# ViST: A Ubiquitous Model with Multimodal Fusion for Crop Growth Prediction

JUNSHENG LI, Department of Computer Science and Technology, Harbin Institute of Technology, China
LING WANG, Department of Computer Science and Technology,Harbin Institute of Technology, China
JIE LIU, Department of Computer Science and Technology,Harbin Institute of Technology, China
JINSHAN TANG, Health Informatics, College of Public Health,George Mason University, USA

Crop growth prediction can help agricultural workers to make accurate and reasonable decisions on farming activities. Existing crop growth prediction models focus on one crop and train a single model for each crop. In this paper, we will develop a ubiquitous growth prediction model for multiple crops, aiming to train a single model for multiple crops. A ubiquitous vision and sensor transformer(ViST) model for crop growth prediction with image and sensor data is developed to achieve the goals. In the proposed model, a cross-attention mechanism is proposed to implement the fusion of multimodal feature maps for reducing the computational cost and balancing the interactive effects between features. For training the model, we combine the data from multiple crops to train a single (ViST) model. A sensor network system is constructed for the data collection on the farm that plants rice, soybean, and maize. Experiment results show that the proposed ViST model has an excellent ubiquitous ability for crop growth prediction with multiple crops.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: crop growth prediction,ubiquitous model, multimodal learning, transformer module,cross-attention mechanism

**ACM Reference Format:**
Junsheng Li, Ling Wang, Jie Liu, and Jinshan Tang. 2018. ViST: A Ubiquitous Model with Multimodal Fusion for Crop Growth Prediction. *J. ACM* 37, 4, Article 111 (August 2018), 22 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

With the development of new technologies such as the Internet of things, big data, and artificial intelligence (AI), modern agriculture has been dramatically changed. One major task in modern agriculture is to predict crop growth. Crop growth prediction is used as the weather vane for agricultural activities. Accurate short-range prediction of crop growth can help farmers manage fertilization, irrigation, and pesticide spraying effectively. With the effective management of these agricultural activities, human and material resources can be reduced as much as possible, and the final yields and economic benefits can be significantly improved. It can also reduce the use of pesticides that pollute the environment and thus protect the ecological environment.

Authors' addresses: Junsheng Li, 22s103187@stu.hit.edu.cn, Department of Computer Science and Technology, Harbin Institute of Technology, No. 92, Xidazhi Street, Nangang District, Harbin, Heilongjiang Province, China, 150000; Ling Wang, wangling@hit.edu.cn, Department of Computer Science and Technology,Harbin Institute of Technology, No. 92, Xidazhi Street, Nangang District, Harbin, Heilongjiang Province, China, 150000; Jie Liu, jieliu@hit.edu.cn, Department of Computer Science and Technology,Harbin Institute of Technology, No. 92, Xidazhi Street, Nangang District, Harbin, Heilongjiang Province, China, 150000; Jinshan Tang, jtang25@gmu.edu, Health Informatics, College of Public Health,George Mason University, Fairfax, VA, USA, 22033.


Manuscript submitted to ACM

There are many existing crop models for crop growth and yield prediction [1–3]. However, These models are designed for a specific crop and specific region and condition, which are not universal models. These models cannot be adaptive to any specific crop in different regions. The parameters and driving variables of the models are derived from the situation of a particular location, which can be measured and available under ideal conditions. Due to the inherent soil heterogeneity and the influence of farming methods on soil properties, the measured parameters will also be different. Because the biological system is too complex and many processes involved have to be fully understood, a universal model based on biology is different to build. A data-driven model with Neural Network makes it possible to build a universal model. Under the depth learning model, any dynamic systems, including crop systems, can be approximated through the design of network depth.

In the past, a lot of research has focused on multi-scale crop images collected by UAV or satellite remote sensing for crop growth and yield prediction [4–6] . These image data reflect the phenotype characteristics of crops. The dynamic changes of crop phenotypes, such as the Leaf Area Index, are used to predict crop growth for a large region. Some researchers combine crop spectral data and soil and meteorological data for crop growth prediction [7–11] . However, these data-driven methods still need to be universal. The results are strongly related to collecting high-precision crop parameters and input data of crop environment. The processing of the data collection is costly, leading to the use for farm management hardly.

To approximate the objective function, a model has to be built to apply to multimodal and multi-dimension data. Most existing multimodal mechanisms exist in automotive drive and medical fields [12, 13]. Our work is the first to fuse crop RGB image and sensor data for low-cost crop growth prediction on farms. The multimodal fusion process fuses information from two or more modalities to realize information complementation and broaden the coverage of information contained in the input data. Different from other fields, the estimation model for crop growth by multimodal fusion has more challenges in generalization as each crop has its character. To solve this problem, we propose the ViST (Vision and Sensor Transformer) model, which can realize efficient information fusion for accurate prediction of crop growth.

Besides the fusion of multiple data resources for crop growth, we also investigate the possibility of hybrid training, which aims at developing a single network to predict the crop growth of multiple crops. In the past, many models for crop growth prediction were developed. However, all of these models were trained using a single crop, which means each crop needs a single model. On a farm, there are generally many crops. If each crop needs a trained model, training is time-consuming and complex for farmers to use. Thus, this paper aims to develop a ubiquitous model that could be used for multiple crops. A sensor network system is constructed for the data collection on the farm that plants rice, soybean, and maize.

The main contributions of this paper are as follows:

• We proposed a ViST (Vision and Sensor Transformer) model for crop growth, which can efficiently utilize a multimodal data fusion mechanism.

• We also investigated the possibility of hybrid training, which aims at developing a single network to predict crop growth of multiple crops.

• The proposed model was compared with other existing models using real data from farms we collected, and the proposed model can obtain better performance than other existing models.

## 2 RELATED WORK

### 2.1 Single-modality approaches for crop growth prediction

Single-modality approaches for crop growth prediction include two types of approaches: image-only and non-image sensor-only approaches. Image-based remote sensing technology for crop growth prediction is one of the image-only approaches. Image-based remote sensing technology has attracted attention as it can estimate crop growth effectively due to its ability to provide timely, dynamic, and macro-scale observations [14] . With the development of machine learning techniques, image-based remote sensing technology for crop growth has developed further. They are often combined with machine learning techniques to estimate crop growth [15–17]. However, the quality of images acquired through traditional remote sensing is often affected by weather and cloud changes and thus affects the prediction. Besides, remote sensing technology generally has high maintenance and operation costs, affecting its vast uses. Recently, convenient image-based techniques, based on UAV drew the researchers' attention[18–22]. By mounting a camera to a UAV, high spatial resolution images of crops can be acquired and thus can be used for crop growth estimation. These techniques are beneficial for small farms.

Non-image sensor-only approaches are also popular for growth prediction. Because crop growth is affected by climate/weather and soil conditions [23–25], thus meteorological and soil sensors have been widely used to predict crop growth attributes. These non-image sensor-only approaches often use machine learning techniques. Dahikar et al. [26]proposed a crop prediction method by sensing various soil parameters and parameters related to the atmosphere and using ANN for crop yield prediction in rural areas. O'Neal et al. [27]designed a fully connected network to predict maize yield using local crop stage weather data and yield data from 1901 to 1996. Morimoto et al. [28]used a deep learning model to identify changes in citrus sugar and citric acid content based on rainfall and sunshine duration data. Drummond et al. [29]applied feedforward neural networks to estimate nonlinear relationships between soil parameters and crop yields. Kitchen et al. [30]found that neural networks could provide the most accurate empirical model of the data and fit the yield data well to soil and terrain features.

### 2.2 Multimodal learning for crop growth prediction

Image-only and non-image sensor-only approaches have shown impressive results in predicting crop growth [31]. However, regarding the integrity of information expression, the model obtained by a single modality still has certain defects for missing information. One solution is to integrate the representations of these two modalities to take advantage of their complementary advantages in crop growth prediction.

Many deep learning-based approaches have been developed to handle multimodal data [32]. Multimodal machine learning has led to a wide range of applications: from audiovisual speech recognition (AVSR)[33], multimedia content indexing and retrieval [34], understanding human multimodal behavior, multimodal emotion recognition [35], image and video captioning [36], VQA[37], multimedia retrieval [38], to health analytics [39], etc. Huang et al. [40] proved from a theoretical point of view that multimodal learning could fuse the information of single modalities and complement each other so that the final effect of the model is better than that of a single modality.

In agriculture, there is some research on multimodal learning. Dang et al. [41] used DNN with a multilayer feedforward perceptron(MLP) model for crop yield prediction. Chu et al. [42] proposed an end-to-end prediction model for summer and winter rice yield based on MLP deep learning fusion. Two simple MLPS were used to extract spatial and temporal features, and then these two simple MLP models were combined to mine the relationship between features and rice yield. The model maintained stable convergence after 100 iterations. Maimaitijiang et al. [43] tried to use multimodal

data fusion to complete tasks related to crop growth. The combined multimodal information, such as canopy spectral, structural, thermal, and texture features, are extracted. Input-level and middle-level feature fusion by MLP are used to predict crop grain yield.

However, the previous work uses MLP for data fusion in the NN models. The problem is that it is suitable for small-scale learning. When the model scale is enlarged, it will suffer from serious overfitting. Due to the large amount of data in images, it is difficult for MLP to extract features efficiently [44]. In addition, the learning efficiency of fully connected architectures is very low, which has long been confirmed by machine learning experiments. Inefficiency means that more training data are needed to reach a certain level of performance. Many application scenarios cannot provide enough data support, so it is necessary to introduce assumptions to improve the utilization efficiency of limited data. Therefore, the application scenarios of the fully connected architecture are limited, and there are also problems of poor interpretability and robustness.

With the success of Transformers and self-supervised learning, there has been increasing research on cross-modal learning, such as vision-language pre-trained models (VLP)[45], images and lidar fusion[46], Audio set, Epic-Kitchens, and VGGSound classification [47]. The attention adaptively generated by Transformer has good adaptability. Attention can filter out a small amount of important information from a large amount of data, focus on this vital information, and ignore the most unimportant information. The information is critical, and more weight can be assigned. That is, the weight represents the importance of the data [48].

The multimodal fusion process fuses information from two or more modalities to realize information complementation and broaden the coverage of information contained in the input data. However, it inevitably adds much redundant information. Therefore, a more effective way of information fusion and expression is needed.

A multimodal learning method based on Transformer is proposed to complete the task of crop growth prediction. Compared with MLP, Transformer can extract features efficiently on tasks with a large amount of data, such as images and sensors. As a result, unnecessary calculations are reduced, and the efficiency and robustness of the prediction model are improved. A self-attention mechanism is proposed for fusing agricultural sensor and image data with fewer redundancy calculations.

## 3 VISION-AND-SENSOR TRANSFORMER MODEL

The proposed ViST overall network structure is shown in Figure 1. The inputs are crop images and sensor data. In the framework, the MLP module and Linear Projection Flattened Patches module(LPFP) extract feature from sensor and image data, respectively. The transformer encoder is designed for data fusion. Pooler and Linear modules are intended to reduce the dimension of features.

The input of the ViST is sensor and image data. They are processed independently at MLP and LPFP modules, respectively. The features from the two modules are input to one Transformer Encoder for feature fusion. The output of the encoder is given to the Concat module with the output for Multiple Modalities (MM). At the same time, the features are sent separately to the other two Transformer encoders for self-attention mechanics. The output of these two transformer encoders with cross-attention mechanics are Single Modality with Image(SMI) and Single Modal with Sensor(SMS), respectively. The output MM, SMI, and SMS are then input into the Pooler module to reduce the dimension of the features. Finally, the features were input to the linear layer module to output the leaf area index (LAI) value (in the range of [0,1]). The specific details of each module in the framework is described below.

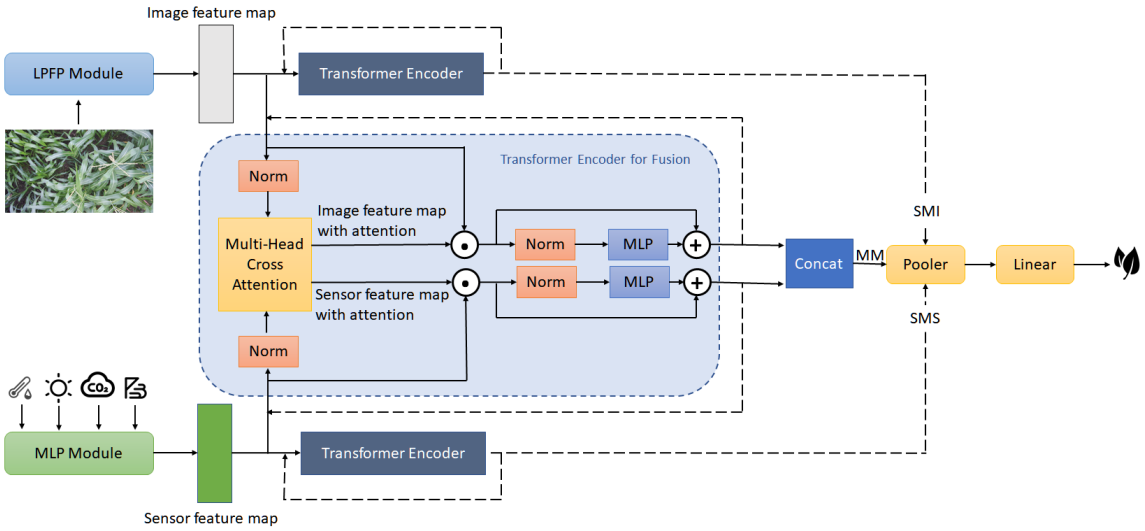



Fig. 1. The framework of ViST for growth prediction.

### 3.1 MLP Module

Generally speaking, image data has three channels for RGB, and each pixel has a value. But sensor data only has dozens of values. Therefore, the number of values for Sensor data is much less than those for image data. Image data will dominate if the two data types are directly integrated; At the same time, the sensor data will not be well expressed.

To solve this problem, sensor data was converted into feature maps. MLP module can amplify the characteristics of the sensor data. The sensor data are composed of weather data and soil data. The data are numerical and have 19 data items. After data preprocessing, the sensor data were arranged into one-dimensional vectors and input into the multilayer feedforward perceptron(MLP) module. The MLP module is a multilayer perceptron that contains 19 neurons in the input layer and 768 neurons in the output layer. The specific structure of the MLP module is shown in Figure 2. The two hidden layers contain 32 and 64 neurons, respectively.

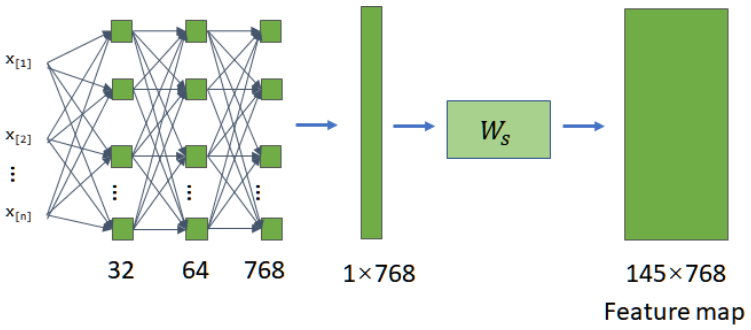



Fig. 2. MLP Module.

After the MLP module outputs a one-dimensional vector containing 768 elements, the size of the vector is 1×768. This 1×768 vector will be converted to a 145 x768 matrix using the following equation

$$M_{out} = W_s \times M_{in} \tag{1}$$

where $M_{in}$ is the 1×768 vector from ML module and the output $M_{out}$ is a 145×768 matrix.$W_s$ is a matrix with a size of 145×1. The element of $W_s$ is obtained by training.

In this paper, the sensor data is mapped to the same dimension as the image features of the LPFP module to facilitate subsequent feature fusion operations. It is worth noting that the sensor features based on the MLP module do not require position information, and the input order of sensor data can be arbitrarily scrambled.

### 3.2 Linear Projection Flattened Patches module

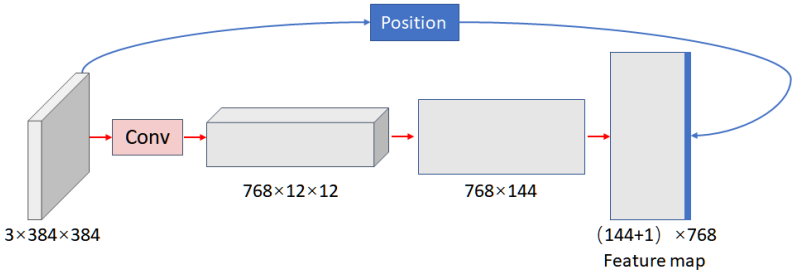



Fig. 3. Linear Projection Flattened Patches module.

The structure of the linear projection flattened patches module is shown in Figure 3. The input of the module is an image with a size of CHANEL×HEIGHT×WIDTH (C×H×W). The 3D tensor corresponds to an RGB image. Before inputting the image, the image is adjusted to the size of 3×384×384, as shown in the input in Figure 3. The image is divided into 768 12×12 image blocks by 768 32×32 convolution cores and 32 convolution operation steps. Each small image is flattened into 144 one-dimensional vectors, and 768 flattened one-dimensional vectors are successively connected into 768×144 vectors. The 768×144 vectors is transposed to the 144×768 vectors to be fused with the feature vector from the sensor. Then, the 144×768 feature vector is spliced with the position information vector to generate 145×768 image features with position information. Linear Projection Flattened Patches divide the image into several equal small images and align the image feature map with the sensor features through matrix transformation.

The image and sensor feature vectors are stacked together to form a sequence of the same dimensions, forming a compact representation of the environment encoded by the global context. The stacked vectors are ready to be input to the next transformer encoder for complete feature fusion.

### 3.3 Transformer Encoder for fusion

The Transformer Encoder for Fusion (TEF) module structure consists of multiple alternating layers of Multi-Head Cross Attention (MHCA, shown in Figure 4) and multiple MLP blocks. Layernorm (LN) is applied before every block in TEF. MLP module in TEF is referenced from ViT. The input of TEF is an Image Feature(IFM) for and Sensor Feature Map(SFM), and the output of TEF is a feature map of fusion. The calculation process is as follows.

$$I_{attn}, S_{attn} = MHCA\,(I_{in}, S_{in}) \tag{2}$$

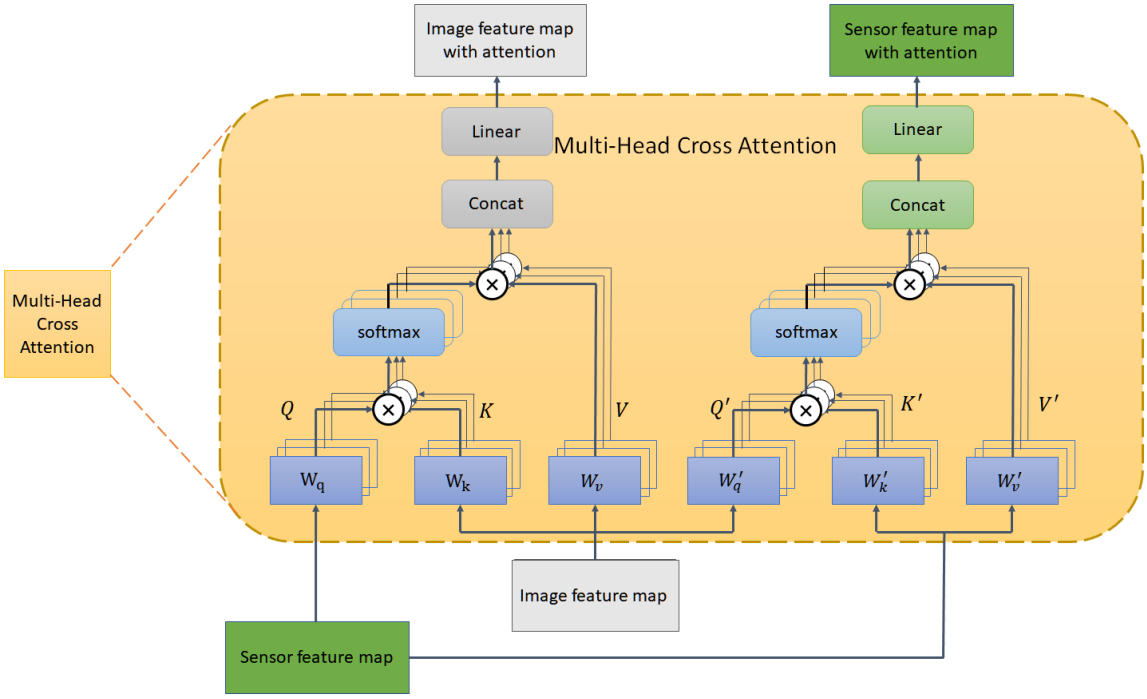



Fig. 4. Multi-Head Cross Attention Module structure.

$$I_{out} = MLP\left(LN\left(I_{in} \cdot I_{attn}\right)\right) + I_{in} \cdot I_{attn} \tag{3}$$

$$S_{out} = MLP\left(LN\left(S_{in} \cdot S_{attn}\right)\right) + S_{in} \cdot S_{attn} \tag{4}$$

where $I_{in}$ is input of Feature map for image. $I_{out}$ is the output of Feature Map. There are 12 layers of calculation for $I_{out}$. The calculation method of $S_{in}$ and $S_{out}$ is the same as $I_{in}$ and $I_{out}$.$I_{attn}$ is attention feature map of Image and $S_{attn}$ is attention feature map of Sensor data.$I_{out}$ and $S_{out}$ are outputs of TEF. Finally, $I_{out}$ , and $S_{out}$ are spliced into the output of MM.

A key idea of this paper is to use MHCA to complete feature fusion. The Cross Attention mechanics is proposed to balance interactive effects between two features from image and sensor data. Two part query the image and sensor feature maps for determining the parameters, respectively. The structure of MHCA is shown in Figure 4. One part takes the SFM as query(Q) and the IFM as the target of query to compute key(K) and value(V). The other part uses IFM as $Q'$, the specific calculation formula is as follows.

$$Q = SW_q, K = IW_k, V = IW_v \tag{5}$$

$$A = softmax\left(\frac{QK^T}{\sqrt{C/h}}\right) V \tag{6}$$

$$Q' = I'W'_q, K' = SW'_k, V' = SW'_v \tag{7}$$

$$A' = softmax\left(\frac{Q'K'^T}{\sqrt{C/h}}\right) V' \tag{8}$$

where $I$ is IFM and $S$ is SFM. $W_q, W_k, W_v \in R^{C\times(C/h)}$ are learnable parameters, $C$ and $h$ are the embedding dimension and number of heads. Since a single modal feature map is used in the query, the computation and memory complexity of generating the attention map ($A$ and $A'$) in cross-attention are linear rather than quadratic as in all-attention. The entire process is more efficient. And it can reduce the risk of overfitting. In addition, multiple heads are used cross attention. Finally, features from multiple heads are spliced, and they are input for Linear to generate the output of MHCA.

### 3.4 Loss function

This paper adopted supervised learning, and actual LAI data manually collected from real farms oversee the learning. The loss function estimates the degree of inconsistency between the predicted value $f(x)$ and the true value $Y$ of the model. It is a non-negative real-valued function, usually expressed by $L(Y, f(x))$. The loss function is the core of the empirical risk function and an essential part of the structural risk function.

In this paper, crop growth prediction is treated as a regression problem. The regression problem solves the prediction of specific values, such as house price prediction, sales prediction, etc. The neural network to solve the regression problem generally has only one output node, and the output value of this node is the predicted value. The loss function used under the regression problem is the mean square error loss function.

$$MSE\left(y, y'\right) = \frac{\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}{n} \tag{9}$$

where $n$ is the number of samples, $y$ is the true value of the leaf area index, and $y'$ is the predicted value of the leaf area index. The mean square error was used to evaluate the accuracy of the model output and help the model update the parameters further.

## 4 EXPERIMENTS AND RESULTS

In this part, we will introduce data collection, data collection equipment, and data format. Image and sensor data are preprocessed after data collection. The performance measures for evaluating the models and the experimental details are described.

### 4.1 Study area and data

Data include weather, soil, crop image, leaf area index, etc. The data were collected at Farm 290, Suibin County, Hegang City, Heilongjiang Province, China. The data acquisition location is shown in Figure 5. The data were collected from three crops(rice, maize, and soybean). The data were collected from three sites for each crop. The data acquisition equipment is shown in Figure 6. Cameras and sensors were placed in each area of the farm. LAI data were collected periodically using a hand-held device, taking multiple measurements each time to prevent accidental errors. The image and sensor data were collected in real-time and uploaded to the cloud server periodically for storage, which is convenient for remote viewing and data analysis.

Image acquisition takes the form of a fixed camera shot. The image format is RGB with a resolution of 3840×2160. Three points are fixed for each crop, and the crops are photographed from a top view. The height is set at 3 meters. The time interval between each shot was two hours, and 12 RGB image data were taken by a single camera daily. The images of rice, soybeans, and maize captured by the camera on the farm are shown in Figure 7(a)(b)(c), respectively.

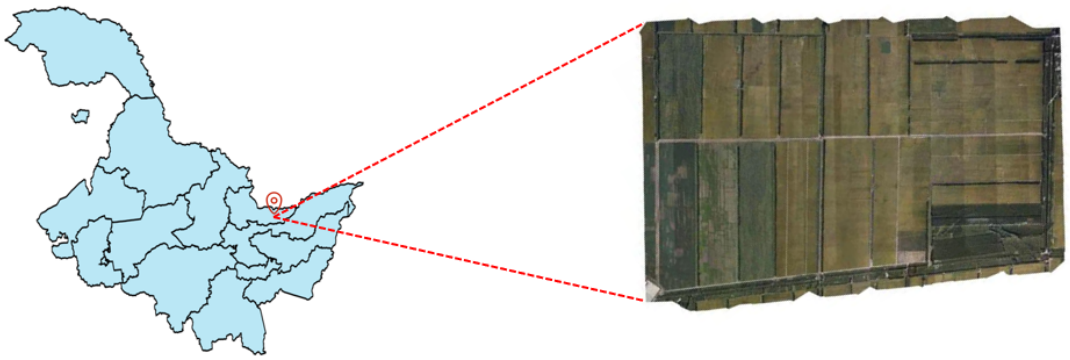


Fig. 5. The location of the farm where data were collected.

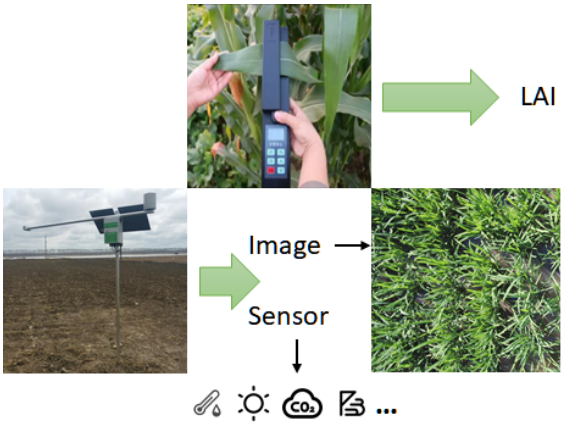



Fig. 6. Data acquisition equipment.

(a) Rice image data. (b) Soybean image data. (c) Maize image data.

Fig. 7. Crop image data.

Table 1. Sensor data

| Data item | Unit |
|---|---|
| Carbon dioxide | PPM |
| Soil temperature | Celsius |
| Soil humidity | % |
| Air temperature | Celsius |
| Air humidity | % |
| Light intensity | Klux |
| Wind direction | degree |
| Wind speed | Meters per second |
| Air pressure | Hpa |
| PM10 | PPM |
| PM2.5 | PPM |

The collected sensor data items are shown in Table 1, which includes eleven sensor data types. The soil sensors were deployed in the ground at 10cm, 20cm, 30cm, 40cm, and 50cm depths, respectively. Sensors were deployed on the top of the IoT Device.

### 4.2 Data preprocessing

**Preprocessing of sensor data:**The numerical differences of each dimension are relatively large for the sensor data. We used the following equation to perform normalization for each size to eliminate the influence of dimensions.

$$y_{ij} = \frac{x_{ij} - min\left(x_j\right)}{max\left(x_j\right) - min\left(x_j\right)} \left(i = 1, 2, \ldots, n; j = 1, 2, \ldots, m\right) \tag{10}$$

where $min(x_j)$ is the minimum value of index $x_j$, and $max(x_j)$ is the maximum value of index $x_j$.

The maximum and minimum values for each metric were saved in a separate file for data preprocessing when they were used for processing actual data.The processed data are shown in Table 2.

**Preprocessing of image data:**Image data were normalized to a specific range, ensuring better convergence in backpropagation. The Z-score method was used to normalize the data of each image.

$$y = \frac{x - \mu}{\sigma} \tag{11}$$

where $\mu$ is the mean value, $\sigma$ is the standard deviation, $x$ is the input data, $y$ is the output. The mean and standard deviation of the three channels of the image were recorded, respectively. The mean value of the normalized image on each channel was 0, and the variance was 1. This method did not apply to the case of a small sample size. Generally speaking, it can be used only when the sample size exceeds 30. The normalized results of image calculation for rice, soybean, and maize are shown in Table 3.

**Processing of LAI data:**Because the LAI data for each crop were collected manually, the time is not continuous, and the amount of data is relatively small. The piecewise cubic Hermite interpolation polynomial (PCHIP) method was used to obtain more LAI data. The LAI data of rice, soybean, and maize were interpolated for each day, as shown in Figure 8(a),(b),(c), respectively. In the figure, the left side is the original data, and the right is the interpolated data. From the curve, it can be seen that the interpolated image is smoother.

Table 2. Sensor input metrics preprocessing

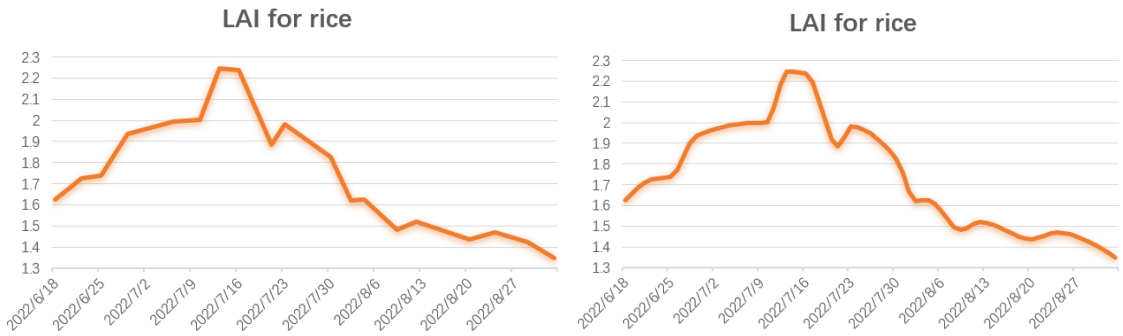
| Indicators | Minimum | Maximum |
|---|---|---|
| Carbon dioxide | 364 | 636 |
| Soil temperature for 10 centimeters depth | 18.1 | 25.1 |
| Soil temperature for 20 centimeters depth | 18.3 | 23 |
| Soil temperature for 30 centimeters depth | 18.3 | 22.1 |
| Soil temperature for 40 centimeters depth | -30 | -30 |
| Soil temperature for 50 centimeters depth | 17.1 | 21.2 |
| Soil humidity for 10 centimeters depth | 46.8 | 80.6 |
| Soil humidity for 20 centimeters depth | 53 | 75.6 |
| Soil humidity for 30 centimeters depth | 55.2 | 79.5 |
| Soil humidity for 40 centimeters depth | 0 | 80.6 |
| Soil humidity for 50 centimeters depth | 67.3 | 81.5 |
| Air humidity | 31 | 98.53 |
| PM10 | 0 | 128 |
| PM2.5 | 0 | 55 |
| Air pressure | 981.1 | 1005.1 |
| Light intensity | 0 | 200 |
| Air temperature | 16.37 | 30.99 |
| Wind direction | 0 | 359.8 |
| Wind speed | 0 | 6.68 |
| LAI | 1.3075 | 1.91 |

Table 3. Image standardized data

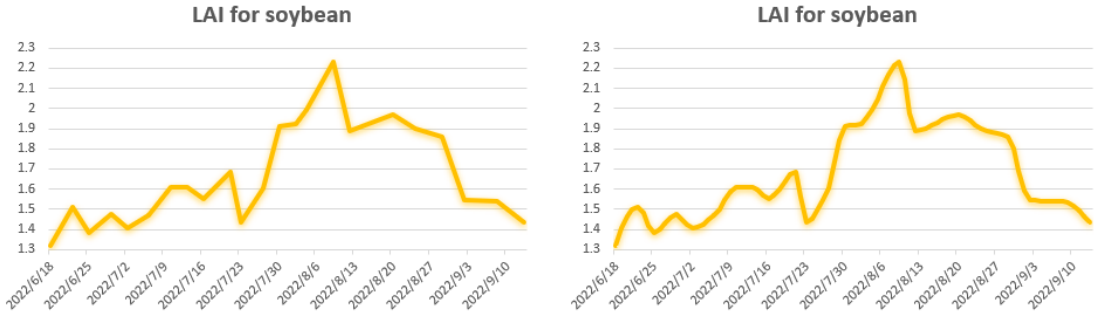
| crop | channel | mean | standard deviation |
|---|---|---|---|
| rice | Channel 1 | 0.4452 | 0.1973 |
| | Channel 2 | 0.5014 | 0.2035 |
| | Channel 3 | 0.4292 | 0.183 |
| soybean | Channel 1 | 0.5 | 0.1517 |
| | Channel 2 | 0.5355 | 0.1641 |
| | Channel 3 | 0.487 | 0.1497 |
| corn | Channel 1 | 0.4452 | 0.1973 |
| | Channel 2 | 0.5014 | 0.2035 |
| | Channel 3 | 0.4292 | 0.183 |

**Data alignment:** The leaf area index data were used as the label of each data to calculate the model error during training and validation. Image data and sensor data were aligned based on time, and all data simultaneously were merged into one training data. The time alignment was based on an hourly basis; there are 24 time points in a day, such as 1:00, 2:00,3:00, etc. There may be multiple training data at the same time point. Since the amount of image data was less than the sensor data, multiple training data records correspond to one image at a one-time point. For the leaf area index data, since it only had one data per day, the leaf area index was labeled with the training data in the unit of days when the data were aligned.

**Training and test sets:**To make the model evaluation more objective and reasonable, the data sets of rice, maize, and soybean were divided into a training set and test set, respectively. 80% of the samples of each crop were used as the




(a) Before and after rice LAI interpolation




(b) Before and after soybean LAI interpolation


(c) Before and after maize LAI interpolation

Fig. 8. LAI data preprocessing.

training set, 10% samples were used as the validation set, and the remaining 10% samples were used as the test set, which was invisible to the model. The only model inference was made during the test process, and model parameters were not updated according to the loss function.

### 4.3 Comparison models

We compared our model with existing DNNF1 and DNNF2, two multimodal models for soybean yield prediction [50]. The inputs to DNNF1 and DNNF2 were multimodal data such as crop canopy spectrum, multi-spectral thermal sensors, and crop surface texture features. For comparison, the inputs to DNNF1 and DNNF2 in our research were modified to make them the same as the input to our ViST model. Figure 9 shows the inputs to the two existing models.

For the DNNF1 model, we used pre-fusion. The images were converted into feature vectors using the Linear Projection Flattened Patches module and flattened into one-dimensional vectors. The sensor data and the one-dimensional vector of the image were spliced and input into the DNNF1 multilayer perceptron. For the DNNF2 model, we used post-fusion. After processing, the data were trained by multilayer perceptron, and the outputs of the two-layer MLP were concatenated and then input into the three-layer MLP of DNNF2. The final output result of the model was an LAI value, which represented the growth of crops.

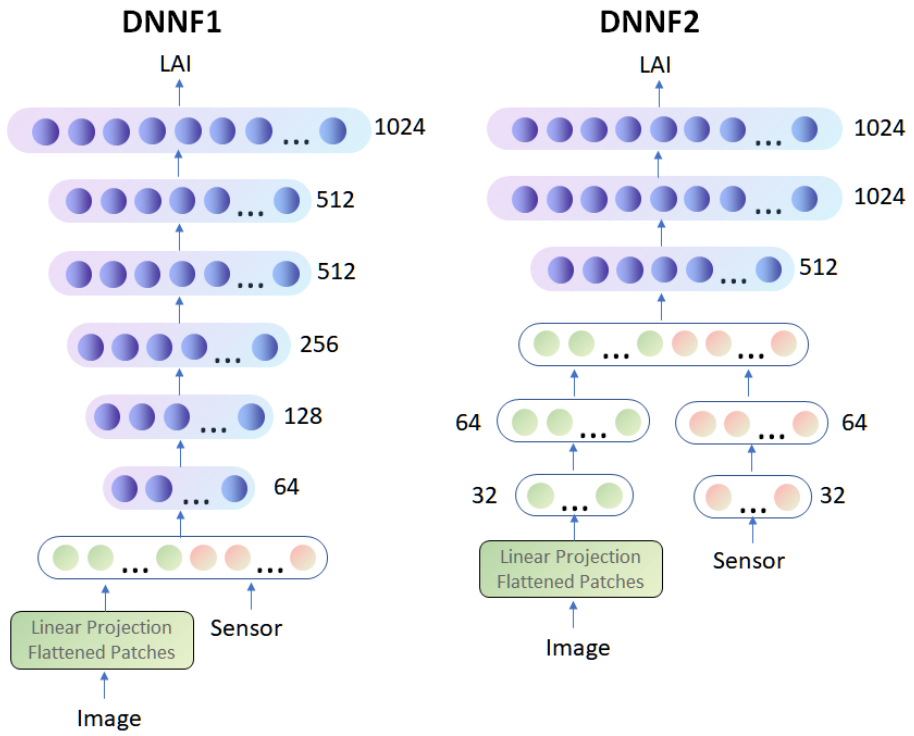



Fig. 9. The structure of the models compared.

### 4.4 Evaluation Metrics

Mean Absolute error (MAE) and mean square error (MSE) were used to evaluate the performance of various models where MSE is the same as the loss function. MAE is the sum of the absolute differences between the target and prediction values. It measures the average error size in a set of predicted values regardless of their direction and ranges from 0 to

∞.

$$MAE = \frac{\sum_i^n \left| y_i - y_i^p \right|}{n} \tag{12}$$

where $y_i^p$ is the predicted value of the model,$y_i$ is the real value, and $n$ is the total number of samples. One advantage of MAE over MSE is that it is less sensitive to outliers. Because MAE calculates the absolute value of the error,$y - f(x)$, the penalty is fixed for any different size.

The Mean Square Error (MSE) is the mean of the squared differences between the model's predicted value $f(x)$, and the actual sample value $y$. The formula is as follows:

$$MSE = \frac{\sum_{i=1}^n \left( f_{x_i} - y_i \right)^2}{n} \tag{13}$$

where $y_i$ and $f_{x_i}, i$ are the true value and the corresponding predicted value for the first sample, and $n$ is the number of samples

The leaf area index (LAI) was selected as an evaluation index of crop growth. The leaf area index (LAI) is a comprehensive index related to individual and group characteristics of crop growth [49]. LAI cannot reflect all individual and group characteristics and must be supplemented by ground monitoring [50]. Since images and sensors can provide adequate information on crop growth, images and sensors can be used as supplements to ground monitoring. After normalization, the leaf area index (LAI) ranges from [0,1].

### 4.5 Experiments and results

In this section, we describe extensive experiments conducted to show the effectiveness of our proposed ViST over existing methods. The experiments were introduced in two parts. First, we describe experimental results for a single crop. Second, we describe the experimental results for multiple crops. We performed ablation studies on ViST according to different input modes. The input modes were image-only data, sensor-only data, and multimodality data, respectively. The performance of each model was compared using MAE and MSE as given in Equations 12,13.

The preprocessed data were used to train each model. The primary hyperparameters of ViST are given in Table 4. Adam optimizer was used, and the weight decay was set to 0.0001. All models were trained using a GeForce RTX 3090 GPU. The cosine annealing strategy was used to adjust the learning rate dynamically. The maximum number of iterations was adjusted according to the sample number and the epoch. Therefore, the learning rate was monotonically decreasing with increasing the number of epochs in the training process. The primary hyperparameters of the ViST model are shown in Table 4. The parameters of each model were initialized using the random normal distribution. MM, SMI and SMS were used as the model's input for the ablation experiment. In all experiments, 80%, 10%, and 10% of the dataset were used as the train, valid, and test data, respectively.

*4.5.1 Experiments and results for a single crop.* In this section, we tested our model for a single crop. We first compared the performance of the ViST model with single modality data and multimodality data(see Section A). Then we compared the performance of the ViST model with other models(see Section B).

**A.** Performance Comparison of ViST model with single modality data and multimodality data

The ViST model was trained for a single crop with three different data input modes: the first was trained using the image-only data; the second was trained using the sensor data only; the third was training using both of image and sensor data. Thus, three trained models were obtained for each crop.

Table 4. Main hyperparameters of the model

| Parameter | Parameter size |
|---|---|
| hidden size | 768 |
| head num | 12 |
| layer num | 12 |
| MLP ratio | 4 |
| drop rate | 0.1 |
| sensor num | 19 |
| epoch | 50 |
| batch size | 32 |
| weight decay | 0.0001 |
| learning rate | 0.0001 |

Figure 10 shows the validation information of the model trained using three different modes for rice. From Figure 10, we can find that the values of MAE and MSE decreased with the increase in the training rounds. When the input mode for the ViST model was multimodality, the lowest values of MSE and MAE were obtained when the model converged. The minimum MSE of the model with multimodality input mode reached 0.001044, which reduced the error by 52.35%, and 84.62% compared to image-only and sensor-only modes. The minimum MAE value of the model with multimodality input mode reached 0.01999. The MAE of the Model with multimodality input mode was also significantly improved. The error of the MAE was reduced by 64.09% and 39.24% compared to sensor-only and image-only modes, respectively. It can be concluded that multimodal fusion helped performance improvement for crop growth prediction. Figure 11 and Figure 12 show the training information of the model trained using three different modes for Soybean and Maize, respectively. We obtained similar results, which show that training results with multimodality input mode had the lowest MAE and MSE values when the network converged. In addition, the training with multimodality input mode had the fastest convergence speed.

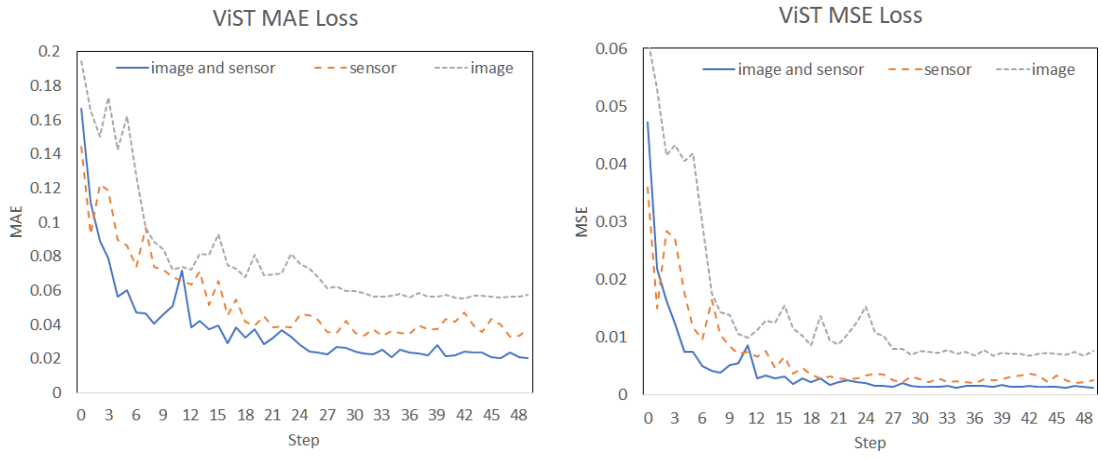



Fig. 10. ViST model for rice training results for different input modes.

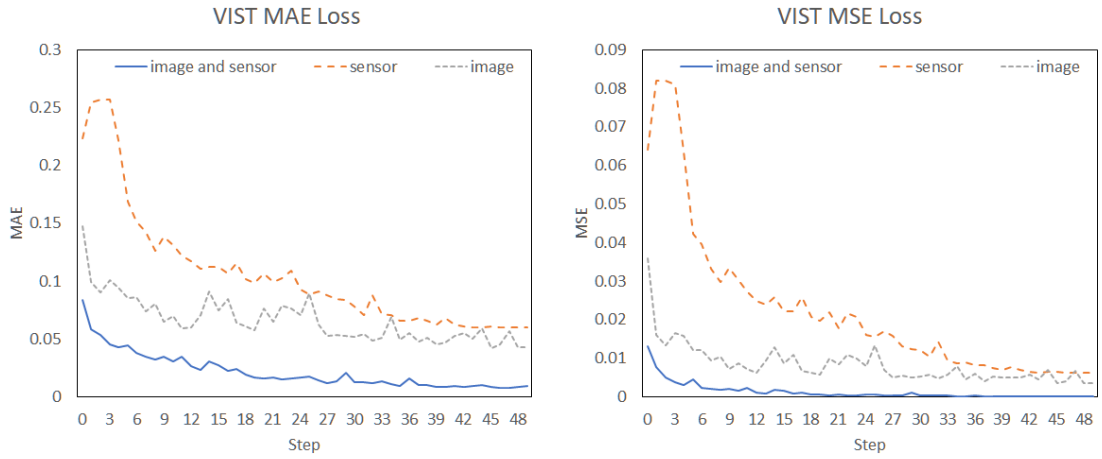



Fig. 11. Soybean ViST model training results for different modes.

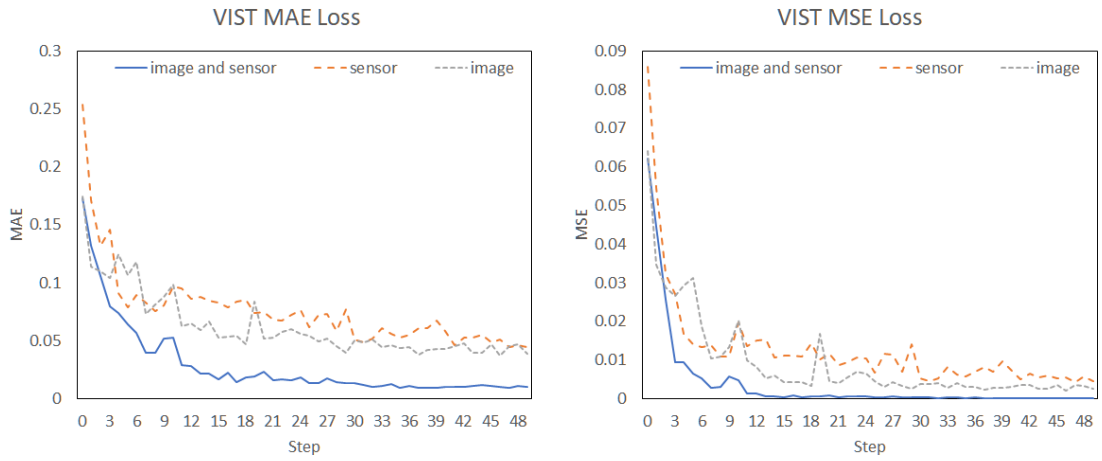



Fig. 12. Maize ViST model training results for different input modes.

After the trained models were obtained, we used the data from the test dataset to test the models. The performance of the ViST model with single modality data and multimodality data for validation dataset and test dataset is shown in table 5. The model obtained using Multimodality data achieved the best performance in three input modes for each crop. Soybean had the best performance. The MSE of soybean for the test dataset reached 0.000158, which reduced the error by 97.47%, and 95.66% in comparison with image data only and sensor data only mode. The error of the MAE was reduced by 84.81% and 78.66% compared with sensor-only data and image-only data, respectively. Obviously, multimodality can bring higher accuracy to the model. The performance of rice was the worst among the three crops.

Table 5. ViST model test results for different input modes.

| Crop | Input mode | Validation | | Test | |
|---|---|---|---|---|---|
| | | MAE | MSE | MAE | MSE |
| rice | Image | 0.05566 | 0.00679 | 0.05764 | 0.007581 |
| | Sensor | 0.0329 | 0.002191 | 0.03707 | 0.002595 |
| | Multimodality | 0.01999 | 0.001044 | 0.02012 | 0.00123 |
| soybean | Image | 0.04185 | 0.003567 | 0.04254 | 0.003638 |
| | Sensor | 0.05961 | 0.006247 | 0.05974 | 0.006251 |
| | Multimodality | 0.007779 | 0.000117 | 0.009076 | 0.000158 |
| maize | Image | 0.03698 | 0.002112 | 0.03844 | 0.002527 |
| | Sensor | 0.04475 | 0.004377 | 0.04455 | 0.004638 |
| | Multimodality | 0.009159 | 0.000223 | 0.01053 | 0.000201 |

But the performance obtained by multimodality data mode also had a good performance. The MSE value of the model with multimodality data mode for rice obtained by the test dataset reached 0.00123, which reduced the error by 52.60%, and 46.30% in comparison with sensor-only data and image-only data. The error of the MAE was reduced by 45.72% and 65.09% in comparison with sensor-only data and image-only data. Therefore, the performance with multimodality data mode is better than that of single mode for the same crop.

**B.Comparison with DNNF1 and DNNF2:** In this Section, we compared the proposed ViST model with two other popular models: DNNF1, and DNNF2. All three models were trained on rice, soybean, and maize data, respectively. Experimental results show that the Loss of the ViST with multimodal input had the lowest value compared with that of DNNF1 and DNNF2 in almost all of the cases. Therefore, the Transformer-based ViST has apparent advantages over DNNF1 and DNNF2.

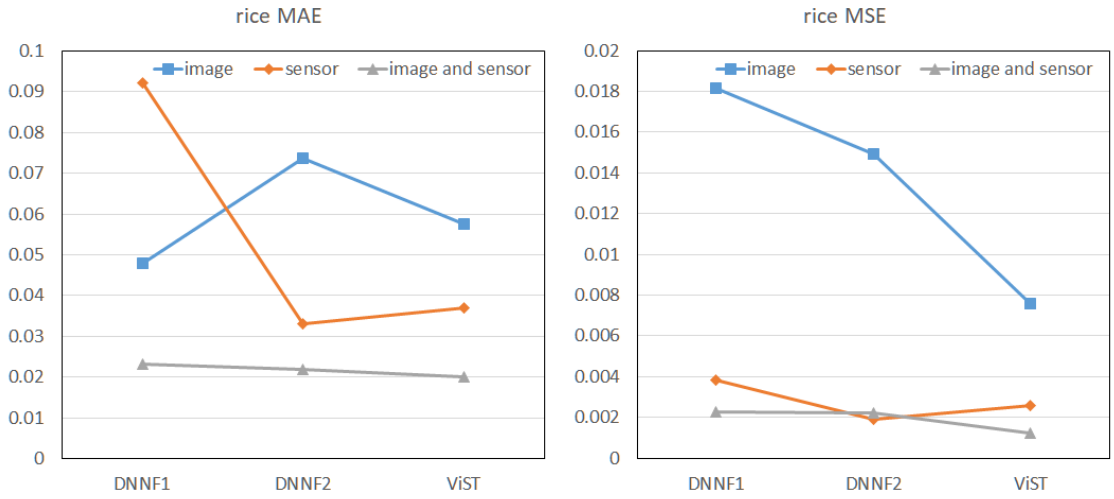



Fig. 13. Comprehensive comparison of three models for rice.

Figure 13 shows the experimental results for rice. Although the performance of the proposed ViST model didn't have the best performance for single input modes, the proposed ViST model had the lowest MAE and MSE values for multiple models mode. From the perspective of MAE, the loss of ViST with multimodal input reaches 0.02012, which reduced the error by 13.39%, and 8.30% compared with DNNF1 and DNNF2, respectively. From the perspective of MSE, the loss of ViST with multimodal input reached 0.00123, which reduces the error by 46.41% and 44.94% compared with DNNF1 and DNNF2, respectively.

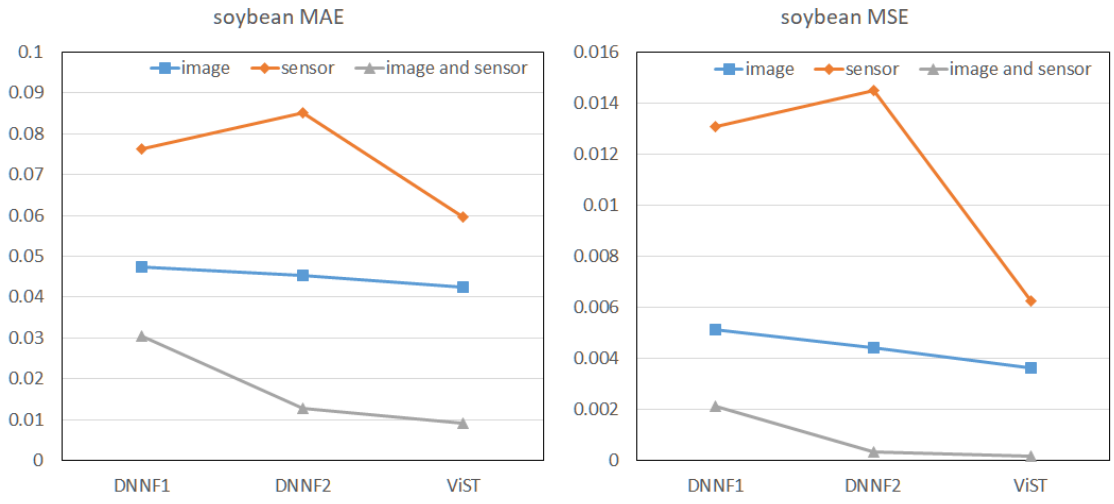



Fig. 14. Comprehensive comparison of three models for soybean.

Figure 14 shows the comprehensive comparison of three models for soybean. ViST performed best under all three input modes. The DNNF2 model outperformed DNNF1 with image-only and multimodal input modes. DNNF1 performed better than DNNF2 with sensor-only mode. For MAE, the error of the ViST reached 0.009076, which reduced the error by 70.20% a 29.26% compared with DNNF1 and DNNF2, respectively. From the perspective of MSE, the error of ViST reached 0.000158, which reduced the error by 92.60%, and 51.29% compared with DNNF1 and DNNF2, respectively. The ViST had a greater improvement in performance than DNNF1 and DNNF2, and DNNF2 was better than DNNF1. The accuracy improvement of the ViST model is relatively significant, and the robustness is better under the condition of multimodal input mode.

The comprehensive comparison of their models for Maize is shown in Figure 15, and the ViST model performed best in all of the three input modes. The accuracy and stability of the DNNF2 model were higher than those of DNNF1. From MAE, the error of ViST reached 0.01053, which reduced the error by 68.04% and 31.98% compared with DNNF1 and DNNF2, respectively. From MSE, the error of ViST reached 0.0002008, which reduced the error by 91.03% and 63.54% compared with DNNF1 and DNNF2, respectively. Therefore, ViST has good precision and stability on Maize data.

*4.5.2 Experiments and results for hybrid network training.* The data from three crops were mixed and used to train the model to test the model's generalization ability for the growth prediction of multiple crops. The results of ViST, DNNF1, and DNNF2 models with the hybrid data set as input are shown in Figure 16. It can be seen that the ViST

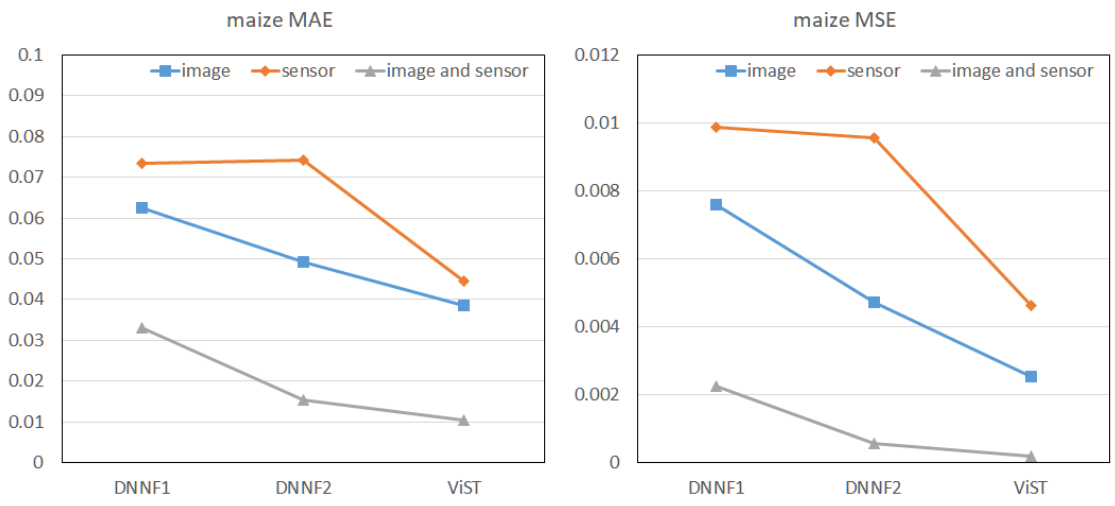



Fig. 15. Comprehensive comparison of three models for maize.

models still had high accuracy compared with the other two models when multiple crops were mixed as the input to train the model. The convergence speed of MAE and MSE loss for the ViST model was significantly faster than that of DNNF1 and DNNF2. Under 50 training iterations, the MSE loss of ViST reached 0.0005848 when the model converged, which reduced the error by about half compared with DNNF1 and DNNF2. For MAE, the ViST model had the least value compared with DNNF1 and DNNF2 when the model converged.

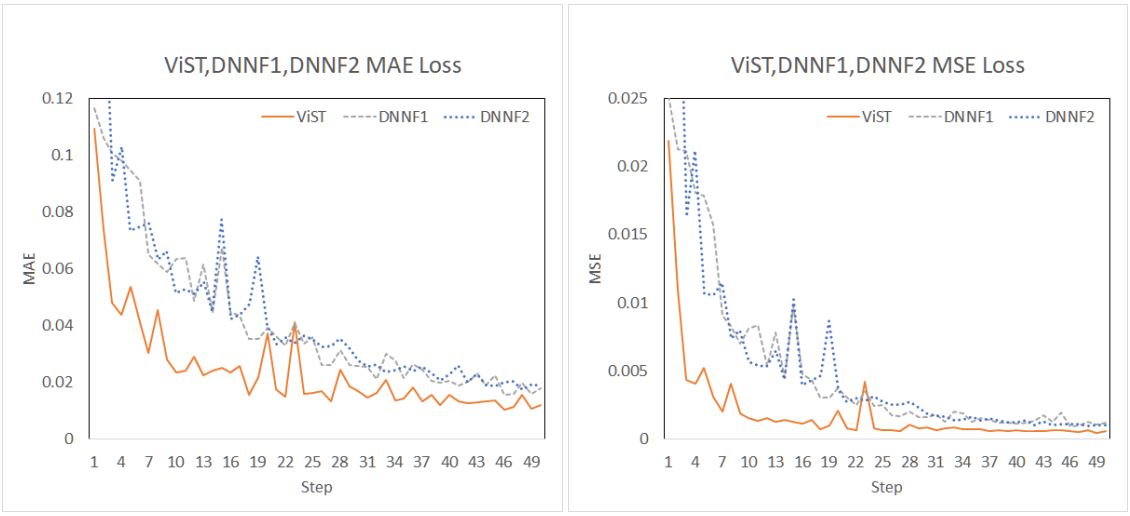



Fig. 16. Comparison of ViST, DNNF1, DNNF2 models for three crops.

The results on the ubiquitous ability of ViST are shown in Table 6. When training and testing were from the same crop, the mean values of MAE were 0.02012, 0.009076, and 0.01053 for rice, maize and soybean, respectively. The MAE value obtained by hybrid training was 0.01198. When training and testing were from the same crop, the mean values of MSE were 0.00123, 0.000158, and 0.000201 for rice, Maize, soybean respectively. The MAE value obtained by hybrid training was 0.000585. From these numbers, we find that the errors obtained by training a hybrid network for multiple crops were similar to the errors obtained by training a single network for a single crop.

Table 6. Comparison Results of ViST model with different crops as input.

| Train data | Test data | MAE | MSE |
|---|---|---|---|
| Rice | Rice | 0.02012 | 0.00123 |
| Maize | Maize | 0.009076 | 0.000158 |
| Soybean | Soybean | 0.01053 | 0.000201 |
| Three crops | Three crops | 0.01198 | 0.000585 |

## 5 CONCLUSION AND FUTURE WORK

This paper proposed ViST, a Transformer-based model for crop growth prediction on a farm. The attention mechanism in the model was used to improve the effect of model fusion. The data of the three crops were trained together as input to the model. The model has high accuracy and good generalization ability. Experiment results show that the model with multimodality data can improve crop growth prediction. The data sets of various crops will be enriched to improve the model's prediction accuracy from the perspective of improving data quality.

We also show that it is possible to train a hybrid network for multiple crops. We will further investigate this issue in the future.

## ACKNOWLEDGMENTS

This research is supported by Heilongjiang NSF funding, No. LH202F022, Heilongjiang research and application of key technologies, No. 2021ZXJ05A03, and New generation artificial intelligent program, No.21ZD0110900 in CHINA.

## REFERENCES

[1] Donald Gaydon and Christian Roth. 2014. *SAC Monograph: The SAARC-Australia Project-Developing Capacity in Cropping Systems Modelling for South Asia.*

[2] N Brisson, C Gary, E Justes, R Roche, B Mary, D Ripoche, D Zimmer, J Sierra, P Bertuzzi, P Burger, F Bussière, Y.M Cabidoche, P Cellier, P Debaeke, J.P Gaudillère, C Hénault, F Maraux, B Seguin, and H Sinoquet. 2003. An Overview of the Crop Model Stics. *European Journal of Agronomy* 18, 3 (2003), 309–332. https://doi.org/10.1016/S1161-0301(02)00110-7

[3] HL Boogaard, CA Van Diepen, RP Rotter, JMCA Cabrera, and HH Van Laar. 1998. WOFOST 7.1; user's guide for the WOFOST 7.1 crop growth simulation model and WOFOST Control Center 1.5. (1998).

[4] Kaili Wang, Keyu Chen, Huiyu Du, Shuang Liu, Jingwen Xu, Junfang Zhao, Houlin Chen, Yujun Liu, and Yang Liu. 2022. New Image Dataset and New Negative Sample Judgment Method for Crop Pest Recognition Based on Deep Learning Models. *Ecological Informatics* 69 (2022), 101620. https://doi.org/10.1016/j.ecoinf.2022.101620

[5] Jibo Yue, Guijun Yang, Qingjiu Tian, Haikuan Feng, Kaijian Xu, and Chengquan Zhou. 2019. Estimate of Winter-Wheat above-Ground Biomass Based on UAV Ultrahigh-Ground-Resolution Image Textures and Vegetation Indices. *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), 226–244. https://doi.org/10.1016/j.isprsjprs.2019.02.022

[6] Mehmet Ozgur Turkoglu, Stefano D'Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. 2021. Crop Mapping from Image Time Series: Deep Learning with Multi-Scale Label Hierarchies. *Remote Sensing of Environment* 264 (2021), 112603. https://doi.org/10.1016/j.rse.2021.112603

[7] Miroslav Trnka, Josef Eitzinger, Pavel Kapler, Martin Dubrovský, Daniela Semerádová, Zdeněk Žalud, and Herbert Formayer. 2007. Effect of Estimated Daily Global Solar Radiation Data on the Results of Crop Growth Models. *Sensors* 7, 10 (Oct. 2007), 2330–2362. https://doi.org/10.3390/s7102330

[8] Tanhim Islam, Tanjir Alam Chisty, and Amitabha Chakrabarty. 2018. A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh. In *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. 1–6. https://doi.org/10.1109/R10-HTC.2018.8629828

[9] Omolola M Adisa, Joel O Botai, Abiodun M Adeola, Abubeker Hassen, Christina M Botai, Daniel Darkey, and Eyob Tesfamariam. 2019. Application of Artificial Neural Network for Predicting Maize Production in South Africa. *Sustainability* 11, 4 (2019), 1145.

[10] Jing Liu, CE Goering, and Lei Tian. 2001. A Neural Network for Setting Target Corn Yields. *Transactions of the ASAE* 44, 3 (2001), 705.

[11] Kanichiro Matsumura, Carlos F Gaitan, Kenji Sugimoto, Alex J Cannon, and William W Hsieh. 2015. Maize Yield Forecasting by Linear Regression and Artificial Neural Networks in Jilin, China. *The Journal of Agricultural Science* 153, 3 (2015), 399–410.

[12] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. 2019. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences* 3, 2 (2019), 162–169.

[13] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. 2022. Multimodal End-to-End Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems* 23, 1 (2022), 537–547. https://doi.org/10.1109/TITS.2020.3013234

[14] Zhongxin Chen, Huajun Tang Jianqiang Ren, Pei Leng Yun Shi, Limin Wang Jia Liu, Yanmin Yao Wenbin Wu, and Hasituya. 2016. Progress and Prospect of agricultural remote sensing research and application. *Journal of Remote Sensing* 20, 05 (2016), 748–767.

[15] Michael D Johnson, William W Hsieh, Alex J Cannon, Andrew Davidson, and Frédéric Bédard. 2016. Crop Yield Forecasting on the Canadian Prairies by Remotely Sensed Vegetation Indices and Machine Learning Methods. *Agricultural and forest meteorology* 218 (2016), 74–84.

[16] Liheng Zhong, Lina Hu, and Hang Zhou. 2019. Deep Learning Based Multi-Temporal Crop Classification. *Remote sensing of environment* 221 (2019), 430–443.

[17] Qi Yang, Liangsheng Shi, Jinye Han, Yuanyuan Zha, and Penghui Zhu. 2019. Deep Convolutional Neural Networks for Rice Grain Yield Estimation at the Ripening Stage Using UAV-based Remotely Sensed Images. *Field Crops Research* 235 (2019), 142–153.

[18] Ulrich Weiss and Peter Biber. 2011. Plant Detection and Mapping for Agricultural Robots Using a 3D LIDAR Sensor. *Robotics and Autonomous Systems* 59, 5 (2011), 265–273. https://doi.org/10.1016/j.robot.2011.02.011

[19] Huilin Tao, Haikuan Feng, Liangji Xu, Mengke Miao, Huiling Long, Jibo Yue, Zhenhai Li, Guijun Yang, Xiaodong Yang, and Lingling Fan. 2020. Estimation of Crop Growth Parameters Using UAV-Based Hyperspectral Remote Sensing Data. *Sensors* 20, 5 (2020), 1296. https://doi.org/10.3390/s20051296

[20] X. Zhou, H. B. Zheng, X. Q. Xu, J. Y. He, X. K. Ge, X. Yao, T. Cheng, Y. Zhu, W. X. Cao, and Y. C. Tian. 2017. Predicting Grain Yield in Rice Using Multi-Temporal Vegetation Indices from UAV-based Multispectral and Digital Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 130 (2017), 246–255. https://doi.org/10.1016/j.isprsjprs.2017.05.003

[21] Maitiniyazi Maimaitijiang, Abduwasit Ghulam, Paheding Sidike, Sean Hartling, Matthew Maimaitiyiming, Kyle Peterson, Ethan Shavers, Jack Fishman, Jim Peterson, Suhas Kadam, Joel Burken, and Felix Fritschi. 2017. Unmanned Aerial System (UAS)-Based Phenotyping of Soybean Using Multi-Sensor Data Fusion and Extreme Learning Machine. *ISPRS Journal of Photogrammetry and Remote Sensing* 134 (2017), 43–58. https://doi.org/10.1016/j.isprsjprs.2017.10.011

[22] Liang Wan, Haiyan Cen, Jiangpeng Zhu, Jiafei Zhang, Yueming Zhu, Dawei Sun, Xiaoyue Du, Li Zhai, Haiyong Weng, Yijian Li, Xiaoran Li, Yidan Bao, Jianyao Shou, and Yong He. 2020. Grain Yield Prediction of Rice Using Multi-Temporal UAV-based RGB and Multispectral Images and Model Transfer – a Case Study of Small Farmlands in the South of China. *Agricultural and Forest Meteorology* 291 (2020), 108096. https://doi.org/10.1016/j.agrformet.2020.108096

[23] Jérôme G Fortin, François Anctil, Léon-Étienne Parent, and Martin A Bolinder. 2011. Site-Specific Early Season Potato Yield Forecast by Neural Network in Eastern Canada. *Precision agriculture* 12, 6 (2011), 905–923.

[24] CA Campbell, RP Zentner, and PJ Johnson. 1988. Effect of Crop Rotation and Fertilization on the Quantitative Relationship between Spring Wheat Yield and Moisture Use in Southwestern Saskatchewan. *Canadian Journal of Soil Science* 68, 1 (1988), 1–16.

[25] David L Ehret, Bernard D Hill, Tom Helmer, and Diane R Edwards. 2011. Neural Network Modeling of Greenhouse Tomato Yield, Growth and Water Use from Automated Crop Monitoring Data. *Computers and electronics in agriculture* 79, 1 (2011), 82–89.

[26] Snehal S Dahikar and Sandeep V Rode. 2014. Agricultural Crop Yield Prediction Using Artificial Neural Network Approach. *International journal of innovative research in electrical, electronics, instrumentation and control engineering* 2, 1 (2014), 683–686.

[27] Monte R O'Neal, Bernard A Engel, Daniel R Ess, and Jane R Frankenberger. 2002. Neural Network Prediction of Maize Yield Using Alternative Data Coding Algorithms. (2002).

[28] T Morimoto, Y Ouchi, M Shimizu, and MS Baloch. 2007. Dynamic Optimization of Watering Satsuma Mandarin Using Neural Networks and Genetic Algorithms. *Agricultural water management* 93, 1-2 (2007), 1–10.

[29] Scott Drummond, Anupam Joshi, and Kenneth A Sudduth. 1998. Application of Neural Networks: Precision Farming. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, Vol. 1. IEEE, 211–215.

[30] NR Kitchen, ST Drummond, ED Lund, KA Sudduth, and GW Buchleiter. 2003. Soil Electrical Conductivity and Topography Related to Yield for Three Contrasting Soil–Crop Systems. *Agronomy journal* 95, 3 (2003), 483–495.

[31] Francisco M Padilla, Marisa Gallardo, M Teresa Peña-Fleitas, Romina De Souza, and Rodney B Thompson. 2018. Proximal Optical Sensors for Nitrogen Management of Vegetable Crops: A Review. *Sensors* 18, 7 (2018), 2083.

[32] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. 2020. A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends. *Knowledge-Based Systems* 194 (2020), 105596.

[33] Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. 1989. Integration of Acoustic and Visual Speech Signals Using Neural Networks. *IEEE Communications Magazine* 27, 11 (1989), 65–71.

[34] Cees GM Snoek and Marcel Worring. 2005. Multimodal Video Indexing: A Review of the State-of-the-Art. *Multimedia tools and applications* 25, 1 (2005), 5–35.

[35] Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. 2021. HEU Emotion: A Large-Scale Database for Multimodal Emotion Recognition in the Wild. *Neural Computing and Applications* 33, 14 (2021), 8669–8685.

[36] Zhou Lei and Yiyong Huang. 2021. Video Captioning Based on Channel Soft Attention and Semantic Reconstructor. *Future Internet* 13, 2 (2021), 55.

[37] Yu Long, Pengjie Tang, Hanli Wang, and Jian Yu. 2021. Improving Reasoning with Contrastive Visual Information for Visual Question Answering. *Electronics Letters* 57, 20 (2021), 758–760.

[38] Rafael Souza, André Fernandes, Thiago SFX Teixeira, George Teodoro, and Renato Ferreira. 2021. Online Multimedia Retrieval on CPU–GPU Platforms with Adaptive Work Partition. *J. Parallel and Distrib. Comput.* 148 (2021), 31–45.

[39] Amir Hossein Yazdavar, Mohammad Saeid Mahdavinejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M Meddar, Annie Myers, Jyotishman Pathak, et al. 2020. Multimodal Mental Health Analysis in Social Media. *Plos one* 15, 4 (2020), e0226248.

[40] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What Makes Multi-modal Learning Better than Single (Provably). arXiv:2106.04538 [cs]

[41] Chaoya Dang, Ying Liu, Hui Yue, JiaXin Qian, and Rong Zhu. 2021. Autumn Crop Yield Prediction Using Data-Driven Approaches:-Support Vector Machines, Random Forest, and Deep Neural Network Methods. *Canadian Journal of Remote Sensing* 47, 2 (2021), 162–181.

[42] Zheng Chu and Jiong Yu. 2020. An End-to-End Model for Rice Yield Prediction Using Deep Learning Fusion. *Computers and Electronics in Agriculture* 174 (2020), 105471. https://doi.org/10.1016/j.compag.2020.105471

[43] Maitiniyazi Maimaitijiang, Vasit Sagan, Paheding Sidike, Sean Hartling, Flavio Esposito, and Felix B. Fritschi. 2020. Soybean Yield Prediction from UAV Using Multimodal Data Fusion and Deep Learning. *Remote Sensing of Environment* 237 (Feb. 2020), 111599. https://doi.org/10.1016/j.rse.2019.111599

[44] Yucheng Zhao, Guangting Wang, Chuanxin Tang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. 2021. A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP. arXiv:2108.13002 [cs]

[45] Jan GPW Clevers and Anatoly A Gitelson. 2013. Remote Estimation of Crop and Grass Chlorophyll and Nitrogen Content Using Red-Edge Bands on Sentinel-2 and-3. *International Journal of Applied Earth Observation and Geoinformation* 23 (2013), 344–351.

[46] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. 2021. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. arXiv:2104.09224 [cs]

[47] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. [n. d.]. Attention Bottlenecks for Multimodal Fusion. ([n. d.]), 14.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[49] Zhiyuan Pei Bangjie Yang. 1999. The definition of crop growth and remote sensing monitoring. *Transactions of the CSAE* 03 (1999), 214–218.

[50] Toby N Carlson and David A Ripley. 1997. On the Relation between NDVI, Fractional Vegetation Cover, and Leaf Area Index. *Remote sensing of Environment* 62, 3 (1997), 241–252.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009