$

as shown in Figure 4. $\mathbf{G} \in \mathbb{R}^{T \times m}$ denotes the query representations for each time step separately and is shared by all training examples. $\mathbf{G}$ is jointly learned during the training process from randomly initialized values.

To empower learning in different representation subspaces, we further split the hybrid query matrix $\mathbf{H}$, key matrix $\mathbf{K}$ and value matrix $\mathbf{V}$ into h parts and attended by h parallel heads separately. For each head $i \in h$, we have its attention score calculated through the scaled dot-product attention:

$$\mathbf{S}_i = \text{softmax}(\frac{\mathbf{H}_i \mathbf{K}_i^T}{\sqrt{m/h}}), \tag{9}$$

where $\mathbf{S}_i \in \mathbb{R}^{T \times T}$ is the attention score matrix which denotes how information contained in representation subspace $\mathbf{V}_i$ can transfer to each other. The scale factor $1/\sqrt{m/h}$ is added to prevent the gradient vanishing problem [29]. The updated spatial representations of i-th head is then obtained by multiplying the attention score matrix with the values matrix $\mathbf{V}_i$:

$$\mathbf{Z}_i^u = \mathbf{S}_i \mathbf{V}_i \tag{10}$$

where $\mathbf{Z}_i^u \in \mathbb{R}^{T \times m/h}$ is the refined subspace representations. With this update, spatial representation $(\mathbf{z}_i^t)^u \in \mathbb{R}^{m/h}$ at a time step t can incorporate information contained in other times steps resulting in meaningful knowledge transferring.

To generate a unifying representation from all subspaces, the updated matrices from h parallel heads are concatenated and once again projected to have:

$$\mathbf{Z}^u = concat(\mathbf{Z}_1, \cdots, \mathbf{Z}_h)\mathbf{W} \tag{11}$$

where $\mathbf{Z}^u \in \mathbb{R}^{T \times m}$ and $\mathbf{W} \in \mathbb{R}^{m \times m}$ is a projection matrix for the concatenated unifying matrix. Following [13], we use a residual connection [26] to add the original spatial representations to the refined representations to enable the propagation of useful features learned at low-level to deeper levels:

$$\mathbf{Z}^u = \mathbf{Z}^u + \mathbf{Z}. \tag{12}$$

To further increase the length of the representations, and thus improve the model's expressive capability [30], we add

position-wise feed-forward network as in [13] to process the updated representations position-wisely. Finally, the refined spatial representations at all time steps are temporally averaged through a pooling layer to generate a global spatiotemporal representation, after which a fully-connected neural network is deployed to generate the final prediction. For a regression task, as an example, the supervised loss for the predictions can be calculated through the mean squared error:

$$L_S = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{13}$$

where $\hat{y}_i$ denotes the prediction from the model for region i. We find in practice that this temporal hybrid attention model is easy to train and less prone to overfitting compared with previously studied models for geo-spatiotemporal data, especially when the training data is scarce. It becomes even more powerful when accompanied with the context-aware unsupervised representation learning.

IV. Experiments

To validate the effectiveness of our proposed approaches, we conduct experiments over a large-scale real-world dataset for the crop yield prediction task. The dataset consist of conflated geospatiotemporal data from various sources. In this section, we first introduce the dataset, including the crop yield dataset and the input datasets. Then we describe the evaluation metrics and baseline methods. Results of model performance are presented last. Codes for the models and data pre-processing are available online[1].

A. Dataset

County-level soybean yield prediction has been an important task and actively researched in previous studies [1], [31]. The ground-truth of the task is average county-level soybean yields harvested in every October and November. We collect the data from the USDA National Agricultural Statistics Service (NASS) Quick Stats Database [32] for years between 2003 to 2018. 13 states of the midwestern United States are selected which account for over 80% national soybean production. There are around 850 data points per year.

We fuse plant growth estimates from remote sensing and environmental factors from various sources as inputs. Specifically, a pair of monthly vegetation indices, the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), are collected from the MODIS satellite product MOD13A3 [33] at 1 km resolution. These two vegetation indices complement each other and have been widely used to monitor plant growth in previous studies [6]. For environmental factors, we consider precipitation and surface temperature to reflect the water and heat stress, to which the crop growing processes were observed to be highly sensitive [34]. Monthly precipitation data at 4 km resolution are from the Parameter-elevation Relationships on Independent Slopes Model (PRISM) dataset where precipitation data are derived from nearly 13,000 stations using climate-based interpolation [35]. Daily surface temperature data during the day and night are collected from the 1 km MODIS satellite product MOD11A1 [36]. They are aggregated to a monthly time step through taking averaged. Additionally, We introduce geographically localized and time-invariant factors such as elevation and soil properties, e.g., soil sand, silt, and clay fractions. They are also important environmental factors to be considered, as they determine how stresses like water and heat influence crop growth. The elevation data is obtained from the NASA Shuttle Radar Topography Mission Global 30 m product [37] while the soil properties data are collected from 1 km Soil Geographic (STATSGO) Data Base [38]. We map all the inputs from various sources to the MODIS product grid at 1 km resolution through averaging or nearest-neighbor search.

Once the data preprocessing steps are finished, anchor images are cropped as 50×50 pixels centering at the counties which have soybean yields, considering that the average size of counties in the U.S. is around 2500 km^2. MODIS landcover product MCD12Q1 is further introduced in this step to decide the center of croplands at each county. Fixing the image size as 50×50 for all the counties might not be optimal but is practical as we are conducting the prediction task in large scale. Spatial and temporal contextual images are sampled corresponding to the anchor image with the same spatial size. We collect before-harvest input data from March to September, resulting in a feature space $\mathbf{X} \in \{\mathbf{A}, \mathbf{SN}, \mathbf{SD}, \mathbf{TN}\}$with four dimensions, $\mathbf{X} \in \mathbb{R}^{7\times50\times50\times9}$. The number of time steps is 7, and the number of features is 9.

Note that we do not apply masks to differentiate the land cover types or crop types inside the geographical regions. Instead, we rely on the model itself to extract necessary information from the sequences of raw images to make accurate crop yield prediction. Another caveat is that the fused heterogeneous geographic data originally have different spatial/temporal resolutions, noise sources (e.g., cloud, urban agglomerations, etc.) and data acquisition errors. This also poses challenges to our model to extract meaningful spatiotemporal signals for the prediction task.

B. Baseline Methods

The baseline methods that are compared with can be classified into three categories based on the dimensions of the input data. The first category includes ridge regression (LR) [6], random forest (RF) [39], and multilayer perceptron (MLP) [9]. These are conventional machine learning models that have been widely used in crop yield prediction tasks. The inputs they take are scalar features in time series which can be obtained by spatially averaging the features at each time step.

The second category includes two deep learning models introduced in [1], LSTM+GP and CNN+GP. Instead of taking average of features in spatial space, they convert the spatial image to histogram to keep the frequency distribution which results in a three-dimensional input space. The input is then fed into a LSTM or CNN for information extraction. Additionally,

[1]https://github.com/facebookresearch/Context-Aware-Representation-Crop-Yield-Prediction

they adopt a Gaussian process (GP) after the prediction of deep models to alleviate the spatial correlations in the prediction errors. The Gaussian process was found to be able to boost model performances in [1].

The third category includes two more advanced deep learning models, 3d convolutional neural networks (C3D) [22], [40] and recurrent convolutional networks (CNN-LSTM) [19] that recognize the spatiotemporal nature of the input as our proposed methods. We adopt a similar architecture for the 3d convolutional neural networks as in [40], with minor modifications to adapt to the height and width of our input. CNN-LSTM deploys the same convolutional networks as our proposed models, but uses the LSTM model, instead of the hybrid attention model, for temporal information extraction. Both C3D and CNN-LSTM take the four-dimensional features as input.

Additional feature standardization processes, including subtraction of mean and division of standard deviation, are applied to the input data of LR and MLP. As for the deep models, LSTM+GP and CNN+GP, raw features are fed to generate histograms, after which subtraction of mean is applied as in [1]. The same standardization process as our proposed models is applied to the input taken by CNN-LSTM and C3D, which will be introduced in Section IV-C.

C. Our Approaches

In addition to the proposed semi-supervised hybrid attention model (SEMI-HA), we evaluate two other models, a supervised-only hybrid attention model (S-HA) and a semi-supervised self-attention model (SEMI-SA). S-HA shares a similar architecture as the proposed SEMI-HA with an exception that no sampled spatial neighbor/distant and temporal neighbor data are provided to constrain the unsupervised representation learning. Evaluations of S-HA is to demonstrate the potential advantage of introducing sampled unlabeled spatial and temporal context, especially under the stress of data scarcity. Comparisons with SEMI-SA instead can be used to show how adding global query impacts the temporal information extraction of the attention model. All three attention models take four-dimensional data as input.

We standardize the input by subtracting from it per-channel mean and dividing it by per-channel standard deviation. The standardization process is applied to each month separately as we observe that there are monthly variations in the mean and standard deviation of all the features. Monthly per-channel mean and per-channel standard deviation are obtained through using input data from the year 2003 to 2013.

D. Evaluation Approach and Metrics

To evaluate the generalization ability of the baseline methods and our proposed approaches to unseen data in future years, we adopt a temporal nested validation approach. We conduct prediction experiments for 5 years between 2014 and 2018 independently. When a year y is selected to collect the test data, data from year $y-1$ are used for validation, while data collected from year $y-N_y-1$ to $y-2$ are used for training. N_y here can be used to control the size of training data and test the performance of a model under the stress of data scarcity.

We report Root Mean Square Error (RMSE) and R^2 as the evaluation metrics. Both RMSE and R^2 have been widely used to evaluate the crop yield prediction performances in previous studies [1], [6], [7]. RMSE measures the consistency of prediction results and ground-truth values. R^2 measures the fraction of variance in ground-truth values that can be explained by predictions. R^2 is not as scale-dependent as RMSE.

E. Hyperparameter Tuning

We tune the hyperparameters of the baseline methods and the proposed approaches based on the performance of the validation dataset. For all the deep models, including our attention models, 50 epochs are run with the best model saved based on validation performance. Grid search from reasonable hyper-parameter combinations is adopted for LR, RF, and MLP.

Generally, we find our attention models easy to tune and less prone to overfitting. Weights for the unsupervised loss, temporal unsupervised loss, and regularization (α, β, and γ as in Equation (2)) are set as 0.2, 0.001, 0.2, respectively. An exception is made for 2016 where we find through validation that it is more sensitive to unsupervised training part than other years. The unsupervised loss weight α is changed from 0.2 to 0.1. Radius to sample spatial neighbor and spatial distant are set as 25 and 100, respectively. As the time step granularity of our feature is relatively coarse, i.e., monthly, we fixed the temporal neighborhood as the region that appears at one time step early. It can be easily adjusted for other tasks that have more sensitive time granularity though. For all the results of our proposed approaches reported in this paper, the aforementioned hyperparameters are adopted.

F. Results of Model Performance

We first set N_y to 10 in comparing all the methods, which means 10 years of past data are used for training. This size of training data has been adopted widely in previous studies [1], [39]. Table I and Table II show the empirical results for the comparison with baselines in terms of RMSE and R^2 respectively. It can be seen that our approaches consistently outperform all the baseline methods with significant margins. A 12.5% improvement in terms of RMSE and 15.1% improvement in terms of R^2 can be seen when comparing SEMI-HA with the best-performing baseline C3D. In comparison with the supervised-only model S-HA, a 3.6% and 3.2% improvements in terms of RMSE and R^2 are observed. Also, SEMI-HA outperforms SEMI-SA by 2.1% in RMSE and 1.5% in R^2. It is worth noting that supervised-only model S-HA outperforms CNN-LSTM with 12.8% improvement in RMSE and 21.0% improvement in R^2. This demonstrates the advantage of the hybrid attention model over LSTM in capturing temporal patterns on this crop yield prediction task.

TABLE I
RMSE COMPARISON OF VARIOUS METHODS WHEN 10-YEAR DATA ARE USED FOR TRAINING.

Method	Year					Avg
	2014	2015	2016	2017	2018	
LR	6.465	7.754	7.589	6.839	8.163	7.362
RF	5.332	6.69	8.134	6.352	7.692	6.840
MLP	5.236	6.076	6.752	6.025	8.242	6.466
LSTM+GP	5.013	5.553	6.761	5.134	5.522	5.597
CNN+GP	4.824	5.540	8.136	5.706	6.235	6.088
C3D	4.969	5.891	5.462	5.736	5.558	5.523
CNN-LSTM	4.948	5.698	6.688	5.596	5.83	5.752
S-HA	4.545	5.453	4.9	4.701	5.467	5.013
SEMI-SA	4.673	5.251	**4.694**	4.587	5.47	4.935
SEMI-HA	**4.502**	**5.19**	4.754	**4.354**	**5.363**	**4.833**

TABLE II
R^2 COMPARISON OF VARIOUS METHODS WHEN 10-YEAR DATA ARE USED FOR TRAINING.

Method	Year					Avg
	2014	2015	2016	2017	2018	
LR	0.495	0.231	0.002	0.458	0.361	0.309
RF	0.657	0.428	-0.15	0.532	0.432	0.380
MLP	0.669	0.528	0.209	0.579	0.347	0.466
LSTM+GP	0.696	0.605	0.207	0.694	0.707	0.582
CNN+GP	0.719	0.607	-0.148	0.622	0.627	0.485
C3D	0.7	0.556	0.481	0.618	0.704	0.612
CNN-LSTM	0.701	0.582	0.224	0.637	0.674	0.564
S-HA	0.75	0.619	0.584	0.744	0.714	0.682
SEMI-SA	0.736	0.646	**0.618**	0.756	0.712	0.694
SEMI-HA	**0.755**	**0.655**	0.607	**0.78**	**0.724**	**0.704**

One thing that deserves to be mentioned is that all the baseline methods perform poorly in the year 2016, i.e., the RMSE values are significantly higher and R^2 values are significantly lower than other years. One potential reason for this is that the year 2016 saw a record high yield across nearly all the midwestern states [41]. The information extracted through the baseline methods fails to catch the causes that lead to such a disparate pattern. Our attention models, including the semi-supervised ones and the supervised-only one, instead, are able to extract necessary signals from past data to generalize well in this case.

To further test the performances of all the approaches under the stress of labeled data scarcity, we decrease the number of years to be used as the training data from 10 to 7, then 4, e.g., when $N_y = 4$, there are only 4 years of data used for training, 1 year for validation, and 1 year for test. This is a relatively extreme case but can shed some light on practical cases where input labeled data are less available or expensive to obtain, such as in developing counties or when a manual survey has to be adopted to collect the data.

Figure 5 shows the results averaged from 2014 to 2018 with the x-axis representing the number of labeled training instances. To have better readability, we separate the comparisons with the baseline methods based on the dimensionality of the input data. It can be seen that our models again outperform all the baseline methods with big margins consistently. To better visualize the trends of the model performances with the decreasing size of training data, we plot the averaged RMSE and R^2 from 2014 and 2018 for our attention models and the top four baselines, C3D, CNN-LSTM, LSTM+GP and CNN+GP, as shown in Figure 6. There are two observations that we would like to mention here. First is that the advantage of the semi-supervised model over the supervised-only model becomes more obvious when the size of training data decreases. The improvements in terms of RMSE/R^2 are increased to 4.1%/4.3% when 7 years of data are used for training and 4.5%/7.1% when 4 years of data are used for training. The second observation is that the performance gap between our models which take full advantage of the data and the models with dimensionality reduction techniques still remains when less data are used for training. While it was claimed in [1] that the dimensionality reduction technique adopted by LSTM+GP and CNN+GP was to alleviate the issue of training data scarcity, we see a 14.5% RMSE and 40.6% R^2 improvement comparing our SEMI-HA model with CNN+GP when 4 years of labeled data are used for training.

G. Feature Importance Analysis

To understand how the proposed SEMI-HA model exploits the geo-spatiotemporal data fused from various sources, we provide a feature importance analysis here by excluding one type of environmental factors at a time from the input. The importance of a feature or a set of features can be demonstrated through the performance drop when they are not included in the input for training. The average performances from 2014 to 2018 of the semi-supervised model is shown in Figure 7. Ten years of training data are used here. Specifically, we group the environmental features into five groups. Group VI is for the input when two vegetation indices are excluded. Group LST is for the land surface temperatures during the day and night. Group PPT, SOIL, and ELE are for the precipitation, soil properties and elevation, respectively. We add group ALL for comparison which uses all available factors. It can be seen from the figure that all groups see performance drops to different extents, which means that our model is capturing information from all environmental factors. Another interesting observation is that our model can still achieve satisfying performance when the two vegetation indices are not included and only climate data are used (group VI). It has been reported in previous studies [7], [42] that either vegetation indices or the source of their data, i.e., the multi-spectral remote sensing images, are necessary to achieve a satisfying crop yield prediction performance. Our semi-supervised hybrid attention model, however, is able to achieve comparable performance solely using the climate data. (Note that group VI even performs similar as the C3D model with all features included.) This achievement is established because our model is more capable of extracting spatiotemporal signals in the climate data for producing more accurate predictions.

V. CONCLUSION

Machine learning approaches that exploit geo-spatiotemporal data have played a crucial role in many environmental sustainability applications. Unfortunately, the

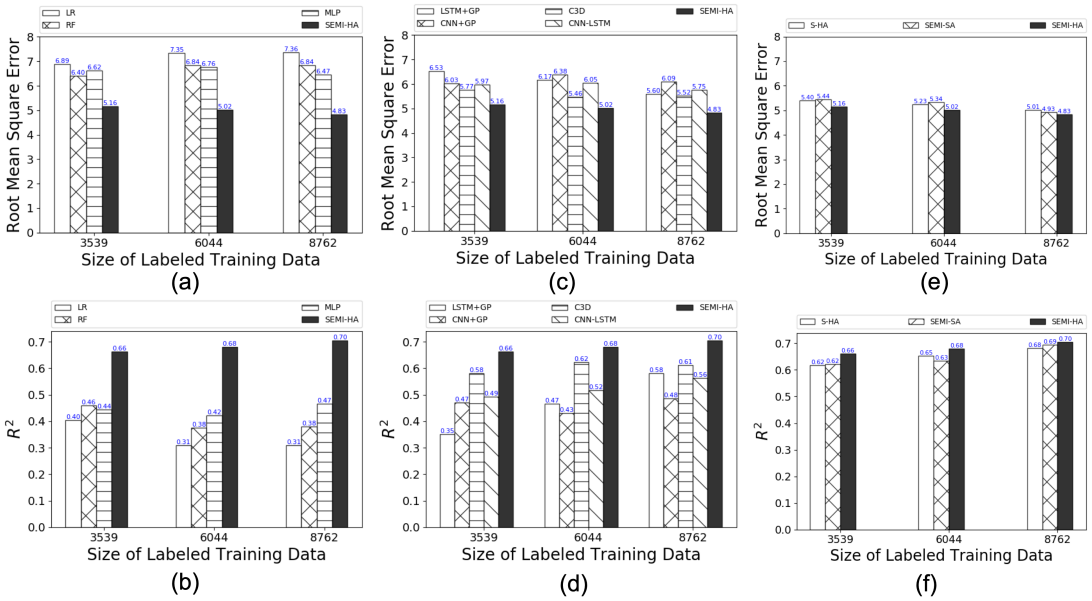

Fig. 5. Model performance with varying numbers of labeled training instances. (a)-(b) Comparison between SEMI-HA and conventional machine learning methods. (c)-(d) Comparison between SEMI-HA and deep learning models (with three/four-dimensional input). (e)-(f) Comparison between three attention models, SEMI-HA, S-HA, and SEMI-SA. All numbers are averaged from 2014 to 2018.

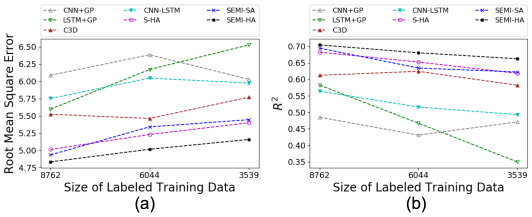

Fig. 6. The trend of RMSE and R^2 for top performing models with decreasing numbers of labeled training instances. (a) RMSE and (b) R^2. All numbers are averaged from 2014 to 2018.

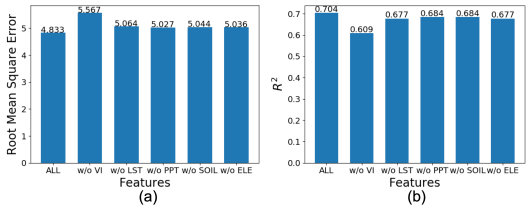

Fig. 7. The performance drop with one type of environmental factor excluded from the input. The ALL group uses all available factors.

data heterogeneity and label scarcity often pose fundamental challenges to such applications. To tackle these challenges, we propose a novel semi-supervised hybrid attention model. This model takes full advantage of the multi-channel spatiotemporal data and is able to learn spatiotemporal representations for downstream prediction tasks. The proposed hybrid attention model improves the classic self-attention model by integrating global trainable query. More complete temporal dependencies adapted to training data/task can thus be captured. To overcome the limitation in label scarcity, we introduce unsupervised representation learning where spatial and temporal context sampled from unlabeled data are utilized. Our model jointly minimizes the unsupervised loss along with the supervised loss for the learning tasks. To evaluate the effectiveness of the proposed methods, we compare its performance with many state-of-the-art baselines in a regression task over large-scale real-world data. The experimental results clearly demonstrate the advantages of the proposed methods. We observe that the advantages become more obvious when less data are utilized in the training phase. To verify the impacts of this unsupervised loss, we compare the performance of the model with vs. without the unsupervised part. The results justify the positive impacts of the unsupervised loss in improving the overall performance of the proposed method.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under award CCF-1934904 granted to Texas AM University, Texas AM University, the Texas Engineering Experiment Station, and Texas AM AgriLife Research.

REFERENCES

[1] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep gaussian process for crop yield prediction based on remote sensing data," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[2] H. Mao, D. Kathuria, N. Duffield, and B. P. Mohanty, "Gap filling of high-resolution soil moisture for smap/sentinel-1: A two-layer machine learning-based framework," *Water Resources Research*, vol. 55, no. 8, pp. 6986–7009, 2019.

[3] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 3967–3974.

[4] S. G. Subramanian and M. Crowley, "Learning forest wildfire dynamics from satellite images using reinforcement learning," in *Conference on Reinforcement Learning and Decision Making*, 2017.

[5] V. Masson, J.-L. Champeaux, F. Chauvin, C. Meriguet, and R. Lacaze, "A global database of land surface parameters at 1-km resolution in meteorological and climate models," *Journal of climate*, vol. 16, no. 9, pp. 1261–1282, 2003.

[6] A. Mateo-Sanchis, M. Piles, J. Muñoz-Marí, J. E. Adsuara, A. Pérez-Suay, and G. Camps-Valls, "Synergistic integration of optical and microwave satellite data for crop yield estimation." *Remote Sensing of Environment*, vol. 234, p. 111460, 2019. [Online]. Available: http://dx.doi.org/10.1016/j.rse.2019.111460

[7] H. Jiang, H. Hu, R. Zhong, J. Xu, J. Xu, J. Huang, S. Wang, Y. Ying, and T. Lin, "A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US corn belt at the county level." *Global Change Biology*, 2019. [Online]. Available: http://dx.doi.org/10.1111/gcb.14885

[8] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, p. 104859, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168169919306842

[9] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, "Soybean yield prediction from UAV using multimodal data fusion and deep learning," *Remote Sensing of Environment*, vol. 237, p. 111599, 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425719306194

[10] H. J. Miller, "The data avalanche is here," *Shouldn't we be*, 2010.

[11] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[12] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] Y. Liu, H. Yuan, and S. Ji, "Learning local and global multi-context representations for document classification," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1234–1239.

[15] J. Lässig, K. Kersting, and K. Morik, *Computational Sustainability*. Springer, 2016, vol. 645.

[16] X. Jia, A. Khandelwal, G. Nayak, J. Gerber, K. Carlson, P. West, and V. Kumar, "Incremental dual-memory lstm in land cover prediction," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 867–876.

[17] D. Entekhabi, N. Das, E. Njoku, J. Johnson, and J. Shi, "Algorithm theoretical basis document l2 & l3 radar/radiometer soil moisture (active/passive) data products," *Jet Propulsion Lab. Report*, vol. 1, 2014.

[18] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 4, pp. 549–553, 2017.

[19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[20] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[21] J. Zhao, B. Du, L. Sun, F. Zhuang, W. Lv, and H. Xiong, "Multiple relational attention network for multi-task learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1123–1131.

[22] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[23] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[24] X. Jia, S. Li, A. Khandelwal, G. Nayak, A. Karpatne, and V. Kumar, "Spatial context-aware networks for mining temporal discriminative period in land cover detection," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 513–521.

[25] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[28] V. Ramanathan, K. Tang, G. Mori, and L. Fei-Fei, "Learning temporal embeddings for complex video analysis," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4471–4479.

[29] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[30] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, "Understanding and improving transformer from a multi-particle dynamic system point of view," *arXiv preprint arXiv:1906.02762*, 2019.

[31] D. M. Johnson, "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the united states," *Remote Sensing of Environment*, vol. 141, pp. 116–128, 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425713003957

[32] U. NASS, "Quick stats database. washington, dc: Usda national agricultural statistics service," 2017.

[33] K. Didan, "Mod13a3 modis/terra vegetation indices monthly l3 global 1km sin grid v006," *NASA EOSDIS Land Processes DAAC*, 2015.

[34] D. B. Lobell, M. J. Roberts, W. Schlenker, N. Braun, B. B. Little, R. M. Rejesus, and G. L. Hammer, "Greater sensitivity to drought accompanies maize yield increase in the us midwest," *Science*, vol. 344, no. 6183, pp. 516–519, 2014.

[35] C. Daly, M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, "Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states," *International Journal of Climatology: a Journal of the Royal Meteorological Society*, vol. 28, no. 15, pp. 2031–2064, 2008.

[36] Z. Wan, S. Hook, and G. Hulley, "Mod11a1 modis/terra land surface temperature/emissivity daily l3 global 1 km sin grid v006 [data set]. nasa eosdis lp daac," 2015.

[37] J. NASA, "Nasa shuttle radar topography mission global 1 arc second," *NASA LP DAAC*, vol. 15, 2013.

[38] D. A. Miller and R. A. White, "A conterminous united states multilayer soil characteristics dataset for regional climate and hydrology modeling," *Earth interactions*, vol. 2, no. 2, pp. 1–26, 1998.

[39] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," *Frontiers in Plant Science*, vol. 10, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpls.2019.01750/full

[40] H. Russello, "Convolutional neural networks for crop yield prediction using satellite images," *IBM Center for Advanced Studies*, 2018.

[41] S. Irwin and D. Good, "Uncertainty about 2016 us average corn and soybean yields persists," *Farmdoc Daily*, vol. 6, no. 132, 2016.

[42] B. Peng, K. Guan, M. Pan, and Y. Li, "Benefits of seasonal climate prediction and satellite data for forecasting u.s. maize yield," *Geophysical Research Letters*, vol. 45, no. 18, pp. 9662–9671, 2018. [Online]. Available: http://doi.wiley.com/10.1029/2018GL079291