MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer

Fudong Lin[1], Summer Crawford[2], Kaleb Guillot[2], Yihe Zhang[2], Yan Chen[3], Xu Yuan[1]*,
Li Chen[2], Shelby Williams[2], Robert Minvielle[2], Xiangming Xiao[4], Drew Gholson[5], Nicolas Ashwell[5],
Tri Setiyono[6], Brenda Tubana[7], Lu Peng[8], Magdy Bayoumi[2], Nian-Feng Tzeng[2]
[1] University of Delaware, [2] University of Louisiana at Lafayette, [3] University of Connecticut,
[4] University of Oklahoma, [5] Mississippi State University, [6] Louisiana State University,
[7] LSU AgCenter, [8] Tulane University

Abstract

Precise crop yield prediction provides valuable information for agricultural planning and decision-making processes. However, timely predicting crop yields remains challenging as crop growth is sensitive to growing season weather variation and climate change. In this work, we develop a deep learning-based solution, namely Multi-Modal Spatial-Temporal Vision Transformer (MMST-ViT), for predicting crop yields at the county level across the United States, by considering the effects of short-term meteorological variations during the growing season and the long-term climate change on crops. Specifically, our MMST-ViT consists of a Multi-Modal Transformer, a Spatial Transformer, and a Temporal Transformer. The Multi-Modal Transformer leverages both visual remote sensing data and short-term meteorological data for modeling the effect of growing season weather variations on crop growth. The Spatial Transformer learns the high-resolution spatial dependency among counties for accurate agricultural tracking. The Temporal Transformer captures the long-range temporal dependency for learning the impact of long-term climate change on crops. Meanwhile, we also devise a novel multi-modal contrastive learning technique to pre-train our model without extensive human supervision. Hence, our MMST-ViT captures the impacts of both short-term weather variations and long-term climate change on crops by leveraging both satellite images and meteorological data. We have conducted extensive experiments on over 200 *counties in the United States, with the experimental results exhibiting that our MMST-ViT outperforms its counterparts under three performance metrics of interest. Our dataset and code are available at* `https://github.com/fudong03/MMST-ViT`.

1. Introduction

Accurate crop yield prediction is essential for agricultural planning and advisory processes [26], informed economic decisions [2], and global food security [36]. However, predicting crop yields precisely is challenging as it requires to consider the effects of i) short-term weather variations, governed by the meteorological data during the growing season, and ii) long-term climate change, governed by historical meteorological factors, on crops simultaneously. Meanwhile, precise crop tracking relies on high-resolution remote sensing data. While process-based prediction approaches [44, 11, 25, 54] exist, they often suffer from high inaccuracies due to their strong assumptions on management practices [15]. On the other hand, motivated by the recent success of deep neural networks [29, 19, 50, 12, 20, 56, 37, 32, 18, 30, 8], deep learning (DL)-based methods have been widely adopted for crop yield predictions, due to their effectiveness in accurate agricultural tracking [27, 26] and their potent capabilities in capturing the spatial and temporal variation of meteorological data [28, 15].

So far, DL-based solutions for crop yield predictions can be roughly grouped into two categories, *i.e.*, remote sensing data-based and meteorological data-based approaches. The former [27, 26, 53, 13, 17, 47, 10] employs such remote sensing data as satellite images, unmanned aerial vehicle (UAV)-based imagery data, or vegetation indices to estimate the annual crop yield, while the latter [16, 1, 36, 42, 48] predicts the crop yield by using meteorological parameter data, including temperature, precipitation, vapor pressure deficit, *etc*. However, the former overlooks the direct impact of meteorological parameters on crop growth, while the latter lacks high-resolution remote sensing data for accurate agricultural tracking.

A recent study [39] has reported that the long-term climate change would gradually decrease the crop yield.

*Corresponding author: Dr. Xu Yuan (xyuan@udel.edu)

Driven by this discovery, follow-up pursuits have attempted to explore the effect of long-term climate change on crops. For example, the CNN-RNN model [28] demonstrates that the crop yield prediction can benefit from historical meteorological data, which is essential for measuring the impacts of climate change. Later, GNN-RNN [15] extends CNN-RNN by framing the crop yield prediction as the Spatial-Temporal forecasting problem. It employs Graph Neural Networks (GNN) and Long Short-Term Memory (LSTM) [22] respectively for learning spatial dependency among neighborhood counties and for capturing the impact of long-term meteorological data on crops. However, both of them only take into account the meteorological data for predictions, failing to leverage the remote sensing data for accurate agricultural tracking.

In this work, we aim to develop a new DL-based solution for predicting crop yields at the county level across the United States, by using both visual remote sensing data (from the Sentinel-2 satellite imagery [41]) and numerical meteorological data (from the HRRR model [24]). Our solution has two main goals. First, it captures the impacts of both short-term growing season weather variations and long-term climate change on crops. Second, it aims to leverage high-resolution remote sensing data for accurate agricultural tracking. To achieve our goals, we propose the Multi-Modal Spatial-Temporal Vision Transformer (MMST-ViT), motivated by the recent success of Vision Transformers (ViT) [12]. To the best of our knowledge, MMST-ViT is the first ViT-based model for real-world crop yield prediction. It advances previous CNN/GNN-based and RNN-based counterparts respectively with better generalization to the multi-model data and with more powerful abilities in capturing long-term temporal dependency. Its three components of a Multi-Modal Transformer, a Spatial Transformer, and a Temporal Transformer are each equipped with one novel Multi-Head Attention (MHA) mechanism [50]. Specifically, the Multi-Modal Transformer leverages satellite images and meteorological data during the growing season for capturing the direct impact of short-term weather variations on crop growth. The Spatial Transformer learns high-resolution spatial dependency among counties for precise crop tracking. The Temporal Transformer captures the effects of long-term climate change on crops. Since ViT-based models are prone to overfitting [12], we also develop a novel multi-modal contrastive learning technique that pre-trains our Multi-Modal Transformer without requiring human supervision. We have conducted experiments on over 200 counties in the United States. The experimental results exhibit that our MMST-ViT outperforms its state-of-the-art counterparts under three performance metrics of interest. For example, on the soybean prediction, our MMST-ViT achieves the lowest Root Mean Square Error (RMSE) value of 3.9, the highest R-squared (R^2) value of 0.843, and the best Pearson Correlation Coefficient (Corr) value of 0.918.

2. Related Work

Vision Transformers. Adopted from Transformers [50] in natural language processing, Vision Transformers (ViT) have exhibited commendable performance in various computer vision tasks. The original ViT [12] first partitions an image into a set of image patches, then applies Multi-Head Attention (MHA) [50] over the patches, and finally utilizes a learnable classification token to capture global visual representation for image classification. Several subsequent methods have been developed, including DeiT [46] for data-efficient ViT through knowledge distillation, Swin [33] for computation-efficient ViT using shifted windows, PVT [51] for dense prediction tasks, and MAE [18] for self-supervised learning, among others [7, 21, 14, 55, 3, 52, 31, 45]. However, applying prior ViT approaches to real-world crop yield prediction is challenging due to its needs of addressing the multi-modal inputs, of learning high-resolution spatial dependency, and of capturing long-range temporal dependency. Our work, based on ViT, advances existing methods by proposing three novel Multi-Head Attention (MHA) mechanisms respectively for leveraging both visual remote sensing data and numerical meteorological data, learning global spatial representation from multiple high-resolution data, and capturing the global temporal representation for measuring the long-term climate change effect. Additionally, we develop a new multi-modal contrastive learning technique to pre-train our multi-modal model for better prediction performance.

Deep Learning (DL) for Crop Yield Prediction. DL has been widely adopted for real-world crop yield predictions. Such prediction studies can be grouped into two categories: remote sensing data-based and meteorological data-based approaches. The former uses satellite images, unmanned aerial vehicle (UAV) data, or vegetation indices to estimate crop yields. Its prominent studies include the use of UAV-acquired RGB images [27] for predicting in-season crop yields, YieldNet [26] which resorts to transfer learning for simultaneously estimating the yields of multiple crop types, among many others [53, 13, 17, 47, 10]. By contrast, the latter utilizes deep neural networks (DNNs) to capture the impact of meteorological parameters on crop yields, including CNN-RNN [28] which incorporates long-term meteorological data, GNN-RNN [15] which extends CNN-RNN by using Graph Neural Networks (GNN) for learning spatial information, and others [16, 1, 36, 42, 48]. However, the remote sensing data-based solutions overlook the impact of meteorological parameters on crops, while the meteorological data-based solutions often fail to incorporate the remote sensing data, known to be crucial for accurate agricultural

Table 1: Overview of USDA Crop Dataset and HRRR Computed Dataset

Dataset	Parameters
USDA	Production, Yield
HRRR	Averaged Temperature, Maximal Temperature, Minimal Temperature, Precipitation, Relative Humidity, Wind Gust, Wind Speed, Downward Shortwave Radiation Flux, Vapor Pressure Deficit

tracking. Our work differs from prior studies in two aspects. First, it leverages both remote sensing data and meteorological data for capturing the impacts of both short-term growing season meteorological variations and the long-term climate change on crops. Second, it is the first to use Vision Transformers for crop yield predictions, advancing previous CNN/GNN- and RNN-based models respectively with better generalization to multi-modal data and with higher abilities for capturing long ranges of temporal dependency.

3. Datasets

In this study, we utilize three types of data for accurate county-level crop yield predictions: i) crop data from the United States Department of Agriculture (USDA), ii) meteorological data from the High-Resolution Rapid Refresh (HRRR), and iii) remote sensing data from the Sentinel-2 satellite, as outlined below.

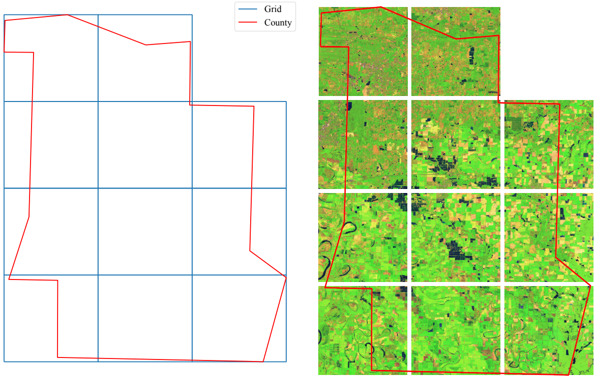
USDA Crop Dataset. The dataset, sourced from the United States Department of Agriculture (USDA) [49], provides annual crop data for major crops grown in the United States (U.S.), including corn, cotton, soybean, winter wheat, *etc*., on a county-level basis. It covers crop information such as production and yield from 2017 to 2022 (as listed in the second row of Table 1).

HRRR Computed Dataset. The dataset, obtained from the High-Resolution Rapid Refresh atmospheric model (HRRR) [24], provides high-resolution meteorological data for the contiguous U.S. continent. It covers 9 weather parameters from 2017 to 2022 (see the third row in Table 1 for more information).

Sentinel-2 Imagery. Sentinel-2 imagery is a set of images captured by the Sentinel-2 Earth observation satellite. It provides agriculture imagery for the contiguous U.S. continent from 2017 to 2022 with a 2-week interval. Since precise agricultural tracking requires high-resolution remote sensing data, the image of a county is partitioned into multiple fine-grained grids (9×9 km). Figure 1 shows an example of county partitioning.

4. Method

In this work, we aim to develop a deep learning (DL)-based model for predicting crop yields at the county level. Our goals are twofold. First, we plan to capture the meteorological effect, including growing season weather variations and climate change, on crop yields. Second, we aim to leverage high-resolution satellite images for precise agricultural tracking. The aforementioned three data sources are utilized for achieving our goals.



(a) Grid Example (b) Sentinel-2 Imagery

Figure 1: Illustration of partitioning a county into multiple grids at the 9km high-resolution. (a) An example of county partitioning, with the red line segments indicating the geometry boundary for the county and the blue line segments representing the high-resolution grids. (b) The resulting satellite images in the Sentinel-2 Imagery, with each composed of 384×384 pixels, depicting an area of 9×9 km.

4.1. Problem Statement

We consider the combination of four types of data $(\boldsymbol{x}, \boldsymbol{y}_s, \boldsymbol{y}_l, \boldsymbol{z})$ for predicting the crop yield at each single U.S. county. Specifically, $\boldsymbol{x} \in \mathbb{R}^{T \times G \times H \times W \times C}$ represents satellite images obtained from Sentinel-2 imagery, which capture the information of field crops on the ground. T and G indicate the numbers of temporal and spatial data, respectively, whereas H, W, and C are the height, width, and number of channels in the satellite image. $\boldsymbol{y}_s \in \mathbb{R}^{T \times G \times N_1 \times d_y}$ and $\boldsymbol{y}_l \in \mathbb{R}^{T \times N_2 \times d_y}$ are meteorological parameters obtained from the HRRR dataset, representing the short-term and the long-term historical data, respectively. Here, N_1 and N_2 are the numbers of daily and monthly HRRR data points, respectively, and d_y indicates the number of weather parameters. Note that the short-term meteorological data is the daily HRRR data during the growing season, while the long-term historical meteorological data is the monthly HRRR data from the past several years (*e.g.*, 2018 to 2020 for predicting crop yields in 2021). $\boldsymbol{z} \in \mathbb{R}^{d_z}$ is the ground-truth crop information obtained from the USDA dataset, with d_z representing the number of parameters for the crop data.

4.2. Challenges

Three challenges exist upon achieving our goals, outlined as follows.

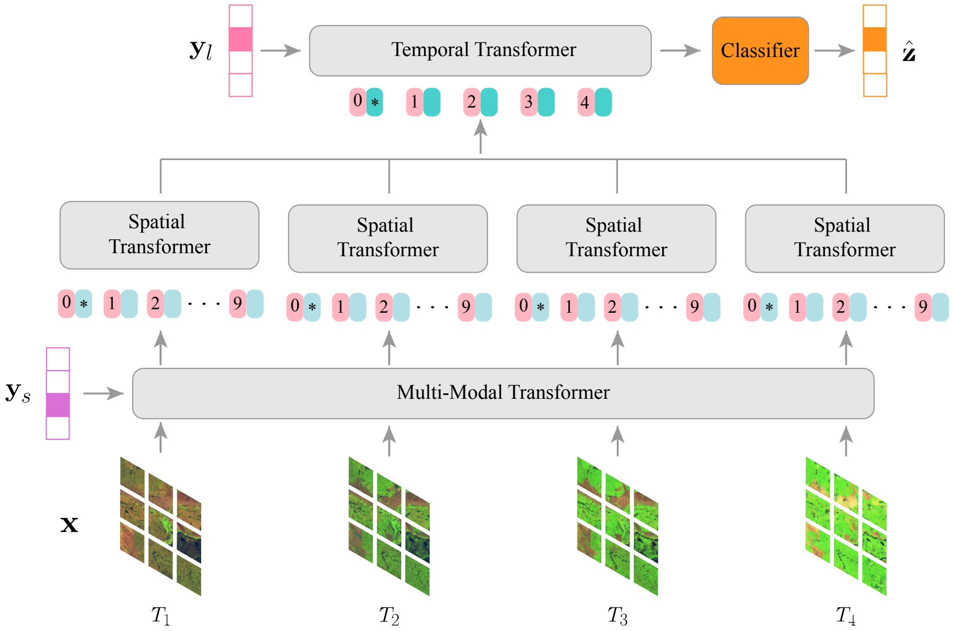


Figure 2: The architecture of our proposed MMST-ViT.

Capturing the Effect of Growing Season Weather Variations on Crop Growth. Previous studies only consider the meteorological data (or the remote sensing data) for crop yield predictions, overlooking the direct impact of growing season weather variations on crop growth. To capture such an impact, we aim to leverage both visual remote sensing and numerical meteorological data. In particular, the Multi-Head Attention (MHA) technique [50] is adopted to capture the meteorological effects on crop growth. However, how to perform multi-modal attention over visual remote sensing data and numerical meteorological data remains open.

Lacking a Mechanism for Pre-training Multi-Modal Model. Deep neural networks (DNNs), especially Vision Transformer (ViT)-based models, are prone to overfitting, requiring appropriate pre-training to achieve satisfactory performance. Unfortunately, conventional pre-training techniques (*e.g.*, SimCLR [9]) only marginally improve crop yield prediction performance as they consider the visual data only, ineffective for pre-training multi-modal models. How to pre-train the multi-modal model for satisfactory crop yield predictions is challenging to be addressed.

Capturing the Impact of Climate Change on Crops. As reported by a prior study [39], the long-term climate change in the atmosphere would gradually decrease the crop yield on the ground. Some studies [28, 15] have demonstrated that taking the climate change effect into account can better crop yield prediction. But their designs overlook the remote sensing data, viewed as essential for precise agricultural tracking. So far, how to develop effective ways to capture the impact of long-term climate change on crop yields is challenging and unanswered yet.

4.3. Our Proposed Approach

To tackle the aforementioned challenges, we develop the <u>M</u>ulti-<u>M</u>odal <u>S</u>patial-<u>T</u>emporal <u>V</u>ision <u>T</u>ransformer (MMST-ViT) for predicting crop yields at the county level, by leveraging the remote sensing data and the short-term and long-term meteorological data.

Model Overview. Our proposed MMST-ViT consists of three key components, *i.e.*, Multi-Modal Transformer, Spatial Transformer, and Temporal Transformer, as shown in Figure 2. The Multi-Modal Transformer is designed to capture the impact of short-term meteorological variations on crop growth, by leveraging the satellite images (*i.e.*, $\boldsymbol{x}$) and the short-term meteorological parameters (*i.e.*, $\boldsymbol{y}_s$). The Spatial Transformer then utilizes the output of the Multi-Modal Transformer for learning the global spatial information of a county. Next, the Temporal Transformer combines the outputs of our Spatial Transformer and the long-term meteorological data (*i.e.*, $\boldsymbol{y}_l$) to capture both global temporal information and the impact of long-term climate change on crop yields. Finally, the output of our Temporal Transformer is used by a linear classifier for predicting the annual crop yields. The details of each component are provided below.

Multi-Modal Transformer. Aiming to capture the direct impact of atmospheric weather variations on crop growth, the Multi-Modal Transformer consists of a visual backbone network and a multi-modal attention layer. The former extracts high-quality visual representations from satellite images for accurate agricultural tracking, while the latter captures the relationship between the visual representation and meteorological parameters, to understand the impact of meteorological parameters on crop growth. Let $f_{\boldsymbol{\theta}}$: $(\mathbb{R}^{T\times G\times H\times W\times C}, \mathbb{R}^{T\times G\times N_1\times d_y}) \rightarrow \mathbb{R}^{T\times G\times d}$ be our Multi-Modal Transformer, with $\boldsymbol{\theta}$ denoting the parameters for DNNs and d representing the dimension for hidden vectors. Notably, all the hidden dimensions in this paper are set to the same size of d. As such, the proposed Multi-Modal Transformer can be expressed as $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}_s) = \boldsymbol{v}_m$, where $\boldsymbol{x}$ and $\boldsymbol{y}_s$ are the Sentinel-2 images and the short-term meteorological data, respectively. $\boldsymbol{v}_m \in \mathbb{R}^{T\times G\times d}$ is the output of our Multi-Modal Transformer.

In this work, we utilize Pyramid Vision Transformer (PVT) [51] as the visual backbone network since it advances ResNet [19] and vanilla ViT [12] respectively by having the global receptive field and by enabling high-resolution feature maps. To capture the direct impact of meteorological parameters on crop growth during the growing season, a naive way is by concatenating visual and numerical representations extracted respectively from the Sentinel-2 images and the meteorological parameters. Unfortunately, our empirical results show that such a naive way cannot achieve satisfactory performance. Inspired by the recent success of multi-head attention mechanisms [50, 12, 40], we devise a novel Multi-Modal Multi-Head Attention (MM-MHA) mechanism to capture the impact of meteorological parameters on crop growth, mathematically expressed as

$$\begin{gathered}\text{MM-MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d})\boldsymbol{V}, \\ \boldsymbol{Q} = \boldsymbol{W}_m^Q \cdot \varphi_m(\boldsymbol{x}),\ \boldsymbol{K} = \boldsymbol{W}_m^K \cdot \pi_m(\boldsymbol{y}_s), \\ \boldsymbol{V} = \boldsymbol{W}_m^V \cdot \pi_m(\boldsymbol{y}_s).\end{gathered} \tag{1}$$

Here, $\varphi_m(\boldsymbol{x}) \in \mathbb{R}^{T\times G\times N_p\times d}$ is the visual representation encoded by PVT, with N_p indicating the number of image patches. $\pi_m(\boldsymbol{y}_s) \in \mathbb{R}^{T\times G\times N_1\times d}$ is the meteorological parameters after the linear projection. Notably, $\boldsymbol{W}_m^Q, \boldsymbol{W}_m^K$, and $\boldsymbol{W}_m^V$ are learnable projection matrices, similar to those in prior studies [50, 40]. To our knowledge, this is the first attempt at developing a multi-modal multi-attention approach for leveraging both visual remote sensing data and numerical meteorological data.

However, ViT-based models are highly sensitive to overfitting, calling for pre-training with millions of visual data [12]. Meanwhile, real-world crop yield prediction usually lacks sufficient crop data for pre-training. Worse still, capturing the impact of short-term weather conditions on crops requires considering both the visual and the numerical data simultaneously. Although self-supervised learning techniques [9, 5, 18] have recently been developed for pre-training DNNs, they cannot tackle the aforementioned issue at the same time. For example, SimCLR [9] fails to address the multi-modal data issue.

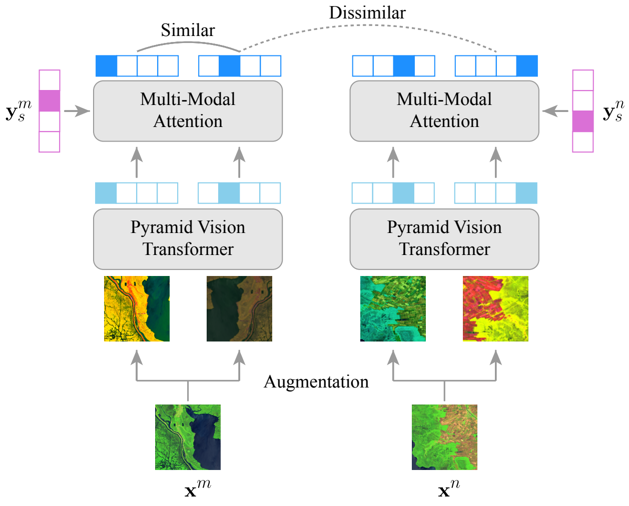


Figure 3: Multi-modal self-supervised learning.

Here, we propose a novel multi-modal self-supervised learning for pre-training our Multi-Modal Transformer without human supervision, which is inspired by SimCLR, but having two differences. First, we utilize the PVT instead of convolutional neural networks (CNNs) as the backbone network since PVT advances CNNs by having a global receptive field [51]. Second, we replace the Multi-Layer Perceptron (MLP) layer in SimCLR with our proposed multi-modal self-attention layer (*i.e.*, Eq. (1)) for simultaneously tackling visual and numerical data. Figure 3 illustrates its architecture. Given a satellite image $\boldsymbol{x}^m$ (or $\boldsymbol{x}^n$), we perform the data augmentation to get its two augmented images, and then feed them to the PVT to get two sets of visual representations. After that, we use our proposed MM-MHA (*i.e.*, Eq. (1)) to perform multi-modal attention between the visual representations and the corresponding short-term parameters $\boldsymbol{y}_s^m$ (or $\boldsymbol{y}_s^n$), arriving at two sets of output sequences. Here, we regard two sets of output sequences out of a given satellite image (*e.g.*, $\boldsymbol{x}^m$) as the positive pair. Meanwhile, two sets of output sequences from different satellite images (*e.g.*, $\boldsymbol{x}^m$ and $\boldsymbol{x}^n$) are regarded as the negative pair. Similar to vanilla ViT [12], the head of output sequences (*i.e.*, the classification token) is taken as the output of our Multi-Modal Transformer, *i.e.*, the hidden vector $\boldsymbol{v}_m$. As such, our multi-modal contrastive loss is defined as,

$$\begin{gathered}\ell(i,j) = -\log\frac{\exp(\text{sim}(\boldsymbol{v}_m^i, \boldsymbol{v}_m^j)/\tau)}{\sum_{i\neq k, k=1,\ldots,2B}\exp(\text{sim}(\boldsymbol{v}_m^i, \boldsymbol{v}_m^k)/\tau)}, \\ \mathcal{L}_{\text{cl}} = \frac{1}{2B}\sum_{k=1}^{B}\left[\ell(k, k+B) + \ell(k+B, k)\right],\end{gathered} \tag{2}$$

where $B = T \times G$. Here, $\boldsymbol{v}_m^i$ and $\boldsymbol{v}_m^j$ are the positive pair, and τ is the temperature parameter. Intuitively, our

Eq. (2) encourages two hidden vectors from the same satellite image (*i.e.*, representing the same region) to be similar, and two hidden vectors from different satellite images (*i.e.*, representing different regions) to be dissimilar. As such, we can pre-train our Multi-Modal Transformer to capture the impact of short-term meteorological parameters on crop growth without labor-intensive human supervision.

Spatial Transformer. Recall that the Sentinel-2 Imagery dataset of a county is partitioned into multiple fine-grained grids for precise agricultural tracking. To learn spatial dependency among those grids, we propose a Spatial Transformer $g_{\phi} : \mathbb{R}^{T \times G \times d} \to \mathbb{R}^{T \times d}$, whose purpose is to capture the global spatial representations for counties. The design of our Spatial Transformer is inspired by the vanilla ViT [12] but with a flexible number of positional embeddings. It is used to tackle the scenario that the number of grids varies among counties as it depends on county sizes. To learn global spatial information for counties, the Spatial Transformer prepends a learnable classification token $\boldsymbol{v}_m^{\text{cls}} \in \mathbb{R}^{T \times 1 \times d}$, arriving at

$$\varphi_s(\boldsymbol{v}_m) = [\boldsymbol{v}_m^{\text{cls}};\ \boldsymbol{v}_m^1;\ \boldsymbol{v}_m^2;\ \ldots\ ;\ \boldsymbol{v}_m^G] + \mathbf{E}_{\text{pos}}, \tag{3}$$

where $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{T \times (G+1) \times d}$ is the flexible positional embeddings. We propose a Spatial Multi-Head Attention (S-MHA) to learn spatial dependency among grids. Mathematically,

$$\begin{gathered}\text{S-MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d})\boldsymbol{V}, \\ \boldsymbol{Q} = \boldsymbol{W}_s^Q \cdot \varphi_s(\boldsymbol{v}_m),\ \boldsymbol{K} = \boldsymbol{W}_s^K \cdot \varphi_s(\boldsymbol{v}_m), \\ \boldsymbol{V} = \boldsymbol{W}_s^V \cdot \varphi_s(\boldsymbol{v}_m).\end{gathered} \tag{4}$$

Here, $\boldsymbol{W}_s^Q$, $\boldsymbol{W}_s^K$, $\boldsymbol{W}_s^V$ are three learnable matrices, similar to those given in Eq. (1). Finally, we regard the head of output sequences as global spatial representations of counties. That is, we have $g_{\boldsymbol{\phi}}(\boldsymbol{v}_m) = \boldsymbol{v}_s$, with $\boldsymbol{v}_s \in \mathbb{R}^{T \times d}$ signifying the hidden vector that incorporates global spatial information.

Temporal Transformer. The Temporal Transformer possesses two goals. The first goal aims to learn temporal dependency among the hidden vectors $\boldsymbol{v}_s$, *i.e.*, the outputs of our Spatial Transformer. The second goal is to capture the effect of long-term climate change on crop yields. To achieve the first goal, we add a classification token $\boldsymbol{v}_s^{\text{cls}}$ for learning the global temporal representation, *i.e.*,

$$\varphi_t(\boldsymbol{v}_s) = [\boldsymbol{v}_s^{\text{cls}};\ \boldsymbol{v}_s^1;\ \boldsymbol{v}_s^2;\ \ldots\ ;\ \boldsymbol{v}_s^T] + \mathbf{E}_{\text{tmp}}, \tag{5}$$

where $\mathbf{E}_{\text{tmp}} \in \mathbb{R}^{(T+1) \times d}$ is the temporal embedding, similar to the positional embeddings $\mathbf{E}_{\text{pos}}$ expressed in Eq. (3). The novel Temporal Multi-Head Attention (T-MHA) is also devised to capture the impact of long-term historical meteorological parameters on crop yields. Its key idea is to incorporate a relative meteorological bias into each head for similarity computation, motivated by prior studies [6, 38, 33]. Mathematically, our T-MHA can be expressed by

$$\begin{gathered}\text{T-MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}/\sqrt{d} + \pi_t(\boldsymbol{y}_l))\boldsymbol{V}, \\ \boldsymbol{Q} = \boldsymbol{W}_t^Q \cdot \varphi_t(\boldsymbol{v}_s),\ \boldsymbol{K} = \boldsymbol{W}_t^K \cdot \varphi_t(\boldsymbol{v}_s), \\ \boldsymbol{V} = \boldsymbol{W}_t^V \cdot \varphi_t(\boldsymbol{v}_s).\end{gathered} \tag{6}$$

Here, $\boldsymbol{y}_l$ represents the long-term meteorological parameters, and $\pi_t(\cdot)$ is a linear projection layer. Similarly, $\boldsymbol{W}_t^Q$, $\boldsymbol{W}_t^K$, $\boldsymbol{W}_t^V$ are three learnable matrices. In summary, our Temporal Transformer $h_{\boldsymbol{\psi}} : (\mathbb{R}^{T \times d},\ \mathbb{R}^{T \times N_2 \times d_y}) \to \mathbb{R}^d$ can be defined as $h_{\boldsymbol{\psi}}(\boldsymbol{v}_s, \boldsymbol{y}_l) = \boldsymbol{v}_t$, with $\boldsymbol{v}_t \in \mathbb{R}^d$ being the hidden vector that incorporates both global temporal information and the impact of climate change on crops simultaneously.

Finally, the hidden vector $\boldsymbol{v}_t$ is fed into a linear classifier for crop yield predictions, *i.e.*, $\hat{\boldsymbol{z}} = \boldsymbol{W}^T\boldsymbol{v}_t + \boldsymbol{b}$, with $\boldsymbol{W}$ and $\boldsymbol{b}$ respectively denoting the weights and the bias for the classifier, and $\hat{\boldsymbol{z}} \in \mathbb{R}^{d_z}$ indicating the crop yield prediction.

5. Experiments

We conduct experiments for the county-level crop yield predictions across four U.S. states, *i.e.*, Mississippi (MS), Louisiana (LA), Iowa (IA), and Illinois (IL). Four types of crops, *i.e.*, corn, cotton, soybean, and winter wheat, are taken into account for performance evaluation.

5.1. Experimental Settings

Datasets. We utilize the Sentinel-2 imagery and the daily HRRR data datasets during the growing season respectively as the remote sensing data and the short-term meteorological data. Meanwhile, the monthly HRRR data from the previous three years are used as the long-term meteorological parameters.

Compared Approaches. We compare our proposed MMST-ViT to three DL-based approaches, with one, *i.e.*, **ConvLSTM** [43], developed for spatial-temporal prediction, and the other two, *i.e.*, **CNN-RNN** [28] and **GNN-RNN** [15], developed for crop yield predictions. Minor rectifications are made to them so that they can admit the Sentinel-2 imagery and HRRR datasets as their inputs. Other hyperparameters, unless specified otherwise, are set to the values as those reported in their original studies.

Metrics. We take three performance metrics, *i.e.*, **Root Mean Square Error (RMSE)**, **R-squared (R^2)**, and **Pearson Correlation Coefficient (Corr)**, for comparative evaluation. Notably, a lower value of RMSE or a higher value of R^2 (or Corr) indicates better performance.

Model Size. Our proposed MMST-ViT builds on the top of Vision Transformer (ViT) [12]. Similar to ViT, it consists of a stack of Transformer blocks [50], where each Transformer block includes a Multi-Head Attention (MHA) block and

Table 2: Model details used in our study

Model		Layer	Hidden Size	Head	MLP Size	Others
Multi-Modal Transformer	PVT-T/4	{2, 2, 2, 2}	{64, 128, 320, 512}	{1, 2, 5, 8}	{512, 1024, 1280, 2048}	SR Ratios = {8, 4, 2, 1}
	MM-MHA	2	512	8	2048	Context Size = 9
Spatial Transformer	S-MHA	4	512	3	2048	-
Temporal Transformer	T-MHA	4	512	3	2048	Context Size = 9

Table 3: Overall comparative crop yield predictions for 2021, with the best results shown in bold. Cotton yields are measured in pounds per acre (LB/AC), whereas other crop yields are measured in bushels per acre (BU/AC)

Method	Corn			Cotton			Soybean			Winter Wheat		
	RMSE (↓)	R^2 (↑)	Corr (↑)	RMSE (↓)	R^2 (↑)	Corr (↑)	RMSE (↓)	R^2 (↑)	Corr (↑)	RMSE (↓)	R^2 (↑)	Corr (↑)
ConvLSTM	18.6	0.611	0.782	65.4	0.715	0.846	7.2	0.616	0.785	7.4	0.511	0.715
CNN-RNN	14.6	0.705	0.839	69.5	0.653	0.808	5.8	0.703	0.839	7.5	0.614	0.783
GNN-RNN	14.2	0.730	0.854	58.5	0.647	0.804	5.4	0.748	0.865	6.0	0.621	0.788
Ours	**10.5**	**0.811**	**0.900**	**42.4**	**0.790**	**0.889**	**3.9**	**0.843**	**0.918**	**4.6**	**0.785**	**0.886**

an MLP block, and both of which incorporate the LayerNorm [4] for normalization. Table 2 presents the details of model sizes used in our experiments. Note that “PVT-T/4” refers to the PVT-Tiny model with a patch size of 4, and “SR Ratios” represents the spatial reduction ratio in PVT. Following the original PVT [51], we divide the PVT backbone into four stages and report the corresponding model sizes at each stage in the second row of Table 2. The other transformers are single-stage, similar to the vanilla ViT [12]. Notably, “Context Size” represents the number of meteorological parameters used in our study.

Hyperparameters. Our model is pre-trained for 200 epochs using AdamW [35] with $\beta_1 = 0.9$, $\beta_2 = 0.95$, a weight decay of 0.05, and a cosine decay schedule [34] with an initial learning rate of $1e-4$, and warmup epochs of 20. To perform data augmentation for pre-training, various techniques such as random cropping, random horizontal flipping, random Gaussian blur, and color jittering are employed. These techniques are similar to those used in SimCLR [9]. After pre-training, we fine-tune our proposed MMST-ViT for 100 epochs using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a cosine decay schedule with an initial learning rate of $1e-3$, and warmup epochs of 5.

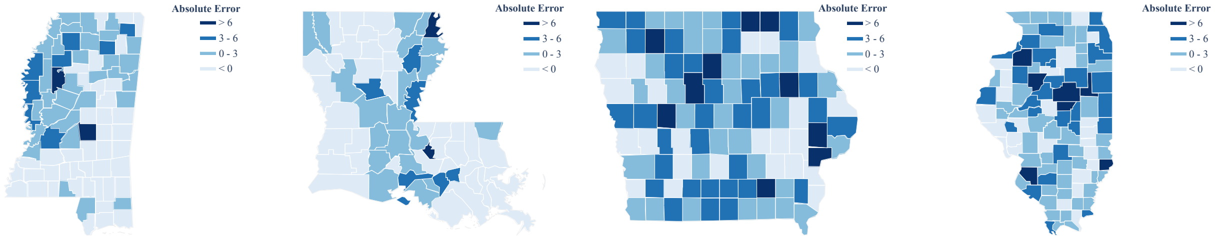
5.2. Comparative Performance Evaluation

We conduct experiments for predicting 2021 crop yields at the county level across four aforementioned U.S. states, with prediction performance measured by the three metrics of RMSE, R^2, and Corr.

Table 3 lists the comparative performance results of our MMST-ViT and its three counterparts, *i.e.*, ConvLSTM, CNN-RNN, and GNN-RNN. Four observations are obtained from the table. First, our approach achieves the best performance under all metrics. In particular, for the soybean yield prediction, our approach achieves the lowest RMSE of 3.9 and the highest R^2 (or Corr) of 0.843 (or 0.918). With the lowest RMSE value, our predicted soybean yields are the closest to the actual amounts. In addition, the highest R^2 and Corr values demonstrate that our predicted soybean yields are best correlated to actual figures. Second, our approach significantly outperforms ConvLSTM, with its RMSE values always lower than those of ConvLSTM markedly, ranging from 2.8 (for winter wheat) to 23.0 (for cotton). The reason is that ConvLSTM overlooks the impact of meteorological parameters on crop growth. Third, CNN-RNN underperforms our approach by 4.1 (for corn), by 27.1 (for cotton), by 1.9 (for soybean), and by 2.9 (for winter wheat), in terms of RMSE, though both models take into account the effect of long-term climate change on crop yields. Because CNN-RNN cannot model the spatial dependency among neighborhood regions. Fourth, MMST-ViT outperforms the most recent state-of-the-art (*i.e.*, GNN-RNN) measurably. Specifically, our RMSE value is lower by 16.1 for cotton, whereas our R^2 and Corr values are better respectively by 0.165 and 0.098 for winter wheat. These results are contributed by leveraging both visual remote sensing and numerical meteorological data in our model, while GNN-RNN only considers the meteorological data.

5.3. Visualizing Crop Yield Prediction Errors

In this section, we conduct experiments to visualize crop yield prediction errors across four U.S. states: Mississippi (MS), Louisiana (LA), Iowa (IA), and Illinois (IL). We use the same experimental settings as those stated in Section 5.2. Figure 4 shows the absolute prediction errors for soybean across counties/parishes in the four states, with navy blue and light blue respectively indicating the high and low absolute errors. Our MMST-ViT model is highly effective in predicting crop yields, with the absolute prediction errors of 57.2% counties below 3 BU/AC. Furthermore, we discover that 95.1% of counties in MS, 92.5% of parishes in LA, 86.8% of counties in IA, and 91.0% of counties in IL achieve decent absolute prediction errors (*i.e.*, ≤ 6 BU/AC). These empirical findings validate the robustness of MMST-



(a) Mississippi (MS) (b) Louisiana (LA) (c) Iowa (IA) (d) Illinois (IL)

Figure 4: Illustration of absolute prediction errors for soybean across four U.S. states. Note that a county/parish with "$<$ 0" indicates its soybean yield data is unavailable.

ViT across various geographic locations.

5.4. Performance of One Year Ahead Predictions

In practice, predicting crop yields well in advance of harvest, can provide valuable information for farmers, agribusinesses, and governments in their agricultural planning and decision-making processes. This information can be used to elevate agricultural resilience and sustainability, make informed financial decisions, and support global food security [23]. A prior study [15] simulated the early prediction scenario by masking partial inputs, successfully making predictions several months before harvest. In this study, we attempt a step further by making crop yield predictions one year before the harvest. In particular, we leverage remote sensing and short-term meteorological data during the growing season in the current year to predict crop yields in the next year. For example, the Sentinel-2 imagery and daily HRRR data during the growing season in 2020 are utilized as remote sensing data and the short-term meteorological data for predicting crop yields in 2021.

Figure 5 illustrates the experimental results. It is observed that our MMST-ViT outperforms three baseline models, *i.e.*, ConvLSTM, CNN-RNN, and GNN-RNN, for the predictions of one year ahead. Among all crops, MMST-ViT achieves the lowest RMSE values ranging from 4.7 (for soybean) to 50.2 (for cotton), the highest R^2 values ranging from 0.711 (for winter wheat) to 0.792 (for soybean), and the highest Corr values ranging from 0.843 (for winter wheat) to 0.890 (for soybean). Additionally, compared to its in-season prediction results (see the 6th row of Table 3), the one year ahead prediction outcomes of our approach are just slightly inferior, degraded respectively by 8.6%, of 15.5%, of 17.0%, and of 14.6% in terms of RMSE for corn, cotton, soybean, and winter wheat. The empirical results confirm the robustness of MMST-ViT.

Interestingly, we also observe that the prediction results for the cotton have a high RMSE value but not a low R^2 (or Corr) value. This is because the cotton yield is measured by pounds per acre (LB/AC), with a high standard deviation value of 250.1, while other crop yields are measured by bushels per acre (BU/AC), with low standard deviation values (*e.g.*, 11.8 for soybean). In scenarios with high variance data, the RMSE value may worsen due to larger residuals, while the R^2 value improves due to the increased proportion of explained variance.

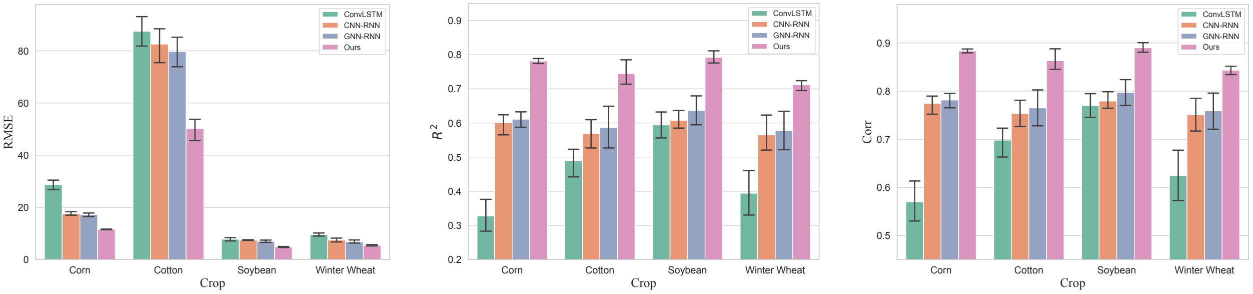
Table 4: Ablation studies for different components, with five scenarios considered and the last row listing the results of MMST-ViT

Component		Corn		Soybean	
		RMSE (↓)	Corr (↑)	RMSE (↓)	Corr (↑)
Temporal Transformer	w/o long-term	14.5	0.843	6.2	0.850
	w/o T-MHA	15.7	0.820	5.6	0.839
Spatial Transformer	w/o S-MHA	13.5	0.856	5.6	0.849
Multi-Modal Transformer	w/o short-term	13.6	0.839	6.1	0.845
	w/o image	15.2	0.809	6.8	0.822
MMST-ViT	-	10.5	0.900	3.9	0.918

5.5. Ablation Studies

Key Components. We next conduct experiments to show how each key component in our MMST-ViT affects prediction performance. Table 4 lists the results of our ablation studies under five different scenarios, with one shown in a row. Here, "w/o long-term" indicates masking the long-term meteorological data, and "w/o T-MHA" represents the absence of our Temporal Transformer and instead resorting to the average pooling to obtain the global temporal representation. Similarly, "w/o S-MHA" denotes utilizing the average pooling to obtain the global spatial representation. The scenarios of "w/o short-term" and "w/o image" represent respectively masking the short-term meteorological data and the remote sensing data from the Multi-Modal Transformer. Note that masking either data makes our model unable to conduct MM-MHA, following Eq. (1). The results of our complete MMST-ViT outcomes are shown in the last row of Table 4.

Three observations can be made from Table 4. First, the long-term meteorological data degrades the Corr value by 0.057 (or 0.068) for corn (or soybean). This validates that crop yield prediction accuracy can benefit from histor-



(a) RMSE

(b) R^2

(c) Corr

Figure 5: Illustration of the performance for predictions of one year ahead under (a) RMSE, (b) R^2, and (c) Corr, with the cotton yield measured by LB/AC and other crop yields measured by BU/AC.

ical meteorological data, which dictates long-term climate change. Second, the absence of either the Temporal Transformer or the Spatial Transformer lowers the Corr value by 0.080 (for corn) and by 0.069 (for soybean), respectively. Third, unable to conduct MM-MHA as the result of masking the short-term meteorological data or the satellite images degrades prediction performance greatly. For example, masking the short-term meteorological data (or satellite images) causes the Corr value degradation by 0.096 (or 0.073) for soybean. This statistical evidence exhibits the importance of our Multi-Modal Transformer for capturing the impact of meteorological parameters on crop growth.

Pre-training Techniques. We conduct experiments to explore the impact of our proposed multi-modal self-supervised pre-training, *i.e.*, Eq. (2), on the crop yield prediction. We consider three scenarios, *i.e.*, MMST-ViT without pre-training, with the pre-training technique described in SimCLR [9], and with our multi-modal pre-training. Table 5 presents the prediction performance outcomes. The table reveals that our MMST-ViT (w/ our multi-modal pre-training) significantly outperforms its counterpart (w/o pre-training), exhibiting a lower RMSE value by 2.7 (or 1.2) and a larger Corr value by 0.043 (or 0.043) for corn (or soybean), due to two reasons. First, Vision Transformer (ViT)-based models are prone to overfitting [12], but our multi-modal self-supervised learning can significantly mitigate this issue by leveraging both visual remote sensing data and numerical meteorological data for pre-training. Second, our multi-modal self-supervised pre-training may can capture the impact of meteorological parameters on the crops. In contrast, pre-training our MMST-ViT with the SimCLR technique only achieves marginal performance improvement (compared to the one w/o pre-training), achieving the Corr improvement of 0.001 (0.875 v.s. 0.876) for soybean. This is because the SimCLR technique fails to leverage numerical meteorological parameters for pre-training. These empirical results validate the necessity and importance of our proposed multi-modal pre-training for accurate crop yield predictions.

Table 5: Ablation studies for different pre-training techniques, with three scenarios considered

Method	Corn		Soybean	
	RMSE (↓)	Corr (↑)	RMSE (↓)	Corr (↑)
w/o pretraining	13.2	0.857	5.1	0.875
w/ SimCLR	12.9	0.857	4.9	0.876
w/ multi-modal pretraining	10.5	0.900	3.9	0.918

6. Conclusion

This paper has proposed the Multi-Modal Spatial-Temporal Vision Transformer (MMST-ViT), a climate change-aware deep learning approach for predicting crop yields at the county level across the United States. MMST-ViT comprises three key components of a Multi-Modal Transformer, a Spatial Transformer, and a Temporal Transformer. Three innovative Multi-Head Attention (MHA) mechanisms are introduced, one for each component. As a result, our MMST-ViT can leverage both the visual remote sensing data and the numerical meteorological data for capturing the impact of short-term growing season weather variations and long-term climate change on crop yields. Additionally, a novel multi-modal contrastive learning technique has been developed, able to effectively pre-train our model without the need of human supervision. We have conducted extensive experiments on 200+ counties/parishes located in 4 U.S. states, with the results demonstrating that our proposed MMST-ViT substantially outperforms its state-of-the-art counterparts consistently under three performance metrics of interest.

Acknowledgments

This work was supported in part by NSF under Grants 2019511, 2146447, and in part by the BoRSF under Grants LEQSF(2019-22)-RD-A-21 and LEQSF(2021-22)-RD-D-07. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agencies.

References

[1] Faezeh Akhavizadegan, Javad Ansarifar, Lizhi Wang, Isaiah Huber, and Sotirios V Archontoulis. A time-dependent parameter estimation framework for crop modeling. *Scientific reports*, 2021.

[2] Javad Ansarifar, Lizhi Wang, and Sotirios V Archontoulis. An interaction regression model for crop yield prediction. *Scientific reports*, 2021.

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

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

[6] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning (ICML)*, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

[8] Liyan Chen, Weihan Wang, and Philippos Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[10] Minghan Cheng, Xiyun Jiao, Lei Shi, Josep Penuelas, Lalit Kumar, Chenwei Nie, Tianao Wu, Kaihua Liu, Wenbin Wu, and Xiuliang Jin. High-resolution crop yield and water productivity dataset generated using random forest and remote sensing. *Scientific Data*, 2022.

[11] Karine Chenu, John Roy Porter, Pierre Martre, Bruno Basso, Scott Cameron Chapman, Frank Ewert, Marco Bindi, and Senthold Asseng. Contribution of crop models to adaptation in wheat. *Trends in plant science*, 2017.

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

[13] Nicola Falco, Haruko M Wainwright, Baptiste Dafflon, Craig Ulrich, Florian Soom, John E Peterson, James Bentley Brown, Karl B Schaettle, Malcolm Williamson, Jackson D Cothren, et al. Influence of soil heterogeneity on soybean plant development and crop yield evaluated using time-series of uav and ground-based geophysical imagery. *Scientific reports*, 2021.

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

[15] Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P. Gomes. A GNN-RNN approach for harnessing geospatial and temporal information: Application to crop yield prediction. In *AAAI*, 2022.

[16] Niketa Gandhi, Owaiz Petkar, and Leisa J Armstrong. Rice crop yield prediction using artificial neural networks. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 2016.

[17] Vivien Sainte Fare Garnot and Loïc Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *International Conference on Computer Vision (ICCV)*, 2021.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] Yi He, Fudong Lin, Xu Yuan, and Nian-Feng Tzeng. Interpretable minority synthesis for imbalanced classification. In *IJCAI*, 2021.

[21] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[23] Takeshi Horie, Masaharu Yajima, and Hiroshi Nakagawa. Yield forecasting. *Agricultural systems*, 1992.

[24] HRRR. The high-resolution rapid refresh (hrrr), 2022.

[25] James W Jones, John M Antle, Bruno Basso, Kenneth J Boote, Richard T Conant, Ian Foster, H Charles J Godfray, Mario Herrero, Richard E Howitt, Sander Janssen, et al. Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agricultural systems*, 2017.

[26] Saeed Khaki, Hieu Pham, and Lizhi Wang. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Scientific Reports*, 2021.

[27] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 2019.

[28] Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 2020.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2012.

[30] Zhengfeng Lai, Sol Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee

Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2023.

[31] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] Fudong Lin, Xu Yuan, Lu Peng, and Nian-Feng Tzeng. Cascade variational auto-encoder for hierarchical disentanglement. In *Conference on Information & Knowledge Management (CIKM)*, 2022.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.

[34] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[36] Spyridon Mourtzinis, Paul D Esker, James E Specht, and Shawn P Conley. Advancing agricultural research using machine learning algorithms. *Scientific reports*, 2021.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020.

[39] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 2022.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[41] Sentinel-Hub. Sentinel hub process api, 2022.

[42] Mohsen Shahhosseini, Guiping Hu, Isaiah Huber, and Sotirios V Archontoulis. Coupling machine learning and crop modeling improves crop yield prediction in the us corn belt. *Scientific reports*, 2021.

[43] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Neural Information Processing Systems (NeurIPS)*, 2015.

[44] Fulu Tao, Masayuki Yokozawa, and Zhao Zhang. Modelling the impacts of weather and climate variability on crop productivity over a large area: a new process-based model development, optimization, and uncertainties analysis. *agricultural and forest meteorology*, 2009.

[45] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning (ICML)*, 2021.

[47] Gabriel Tseng, Ivan Zvonkov, Catherine Nakalembe, and Hannah Kerner. Cropharvest: a global satellite dataset for crop type classification. *Neural Information Processing Systems (NeurIPS)*, 2021.

[48] Matteo Turchetta, Luca Corinzia, Scott Sussex, Amanda Burton, Juan Herrera, Ioannis N Athanasiadis, Joachim M Buhmann, and Andreas Krause. Learning long-term crop management strategies with cyclesgym. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[49] USDA. The united states department of agriculture (usda), 2022.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision (ICCV)*, pages 548–558, 2021.

[52] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media*, 2022.

[53] Xiaocui Wu, Xiangming Xiao, Jean Steiner, Zhengwei Yang, Yuanwei Qin, and Jie Wang. Spatiotemporal changes of winter wheat planted and harvested areas, photosynthesis and grain production in the contiguous united states from 2008–2018. *Remote Sensing*, 2021.

[54] Rongting Xu, Hanqin Tian, Shufen Pan, Stephen A Prior, Yucheng Feng, William D Batchelor, Jian Chen, and Jia Yang. Global ammonia emissions from synthetic nitrogen fertilizer applications in agricultural systems: Empirical and process-based estimates and uncertainty. *Global change biology*, 2019.

[55] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *International Conference on Computer Vision (ICCV)*, 2021.

[56] Yihe Zhang, Xu Yuan, Sytske K. Kimball, Eric Rappin, Li Chen, Paul J. Darby III, Tom Johnsten, Lu Peng, Boisy Pitre, David Bourrie, and Nian-Feng Tzeng. Precise weather parameter predictions for target regions via neural networks. In *ECML-PKDD. Applied Data Science Track*, 2021.