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 to reduce the computational cost and balance 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), 23 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.

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 ubiquitous 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 ubiquitous model based on biology is different to build. A data-driven model with Neural Network makes it possible to build a ubiquitous model. Under the deep learning model, any dynamic systems, including crop systems, can be approximated through network structure design [4, 5].

In the past, much research has focused on multi-scale crop images collected by UAV or satellite remote sensing for crop growth and yield prediction [6–8]. These image data reflect the phenotype characteristics of crops. The dynamic changes of crop phenotypes, such as the Leaf Area Index, predict crop growth for a large region. Some researchers combine crop spectral data and soil and meteorological data for crop growth prediction [9–13]. However, these data-driven methods still need to be ubiquitous. The results are strongly related to collecting high-precision crop parameters and input data of crop environment. The data collection processing is costly, hardly leading to the use for farm management. The leaf area index (LAI) is a comprehensive index related to individual and group characteristics of crop growth [14]. The leaf area index (LAI) is selected as an prediction index of crop growth. LAI cannot reflect all individual and group characteristics and must be supplemented by ground monitoring [15]. Since images and sensors can provide adequate information on crop growth, images and sensors can be used as supplements to ground monitoring.

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 [16, 17]. 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. Training is time-consuming and complex for farmers if each crop needs a trained model. 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 data from farms we collected, and the proposed model can obtain better performance than other current models.

2 RELATED WORK

2.1 Single-modality approaches for crop growth prediction

Single-modality approaches for crop growth prediction include two types of methods: image-only and non-image sensor-only strategies. 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 [18]. 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 [19–21]. 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 [22–26]. 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 [27–29], 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. [30] 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. [31] 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. [32] 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. [33] applied feedforward neural networks to estimate nonlinear relationships between soil parameters and crop yields. Kitchen et al. [34] 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 [35]. 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 [36–38]. Multimodal machine learning has led to a wide range of applications: from audiovisual speech recognition (AVSR)[39], multimedia content indexing and retrieval [40], understanding human multimodal behavior, multimodal emotion recognition [41], image and video captioning [42], VQA[43], multimedia retrieval [44], to health analytics [45], etc. Huang et al. [46] 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. [47] used DNN with a multilayer feedforward perceptron(MLP) model for crop yield prediction. Chu et al. [48] 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. [49] 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 extensive data in images, it is difficult for MLP to extract features efficiently [50]. 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)[51], images and lidar fusion[52], Audio set, Epic-Kitchens, and VGGSound classification [53]. 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 [54].

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. [37] 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[54] is proposed to complete the task of crop growth prediction. Compared with MLP [49], 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. The modular structure of the Transformer model allows for flexible processing of different types of inputs. Different types of features can be inputted into different modules of the Transformer and combined for further processing. This structure makes it easier to expand the model to accommodate different types of data, facilitating the fusion of multimodal data.

A cross-attention mechanism is proposed for fusing agricultural sensor and image data with fewer redundancy calculations. Besides, there are significant differences between the representations and feature spaces of images and sensor data. By using cross-attention mechanisms, images and sensor data can be jointly encoded and effectively complemented based on their common features. This enables the model to handle multimodal data better.

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 features from sensor and image data, respectively. LPFP was introduced by ViT[55]. The transformer encoder is designed for data fusion. Pooler and Linear modules are intended to reduce the dimension of features. The specific details of each module in the framework are described below.

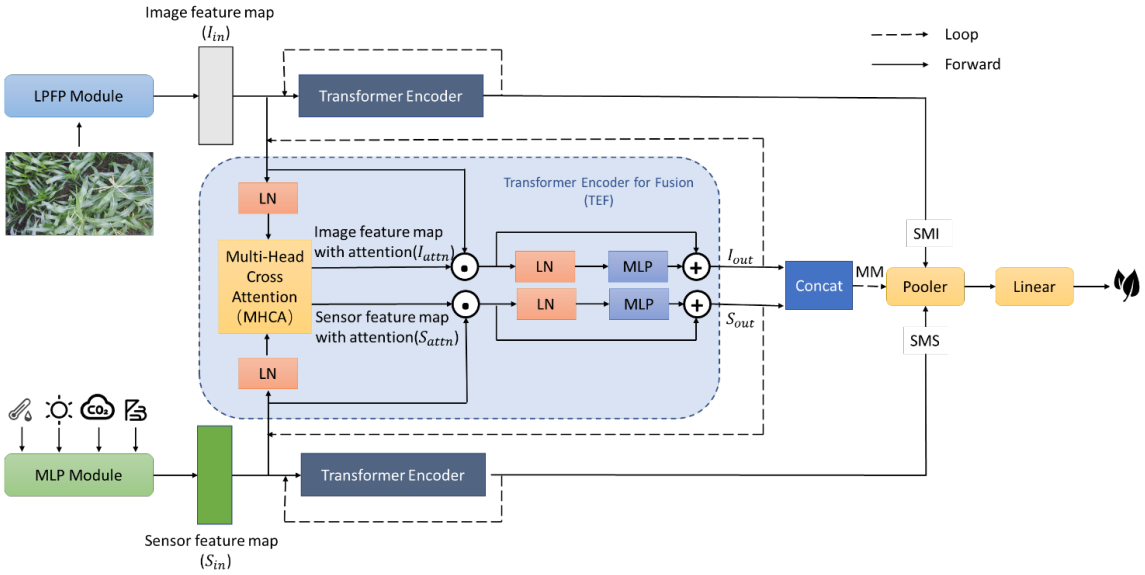


Fig. 1. The framework of ViST for growth prediction. Solid lines represent forward and dashed lines represent loop. The input of the ViST is sensor and image data. They are processed independently at MLP and LPFP modules, respectively. The features from the two modules are input to one Transformer Encoder for feature fusion. The encoder output 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 outputs of these two transformer encoders are Single Modality with Image(SMI) and Single Modal with Sensor(SMS). The results of MM, SMI, and SMS are then input into the Pooler module to reduce the dimension of the features. Finally, the features are input to the linear layer module to output the leaf area index (LAI) value (in the range of [0,1]).

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; Simultaneously, 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 specific structure of the MLP module is shown in Figure 2.

Following the output of the MLP module, a one-dimensional vector containing 768 elements is generated, resulting in a vector size of 1×768. This vector is then transformed into a 145×768 matrix using the following equation:

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

where M_{in} refers to the 1x768 vector generated by the MLP module, which is then transformed into a 145×768 matrix, denoted as M_{out}. W_s is a 145x1 matrix whose elements are obtained through training the model.

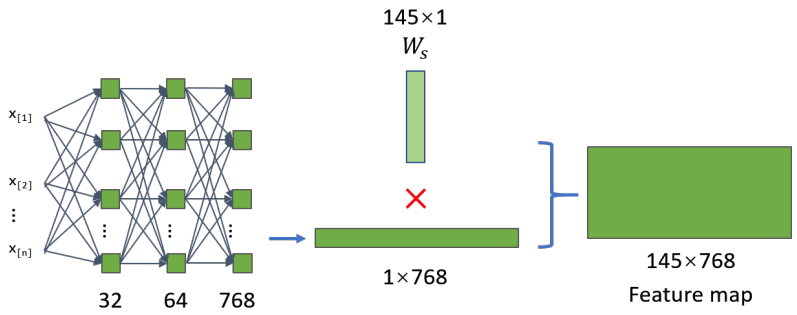


Fig. 2. Sensor input model structure. The MLP module is a multilayer perceptron. It has 19 neurons at the input layer and 768 neurons at the output layer. The two hidden layers contain 32 and 64 neurons, respectively. We found the performance using 2-layer MLP and 3-layer MLP doesn't have significant difference, thus we used the simpler 2-layer MLP.

3.2 Linear Projection Flattened Patches module

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.

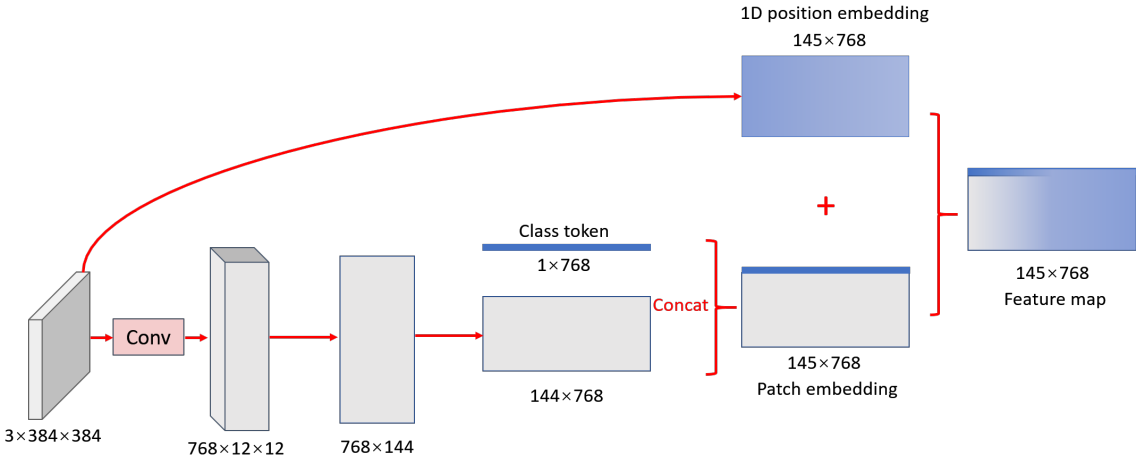


Fig. 3. Linear Projection Flattened Patches module. Linear Projection Flattened Patches divide the image into several equal small images and align the image feature map with the sensor features through matrix transformation.

The 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 blocks with the size of 12×12 by convolution operation. The convolution operation was performed using a kernel of size 32×32. The stride of the convolution operation is 32. 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 vector is transposed to the 144×768 vectors to be fused with the feature vector from the sensor. Then, the 144×768 feature vector is concatenated with the class token vector to generate patch embedding.

The Class Token is a specialized token that can represent the category information of an image. It is a vector that is typically added to the embedded representation of the image. The Class Token can provide additional information by encoding some global features that are shared by all images, which can help the model better understand the content of the images.

The position information of 1D Position embedding is added to patch embedding to retain positional information. Both class token and 1D Position embedding refer to ViT [55]. The position information of each pixel in an image can be transformed into a high-dimensional vector, where each dimension of the vector represents the position of the pixel in a specific dimension of the image. This high-dimensional vector can be concatenated with the feature vector of each pixel (e.g., color information, texture information, etc.), resulting in a higher-dimensional feature vector. ViT has proved that the performance of 2D position embedding is not better than that of 1D position embedding, so 1D position embedding is chosen.

Finally, the Patch embedding is added to the 1D position embedding to obtain the feature map of the image.

The features of the image and the sensor are represented using vectors with the same dimensions. The vectors are ready to be input to the next transformer encoder for complete feature fusion.

3.3 Transformer Encoder for fusion

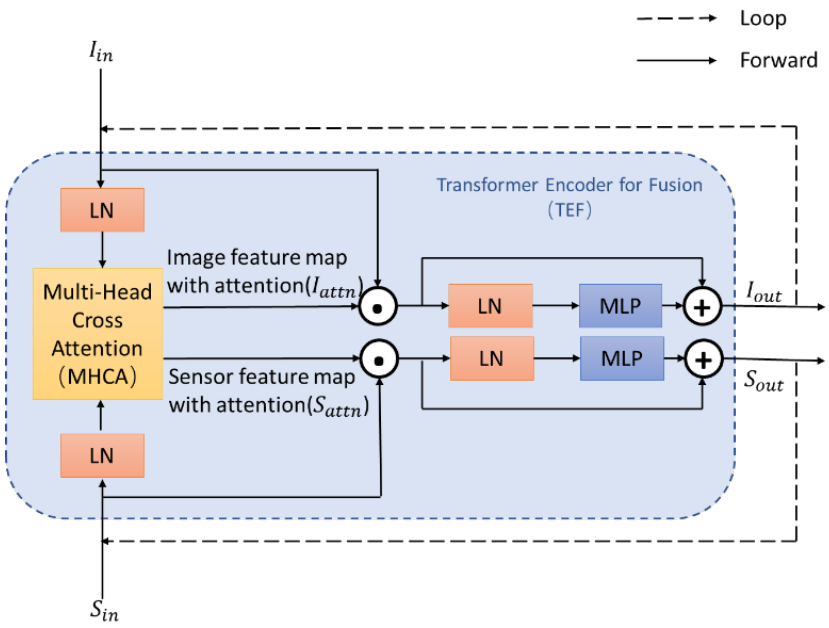


Fig. 4. The structure of the Transformer Encoder for Fusion (TEF) model consists of alternating layers of Multi-Head Cross Attention (MHCA), multiple layer perceptron (MLP) and Layer Normalization (LN) modules. MHCA is the central module responsible for fusing the image and sensor features. The MLP is adopted from ViT, and the LN module is responsible for performing Layer Normalization operations.

The Transformer Encoder for Fusion (TEF) model serves as the central model for multi-modal data fusion. The calculation process for TEF in Figure 4 is given as follows.

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

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

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

Here I_{attn} and S_{attn} are defined as the output of MHCA module, respectively. I_{attn},S_{attn} are Image and Sensor Feature Map with attention, respectively. I_{in},S_{in} are defined as the inputs of TEF, respectively. I_{out} and S_{out} are outputs of TEF, respectively. I_{in} and I_{out} are Image Feature Maps. S_{in} and S_{out} are Sensor Feature Maps. There are 12 loop computations for TEF. The output of TEF I_{out} and S_{out} are the input for TEF at next loop. After 12 loops is finished, I_{out} and S_{out} are given to next module.

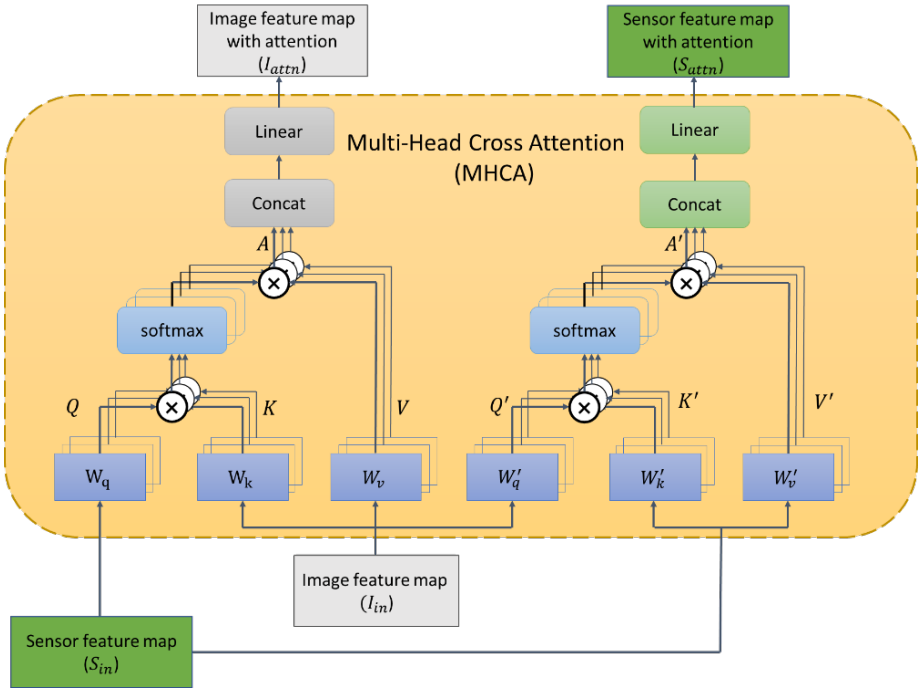


Fig. 5. The Multi-Head Cross Attention Module (MHCA) structure serves as a central component of the Transformer Encoder for Fusion. It operates by taking in both the image feature map (I_{in}) and the sensor feature map (S_{in}) as inputs. The module employs cross-attention mechanisms to investigate the interrelationships and associations among each feature map and between the two feature maps. Through this approach, the MHCA module achieves an improved ability to identify meaningful features from both modalities, generating an image feature map with attention (I_{attn}) and a sensor feature map with attention (S_{attn}) as outputs.

The Cross Attention module is proposed to model intra-modality relationships for image and sensor features. The structure of MHCA is shown in Figure 5. There are two branches to determine the parameters Query(Q, Q'), Key(K, K'), and Value (V, V'). One branch calculates Q, K, V for Image feature map with attention I_{attn}. The other calculates Q',K',V' for Sensor feature map with attention S_{attn}. The calculation process is given as followings:

$$Q = S_{in}W_q, K = I_{in}W_k, V = I_{in}W_v \tag{5}$$

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

$$Q' = I_{in}W'_q, K' = S_{in}W'_k, V' = S_{in}W'_v \tag{7}$$

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

where $W_q, W_k, W_v, W'_q, W'_k, W'_k \in \mathbb{R}^{C\times(C/h)}$ are linear transformation matrix with learnable parameters. These Multiple matrixes are computed by multi-head. C and h are the embedding dimension and number of heads, respectively. A and A' are results of equation 6 and equation 8, respectively. Multiple A and A' are concatenated, respectively. The sensor feature map S_{in} for one branch is used for query Q to compute K' and then update V'. The Image Feature Map I_{in} for another branch is used for query Q' to compute key K and value V. Therefore, these two branches exchanges information between sensor and image features for sematic information sharing. The computation complexity of generating the attention map in cross-attention are linear rather than quadratic as in all-attention. The entire process for MHCA module is more efficient. And it can reduce the risk of overfitting.

3.4 Loss function

The leaf area index (LAI) data was manually collected from the farm and used as labels for the ViST model. Following data normalization, LAI values were scaled to a range of [0,1]. The loss function quantifies the discrepancy between the model's predicted value y' and the actual value y. This non-negative, real-valued function is typically expressed as $L(y, y')$ and serves as the foundation for the empirical risk function and the structural risk function.

In this paper, we approach crop growth prediction as a regression problem. A neural network with a single output node is typically used to solve regression problems, with the output value representing the predicted value. To evaluate the performance of our model, we use the mean square error loss function as the primary metric.

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

where n is the number of samples, y_i is the true value of the leaf area index for the i-th sample, and y'_i is the predicted value of the leaf area index for the i-th sample. The MSE serves as a measure of the accuracy of our model's output and is used to update the model parameters in subsequent iterations.

4 EXPERIMENTS AND RESULTS

In this section, we will present the methods used for data collection, describe the equipment and data format employed. Following data collection, both image and sensor data undergo preprocessing. The evaluation criteria for model performance and experimental details will also be outlined.

4.1 Data collection

The location for data acquisition is shown in Figure 6. The rice, maize, and soybean data were collected on farm. There are three sample points for each crop. There are totally 9 sample points on farm. The data acquisition device is shown

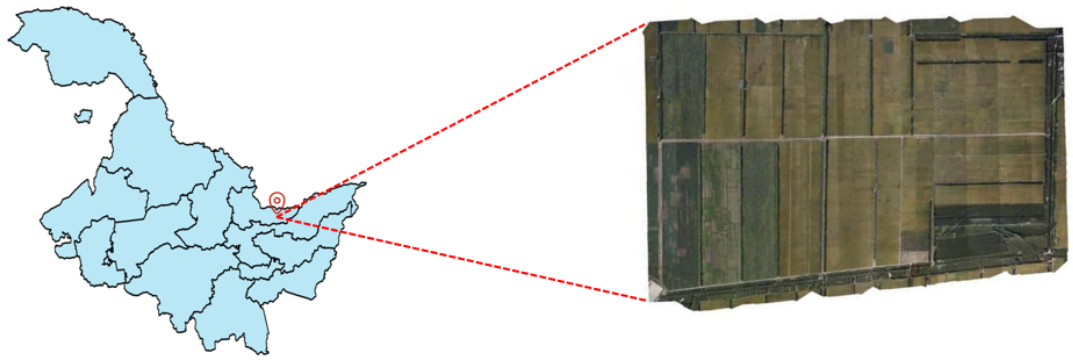


Fig. 6. The data were collected at Farm 290, located in Suibin County, Hegang City, Heilongjiang Province, China.

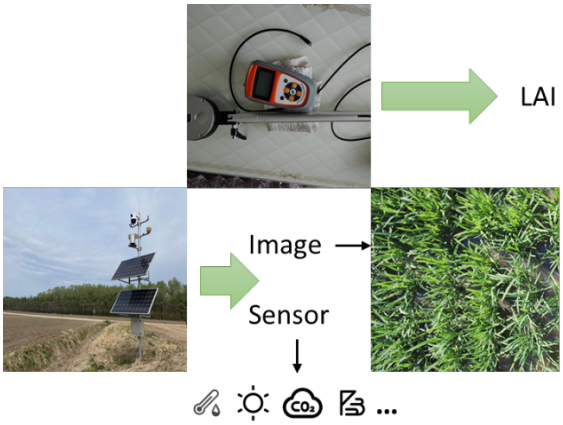


Fig. 7. Data acquisition equipment. The image and sensor data collection device is shown on the left, the leaf area index data collection device is shown in the middle, and image data is shown on the right.

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

Fig. 8. Crop image data.

Table 1. Sensor data. The table lists the data items captured by the sensors and their corresponding units.

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

in Figure 7. One device placed at each sample point on farm includes one camera and 11 number of sensors. LAI data were collected periodically using a hand-held device at each sample point for data alignment.

The crops are photographed from a top view. The images of rice, soybeans, and maize captured by the camera on the farm are shown in Figure 8(a)(b)(c), respectively. The height is set at 3 meters. The image format is RGB with a resolution of 3840×2160. The time interval between each image was two hours.

The collected items of sensors are shown in Table 1, including eleven sensor data types. The soil sensors are deployed in the ground at 10cm, 20cm, 30cm, 40cm, and 50cm depths, respectively. Air, light, and Wind sensors are deployed on the top of the IoT Device. The time interval of data collection for each sensor is half an hour. The image and sensor data were collected and uploaded to the cloud server periodically for storage and data analysis.

4.2 Data preprocessing

Preprocessing of image data: To ensure better convergence in backpropagation, the image data was normalized to a specific range using the Z-score method on each image's data.

$$z = \frac{x - \mu}{\sigma} \tag{10}$$

In this study, where μ denotes the mean value, σ represents the standard deviation, x denotes the input data, and z denotes the output data. The mean and standard deviation of the three channels of the image were recorded, respectively. Specifically, the mean value of the normalized image was set to 0 on each channel, and the variance was set to 1. However, it should be noted that this normalization method may not be applicable to cases with small sample sizes, and is generally recommended to be used only when the sample size exceeds 30. Table 2 presents the normalized results of image calculation for rice, soybean, and maize.

Preprocessing of sensor data: To eliminate the influence of dimensions, we normalized each size using the following equation, as the numerical differences between dimensions in the sensor data were relatively large.

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{11}$$

where x_{min} is the minimum value, and x_{max} is the maximum value. $x_{normalized}$ is the normalized value.

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 3.

Table 2. Image standardized data. The table presents the mean and standard deviation values for the RGB channels of the images of three crops, namely rice, soybean, and corn. These values were obtained through the normalization process described in the previous section.

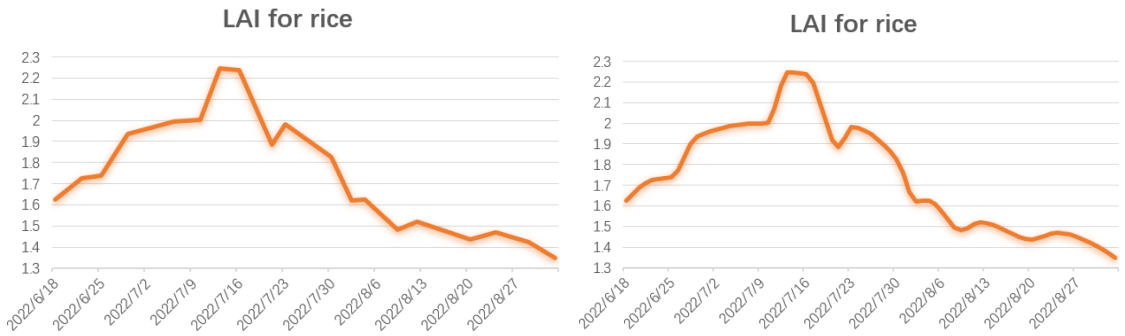
crop	channel	mean	standard deviation
rice	Channel 1	0.4452	0.1973
	Channel 2	0.5014	0.2035
	Channel 3	0.4292	0.183
soybean	Channel 1	0.5	0.1517
	Channel 2	0.5355	0.1641
	Channel 3	0.487	0.1497
corn	Channel 1	0.4452	0.1973
	Channel 2	0.5014	0.2035
	Channel 3	0.4292	0.183

Table 3. Sensor input metrics preprocessing. The table includes the indicators for sensor data processing, as well as their maximum and minimum values.

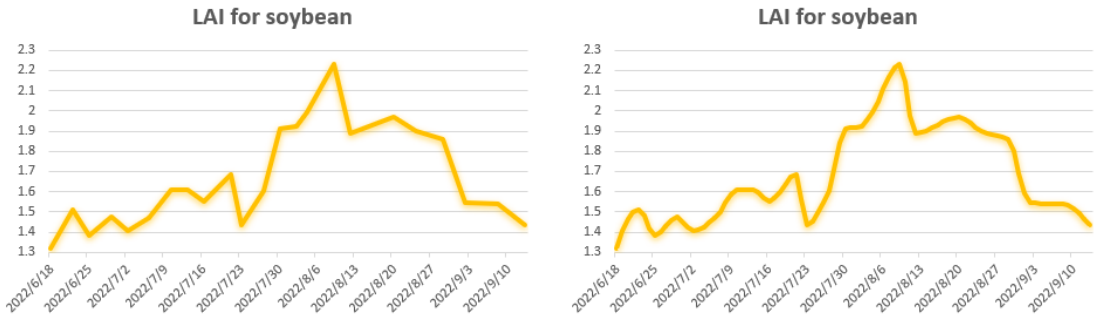
Indicators	Minimum	Maximum
Carbon dioxide	364	636
Soil temperature for 10 centimeters depth	18.1	25.1
Soil temperature for 20 centimeters depth	18.3	23
Soil temperature for 30 centimeters depth	18.3	22.1
Soil temperature for 40 centimeters depth	-30	-30
Soil temperature for 50 centimeters depth	17.1	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

Processing of LAI data: In order to ensure better convergence in backpropagation, the image data were normalized to a specific range using the Z-score method. For each crop, three sets of image and sensor devices were placed at the same position as the camera when measuring the LAI. During the collection of LAI data, hand-held devices capture a part of areas covered by the camera device. LAI data were collected every 5 days. In order to obtain more data, the piecewise cubic Hermite interpolation polynomial (PCHIP) method was used, as described by Fritsch and Carlson [56]. LAI data were available for each day and were interpolated for three sample points of rice, soybean, and maize as

shown in Figure 9(a),(b),(c), respectively. The left side of the figure shows the original data and the right side shows the interpolated data. The interpolated data appear smoother, as can be seen from the curves.



(a) Before and after rice LAI interpolation



(b) Before and after soybean LAI interpolation

(c) Before and after maize LAI interpolation

Fig. 9. LAI data preprocessing. The figure shows the leaf area index data before and after interpolation. The horizontal axis represents time, and the vertical axis represents the leaf area index.

Data alignment: One sample includes image data, sensor data, and a label – LAI data. The image data, sensor data, and LAI data were aligned based on their respective sampling location and time. Each image was captured at a two-hour interval, while the sensors' sampling frequency was every half hour. The time alignment between the data sets was

established on a half-hour basis. Due to the smaller quantity of image data relative to the sensor data, four samples collected over a two-hour period corresponded to a single image. Notably, there is only one LAI data point per day; this means that forty-eight samples share the same LAI label.

Training and test sets: For each crop, three sample points were selected, yielding a total of nine observation points on the farm. Two of the sample points were utilized for model training, while the remaining sample point was used for model testing. Cross-validation was carried out across different sample points for each crop to prevent the problem of using the same sample data for both training and testing, which could result in a smaller difference in the data gradient.

The amount of data utilized in the experiment is illustrated in Fig. 4, with "days" denoting the duration of data collection. "Train sample" and "Test sample" refer to the number of data instances employed for training and testing, respectively.

Table 4. The quantity of experimental data.

Crop	Days	Train sample	Test sample
Rice	83	7968	3984
Soybean	87	8352	4176
Maize	83	7968	3984

4.3 Evaluation Metrics

Mean square error (MSE), mean Absolute error (MAE), Mean Absolute Percentage Error(MAPE) and Symmetric Mean Absolute Percentage Error(SMAPE) are used to predict the performance of various models. MSE is the same as the loss function. MAE is the sum of the absolute differences between the target and prediction values.

$$MAE = \frac{\sum_{i}^{n} |y_i - y_i'|}{n} \tag{12}$$

where y_i' is the evaluated value of the model, y_i is the real value, and n is the total number of samples.

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

$$MSE = \frac{\sum_{i=1}^{n} (y_i' - y_i)^2}{n} \tag{13}$$

where y_i and y_i',i are the true value and the corresponding evaluated value for the first sample, and n is the number of samples.

MAPE is a commonly used performance metric in estimation. It measures the accuracy of a estimation by calculating the absolute percentage difference between the estimation value y_i' the actual value y_i, and then take the average of all the absolute percentage differences. The formula for calculating MAPE is as follows:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i' - y_i}{y_i} \right| \tag{14}$$

where n is the number of samples, y_i is the actual value, and y_i' is the estimate value.

SMAPE is a widely used metric for evaluating the accuracy of forecasting models. It is a symmetric error that calculates the average percentage difference between the actual and evaluated values . Unlike other error measures,

SMAPE takes into account the magnitude of the actual values, making it a more reliable metric for comparing the accuracy of different forecasting models. The formula for calculating SMAPE is as follows:

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{\left|y_i' - y_i\right|}{\left(\left|y_i'\right| + \left|y_i\right|\right)/2} \tag{15}$$

where n, y_i and y_i' are same as the symbols in MAPE formula. SMAPE ranges from 0% to 200%, with lower values indicating better accuracy.

4.4 Experiments and results

In this section, we present the extensive experiments conducted to demonstrate the effectiveness of our proposed ViST over existing methods. Our model was compared with the state-of-the-art models, such as Rice-Fusion [57], Rice-Transformer [58], CNN-Transformer [59], DNNF1, and DNNF2 [49]. While both DNNF1 and DNNF2 are multimodal models developed for soybean yield prediction, Rice-Fusion and Rice-Transformer employ multimodal approaches to diagnose rice diseases. The CNN-Transformer model was proposed to classify crops based on multitemporal data. In terms of comparison, the input and output to these models were kept the same as for our ViST model. Finally, the output of all models was evaluated based on the LAI value representing the crop growth status.

ViST was subjected to ablation studies with regards to its input modes, including image-only data, sensor-only data, and multimodality data. The performance of each model was evaluated using a range of metrics, including MAE, MSE, MAPE, and SMAPE which are presented in Equations 12,13,14, and 15, respectively.

The preprocessed data were utilized to train each model using the primary hyperparameters of ViST, which are outlined in Table 5. An Adam optimizer was employed, with a weight decay of 0.0001, and all models were trained on a GeForce RTX 3090 GPU. To dynamically adjust the learning rate, the cosine annealing strategy was implemented. The maximum number of iterations was determined based on the sample size and epoch, resulting in a monotonically decreasing learning rate with increasing epochs during the training process. The ViST model's primary hyperparameters can also be found in Table 5. Random normal distribution was used to initialize the parameters of each model. For the ablation experiment, the MM, SMI, and SMS were used as inputs for the model.

Table 5. 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

4.4.1 Experiments and results for a single crop. In this section, we evaluated the performance of our model on three crops - rice, soybean, and maize. Firstly, we compared the model's performance using single modality data and multimodality

data, as discussed in Section A. Secondly, we compared the performance of our ViST model with that of other models, as outlined in Section B.

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

The ViST model was trained for each crop with three different data input modes: image data, sensor data, both of image and sensor data. Thus, three trained models were obtained for each crop.

Table 6. ViST model test results for different input modes.

Crop	Input mode	MSE	MAE	MAPE	SMAPE
rice	Sensor	**0.00227**	0.03469	0.14928	13.1395
	Image	0.00548	0.0527	0.27484	20.19456
	Multimodality	0.00244	**0.01344**	**0.05614**	**5.26049**
soybean	Sensor	0.03201	0.12261	0.67907	34.33472
	Image	0.00536	0.04973	0.31889	15.41658
	Multimodality	**0.00007**	**0.00351**	**0.02644**	**2.06496**
maize	Sensor	0.00183	0.03185	0.14455	12.50781
	Image	0.00409	0.04577	0.19804	13.61606
	Multimodality	**0.00003**	**0.00304**	**0.01768**	**1.53183**

After the trained models were obtained, we used the data from the test dataset to test the models. The performance of the ViST model for test dataset is shown in table 6.

It is shown that the multimodality mode has achieved the lowest value of the prediction indictors among three modes for three crops. As four indictors have shown the same trend on the experiment results, MAE is used to analysis the performance of the ViST model. In the rice experiment, the value of MAE decreases by, 61.25%, 74.49% compared to that of the sensor and image mode, respectively. In the soybean experiment, the value of MAE decreases by 97.13%, 92.94%, compared to that of the sensor and image mode, respectively. In the maize experiment, the value of MAE decreases by 90.45%, 93.35%, compared to that of the sensor and image mode, respectively. Therefore, it can be concluded that combining image and sensor data significantly improves the performance for ViST model. And both sensor and image data have almost equal contribution for model training.

For rice, sensors are more effective than images because there is little difference in rice images during growth and leaf features are not prominent. For soybeans, images are more effective than sensors because the image features change significantly during growth. There is little difference between image and sensor data for corn, and both are effective in reflecting crop growth because error rates for various evaluation indices are relatively small.

Using image analysis, crop surface texture features, such as color, texture, and shape, can be analyzed to infer the healthy status of the crop, including whether it is growing well and whether it is affected by pests and diseases. Sensor data can capture changes in environmental factors, such as temperature, humidity, light, and soil, to assess their impact on crop growth. For example, excessive drought or excessive moisture can have a negative effect on crop growth. By comprehensively analyzing these two aspects of information, crop growth can be more accurately evaluated, and appropriate agricultural management measures can be adjusted accordingly.

In the experiments on rice, soybeans, and corn, the multi-modality results were better than the single-modality results. Multi-modal data can supplement the information that is lacking in single-modality data, completing the information gap. For example, images only reflect surface information of the crop, while sensors only reflect environmental information. If

only single modality data is used, many important pieces of information will be ignored. By integrating multiple types of information, different types of information can complement each other, improving the accuracy and comprehensiveness of data analysis.

The fusion of image and sensor multi-modal data can complete noise suppression. Noise on the data is common, especially in sensor data processing. However, multi-modal data can reduce noise on the data by fusing the data, which is more robust than single modality data analysis.

Single-modality data is more susceptible to interference than multi-modality data, which is more robust. Sensor data is often affected by various external environmental factors (such as rain, dust, and occlusion), which can affect data accuracy. Conversely, image data is less affected by external factors. Therefore, in multi-modality data analysis, the combination of image data significantly improves data accuracy and robustness.

Thus, the complementary nature, noise suppression, and robustness of multi-modal data analysis can more comprehensively understand things' essence and more accurately describe the actual situation.

B.Comparison with other models

In this Section, we compared the proposed ViST model with other models: DNNF1, DNNF2, CNN-Transformer, Rice-Fusion, Rice-Transformer . All three models were trained on rice, soybean, and maize data, respectively.

Table 7. The comparative experimental test results of rice.The "Mean" and "Std" represent the mean value and standard deviation, respectively, of the results obtained from five experimental groups.

Model	MSE		MAE		MAPE		SMAPE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CNN-Transformer	0.00024	2.71×10^{-5}	0.01368	2.69×10^{-5}	0.04287	2.52×10^{-5}	4.3371	2.75×10^{-4}
DNNF1	0.00177	2.47×10^{-5}	0.01775	2.14×10^{-5}	0.08548	3.47×10^{-5}	7.21533	2.81×10^{-5}
DNNF2	0.00333	2.18×10^{-5}	0.02816	2.6×10^{-5}	0.11206	2.05×10^{-5}	10.303	1.63×10^{-3}
Rice-Fusion	0.00024	$\mathbf{1.92 \times 10^{-5}}$	0.01065	1.93×10^{-5}	0.04215	1.84×10^{-5}	3.95636	2.55×10^{-5}
Rice-Transformer	0.00244	2.99×10^{-5}	0.01344	$\mathbf{1.88 \times 10^{-5}}$	0.05614	$\mathbf{1.5 \times 10^{-5}}$	5.26049	$\mathbf{1.33 \times 10^{-5}}$
ViST	**0.00023**	2.78×10^{-5}	**0.00923**	2.13×10^{-5}	**0.03263**	2.17×10^{-5}	**3.15724**	3.02×10^{-5}

Table 7 shows the experimental results of various models when applied to the rice dataset. The ViST model showed the best performance on all prediction metrics, implying its excellent performance in estimating rice growth. In the average, ViST lowered MSE by 55.80%, MAE by 38.48%, MAPE by 44.21%, and SMAPE by 42.59%. Compared to the other models, the ViST model shows better performance in accuracy and stability, especially when compared to DNNF1 and DNNF2. For instance, compared to DNNF1, ViST lowered MSE by 87.01%, MAE by 48.00%, MAPE by 61.83%, and SMAPE by 56.24%. It is also shown that ViST has almost same value on MSE with CNN-Transformer and Rice-Fusion model.

Similarly to Table 7, Table 8 presents the experimental results of ViST and other comparative models. The experiments in Table 8 use soybean data as the dataset. By considering the improvement ratios relative to the other models, it is evident that ViST achieves significant enhancements across all four metrics, with the most notable improvements visible on the MAE and MAPE metrics. Regarding the MSE metric, ViST achieves reductions of 56.25%, 92.31%, 61.11%, and 50.00% compared to DNNF1, DNNF2, Rice-Fusion, and Rice-Transformer, respectively. Furthermore, ViST decreases the error by 46.15% relative to CNN-Transformer.

Table 9 shows the experimental results of various models when applied to the maize dataset. The Rice-Fusion model exhibits good performance across four performance metrics, though slightly worse than ViST in terms of MAPE

Table 8. The comparative experimental test results of soybean.The "Mean" and "Std" represent the mean value and standard deviation, respectively, of the results obtained from five experimental groups.

Model	MSE		MAE		MAPE		SMAPE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CNN-Transformer	0.00013	1.75×10^{-5}	0.00749	2.69×10^{-5}	0.0705	1.26×10^{-5}	3.43951	2.47×10^{-5}
DNNF1	0.00016	$\mathbf{1.32 \times 10^{-5}}$	0.00958	2.56×10^{-5}	0.05863	$\mathbf{1.31 \times 10^{-5}}$	4.10262	2.24×10^{-5}
DNNF2	0.00091	1.81×10^{-5}	0.01597	2.38×10^{-5}	0.08465	2.71×10^{-5}	6.45109	2.74×10^{-5}
Rice-Fusion	0.00018	2.63×10^{-5}	0.00656	2.76×10^{-5}	0.0969	1.41×10^{-5}	3.74191	$\mathbf{1.06 \times 10^{-5}}$
Rice-Transformer	0.00014	2.35×10^{-5}	0.00478	1.47×10^{-5}	0.08859	1.32×10^{-5}	3.17471	2.48×10^{-5}
ViST	**0.00007**	2.95×10^{-5}	**0.00351**	$\mathbf{1.03 \times 10^{-5}}$	**0.02644**	2.73×10^{-5}	**2.06496**	1.64×10^{-5}

Table 9. The comparative experimental test results of maize.The "Mean" and "Std" represent the mean value and standard deviation, respectively, of the results obtained from five experimental groups.

Model	MSE		MAE		MAPE		SMAPE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CNN-Transformer	0.00013	$\mathbf{1.05 \times 10^{-5}}$	0.00805	1.82×10^{-5}	0.03245	2.29×10^{-5}	4.27382	1.93×10^{-5}
DNNF1	0.00026	2.03×10^{-5}	0.00926	2.71×10^{-5}	0.05208	1.89×10^{-5}	4.83925	3.28×10^{-5}
DNNF2	0.00009	5.36×10^{-5}	0.00775	2.53×10^{-5}	0.02958	3.39×10^{-5}	3.10525	2.7×10^{-5}
Rice-Fusion	0.00007	1.91×10^{-4}	0.00609	2.74×10^{-5}	0.01712	1.65×10^{-5}	1.71803	2.69×10^{-5}
Rice-Transformer	0.00012	2.31×10^{-5}	0.00672	$\mathbf{1.29 \times 10^{-5}}$	0.02436	3.02×10^{-5}	2.39245	1.77×10^{-5}
ViST	**0.00003**	1.63×10^{-5}	**0.00304**	1.37×10^{-5}	**0.01768**	$\mathbf{1.51 \times 10^{-5}}$	**1.53183**	$\mathbf{1.41 \times 10^{-5}}$

and SMAPE. Despite this, it still outperforms other models. The CNN-Transformer model demonstrates intermediate performance across the four metrics, slightly outperforming the DNNF1 and DNNF2 models with only small variations. It is apparent that the ViST model exhibits a reduction in error percentages greater than 50% for both MAE and MSE metrics. This suggests that the ViST model is able to more accurately predict experimental data and is more stable and reliable in comparison to other models.

The result for three crops compared with other models indicates that the visual and spatiotemporal information captured by ViST is critical in estimating rice growth accurately. Cross attention mechanics can effectively extract texture from leaf image and fusion sensor data with image feature. This suggests that the ViST model is able to more accurately evaluate experimental data and is more stable and reliable in comparison to other models.

4.4.2 Experiments and results for combined data training. The data from three crops were combined and used to train the model to test the model's generalization ability for the growth prediction of multiple crops. If only a single crop dataset is used to train the model, the model will focus too much on the characteristics of that crop, increasing the risk of overfitting. However, training the model with multiple crop datasets can reduce the risk of overfitting and improve the model's stability. With the application of many artificial intelligence technologies, data in the field of agriculture has become very rich. However, the amount of data for each crop itself is not large. Therefore, training the model with multiple agricultural datasets can better utilize limited data resources and improve the model's applicability.Table 10 shows the results of ViST, DNNF1, DNNF2, CNN-Transformer, Rice-Fusion, Rice-Transformer models with the hybrid data set. The ViST model outperforms other models, particularly in terms of MAPE and SMAPE with higher

Table 10. The comparative experimental test results for hybrid.The "Mean" and "Std" represent the mean value and standard deviation, respectively, of the results obtained from five experimental groups.

Model	MSE		MAE		MAPE		SMAPE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CNN-Transformer	0.00019	3.09×10^{-5}	0.00822	$\mathbf{1.98 \times 10^{-5}}$	0.03928	1.66×10^{-5}	2.994976	2.27×10^{-5}
DNNF1	0.00141	2.63×10^{-5}	0.01928	2.07×10^{-5}	0.09484	1.59×10^{-5}	7.311991	1.51×10^{-5}
DNNF2	0.00083	3.07×10^{-5}	0.01388	1.99×10^{-5}	0.04452	$\mathbf{1.47 \times 10^{-5}}$	4.104376	2.31×10^{-5}
Rice-Fusion	0.00024	3.11×10^{-5}	0.00857	3.09×10^{-5}	0.04541	2.14×10^{-5}	3.579631	1.62×10^{-5}
Rice-Transformer	0.00038	2.01×10^{-5}	0.00614	2.08×10^{-5}	0.03564	2.54×10^{-5}	2.467895	$\mathbf{1.43 \times 10^{-5}}$
ViST	**0.00014**	$\mathbf{1.91 \times 10^{-5}}$	**0.00591**	2.85×10^{-5}	**0.02775**	1.65×10^{-5}	**2.408083**	1.51×10^{-5}

accuracy. ViST's MAE performance is superior to other models, showing a 28.10%, 3.75%, and 31.04% reduction compared to the CNN-Transformer, Rice-Transformer, and Rice-Fusion models, respectively. Furthermore, ViST's outstanding performance in MSE indicates its ability to capture data features effectively, with reductions of 26.32%, 63.16%, and 41.47% compared to the CNN-Transformer, Rice-Transformer, and Rice-Fusion models, respectively.

The cross-attention mechanism adopted in ViST can fuse information from input data better, thus improving model performance and generalization capability. The performance of DNNF1 and DNNF2 based on multi-layer perceptron is worst due to their lack of complexity and flexibility in model structure, resulting in ineffective handling of various crop growth estimation tasks compared to ViST.

Table 11. Comparison Results of ViST model with different crops as input.

Train data	Test data	MSE	MAE	MAPE	SMAPE
Rice	Rice	0.00244	0.01344	0.05614	5.26049
Soybean	Soybean	0.00007	0.00351	0.02644	2.06496
Maize	Maize	**0.00003**	**0.00304**	**0.01768**	**1.53183**
Three crops	Three crops	0.00014	0.00591	0.02775	2.40808

The table presents the results of ViST for four sets of experiments involving rice, soybean, corn, and a mixture of the three crops data. The results on the ubiquitous ability of ViST are shown in Table 11. The model for Maize data exhibits relatively lower prediction errors compared to other crop data. For example, in terms of the SMAPE metric, the error value of maize is 1.53%, while the error values of rice, soybean, and a mixture of three crops are 5.26%, 2.06%, and 2.41%, respectively. It can be seen that maize shows better performance. The reason is that Maize has the most feature.

In the mixed experiment of the three crops, the error values of all prediction indicators are slightly increased relative to the single crop experiment, but the magnitude of the increase is small. This indicates that ViST exhibits a high level of prediction accuracy and robustness in handling multi-crop information, and therefore has high practical value for agricultural production applications. As training data is derived from various crops, inter-crop differences may exist and may affect the accuracy and stability of the model. The efficacy of combined training, with the exception of training with rice data, is comparatively suboptimal in contrast to that of other training with soybean and maize data.

4.4.3 Experiments and results for speed test. From experimental results in Figure10, it is shown that the inference speed of ViST is much close to that of Rice-Fusion and Rice-Transformer. Compared to CNN-Transformer, DNNF1, and

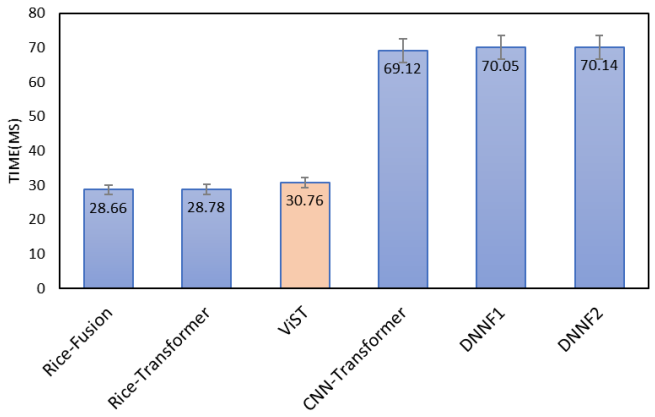


Fig. 10. Speed test results of the models.To compare the inference speed of models, each model performed 1000 inference runs, and the average result was taken as the final outcome. The horizontal axis represents the model name, and the vertical axis represents the average time for each inference run, in milliseconds. Each experimental result has an error rate of ±5%.

DNNF2, ViST's inference time is approximately 50% of theirs. The CNN model requires more convolution operations, resulting in a relatively greater computational load and slower inference speeds. The multilayer perceptron model entails a larger number of parameters and computational load, leading to slower inference speeds. The performance of DNNF1 and DNNF2 based on multi-layer perceptron is worst due to their lack of complexity and flexibility in model structure, resulting in ineffective handling of various crop growth estimation tasks compared to ViST. Although DNNF1 and DNNF2 are simple, the fusion of images and sensors results in unsatisfactory performance.

5 CONCLUSION AND FUTURE DIRECTIONS

This paper has proposed a Transformer-based ViST model for crop growth prediction by image and sensor data on a farm. The cross-attention mechanism in the model was used to improve the effect of model data fusion. The data of the three crops were trained together as input to the model. The model not only achieves high accuracy but also maintains a relatively fast speed. Experiment results show that the model with multimodality data can improve crop growth predication. Although the ViST model has achieved good results, there are still a number of limitations. As the model requires cross-modal feature fusion, this may increase the computational cost and training time of the model. And the ubiquitous of the model needs to be further validated in different crops to determine its applicability and generalizability. In the future, more crop growth data will be collected for model optimization. Furthermore, investigating the model's use in practical agricultural production can assist farmers in better decision-making and production management.

ACKNOWLEDGMENTS

This research is supported by New generation artificial intelligent program, No.21ZD0110900 in CHINA, Heilongjiang NSF funding, No.LH202F022 and Fundamental Research Funds for the Central Universities No.2023FRFK06013.

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] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *nature* 521, 7553 (2015), 436–444.

[5] Zilong Hu, Jinshan Tang, Ping Zhang, and Jingfeng Jiang. 2020. Deep learning for the identification of bruised apples by fusing 3D deep features for apple grading systems. *Mechanical Systems and Signal Processing* 145 (2020), 106922–.

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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

[18] 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.

[19] 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.

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

[21] 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.

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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).
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
[38] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2023. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. arXiv:2209.03430 [cs]
[39] 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.
[40] 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.
[41] 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.
[42] Zhou Lei and Yiyong Huang. 2021. Video Captioning Based on Channel Soft Attention and Semantic Reconstructor. *Future Internet* 13, 2 (2021), 55.
[43] 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.
[44] 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.
[45] 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.
[46] 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]
[47] 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.
[48] 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
[49] 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
[50] 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]
[51] 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.
[52] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. 2021. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. arXiv:2104.09224 [cs]
[53] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. [n. d.]. Attention Bottlenecks for Multimodal Fusion. ([n. d.]), 14.

[54] 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.

[55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and Dirk Weissenborn. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs]

[56] Frederick N Fritsch and Ralph E Carlson. 1980. Monotone Piecewise Cubic Interpolation. *SIAM J. Numer. Anal.* 17, 2 (1980), 238–246.

[57] Rutuja R. Patil and Sumit Kumar. 2022. Rice-Fusion: A Multimodality Data Fusion Framework for Rice Disease Diagnosis. *IEEE access : practical innovations, open solutions* 10 (2022), 5207–5222. https://doi.org/10.1109/ACCESS.2022.3140815

[58] Rutuja Rajendra Patil and Sumit Kumar. 2022. Rice Transformer: A Novel Integrated Management System for Controlling Rice Diseases. *IEEE access : practical innovations, open solutions* 10 (2022), 87698–87714. https://doi.org/10.1109/ACCESS.2022.3200688

[59] Zhengtao Li, Guokun Chen, and Tianxu Zhang. 2020. A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 847–858. https://doi.org/10.1109/JSTARS.2020.2971763

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