Common Feature Representation Learning for Multi-Crop Seed Classification via Hyperspectral Images

Zhiyi Zhang[a], Jiading Yuan[b], Pengfei Zhang[b,c], Zhuopin Xu[b] and Qi Wang[b,*]

[a]University of Science and Technology of China, Hefei, Anhui, China
[b]Hefei Institutes of Physical Science, Chinese Academy of Science, Hefei, Anhui, China
[c]Zhongke Technology Achievement Transfer and Transformation Center of Henan Province, Zhengzhou, Henan, China

ARTICLE INFO

Keywords:
Seed classification
Hyperspectral image processing
Deep learning
Common Feature Representation
Transfer Learning

ABSTRACT

Hyperspectral Imaging (HSI) has emerged as a powerful approach for rapid and non-destructive seed classification; however, current HSI models are often limited by data scarcity, alongside domain specificity from insufficient varieties and single-crop specialization, ultimately hindering model capacity and performance. To tackle this limitation, we exploit Common Feature Representations (CFRs) across multiple crops, forming a large-scale dataset with 296208 samples across 79 classes. This dataset promotes training deeper convolution-attention networks, that accurately process full HSI data end-to-end. Models are pre-trained on multi-crop data to learn generic CFRs, and then fine-tuned for target crops. We further optimize architectures and training pipelines for HSI, with hardware acceleration support. Experiments show that, compared to single-crop training baselines, CFRs enhance both accuracy and convergence. For maize, rice, sorghum, and wheat, the maximum accuracy is improved by 1.9%, 4.0%, 0.7%, and 0.9%, respectively, achieving final accuracy of 96.3%, 92.6%, 92.6%, and 98.6%——substantially outperforming small machine-learning models. CFRs also reduce performance variance and test-validation accuracy gap, confirming higher robustness. This work validates the scaling law for HSI Deep Learning: multi-crop data integration via CFRs expands knowledge domains, paving the way for larger and more powerful models.

1. Introduction

Seed variety is a crucial determinant of nutrition, agricultural yield, and market value, directly affecting crop performance and food security. Therefore, efficient seed classification is important for various applications, like food production, crop breeding, and animal husbandry. While accurate seed classification has been achieved through laboratory-based methods, such as protein electrophoresis, and DNA molecular marker; these methods are destructive, time-consuming, and labor-intensive, limiting their use for large-scale and online real-time analysis. To overcome these drawbacks, hyperspectral imaging (HSI) has become a promising tool for rapid and non-destructive seed classification.

Unlike normal RGB images and near-infrared spectroscopy, hyperspectral images integrate both spatial and spectral features across numerous wavelengths. This high-dimensional data can be modeled to infer biochemical and chemical features of seeds. To model HSI data, previous studies have explored machine learning (ML) methods, including K-Nearest Neighbors (KNN) [1–3], Random Forest (RF) [2, 4–6], and Support Vector Machines (SVM) [2, 7–11]. However, traditional ML methods depend on manually extracted features from selected spectral bands. This manual extraction process is labor-intensive, requires expert knowledge, and may lose fine-grained information. For more automated and accurate modeling, Deep Learning (DL) has been recently utilized, and demonstrated state-of-the-art performance on many important crops, such as rice [12, 35], wheat [13–16], maize [17–20], and sorghum [21]. In addition to Convolutional Neural Networks (CNNs), some studies [12, 17, 20] incorporate attention mechanisms to further improve performance.

However, there still remain challenges in HSI seed modeling:

(1) Data Scarcity. Unlike the millions of samples in large RGB datasets [22, 23], typical HSI datasets range from hundreds [2–5, 7–9, 11, 21, 24, 25] to thousands [10, 12–15, 17, 18, 35], with the largest open-source HSI seed dataset, RSHI60K [1], containing 60000 samples. This limited data size increases overfitting risks, especially for large models that demand extensive training data. Consequently, many studies have opted for smaller models, such as ML models or limited-depth DNNs [13–21], potentially sacrificing achievable performance. In contrast, many computer vision (CV) models [26–33] boast hundreds of layers and over 20 million parameters, allowing them to learn more intricate patterns and achieve state-of-the-art performance. Furthermore, as smaller datasets tend to have greater distributional bias and lower diversity, accuracy metrics may be inflated and imprecisely reflect real-world performance.

(2) Domain Specificity. Current models are frequently trained on single-crop data, with involved variety count typically less than 21 [12, 16–18], and even fewer than 11 [1–4, 7, 9, 10, 13–15, 21, 25, 34, 35]. This narrow focus restricts the knowledge domain and model generalization. In contrast, large RGB datasets [22, 23] contain over 1000 classes, encompassing a much broader range of visual patterns that enable robust learning. Within a narrow domain, HSI data can be both expensive to collect and limited in size, exacerbating the challenge of data scarcity. To leverage knowledge from related datasets. a few studies [34–36] have explored HSI transfer learning. However, these transfer approaches are constrained to the single-crop scope, and are insufficient to drive models comparable to those in CV.

To tackle these challenges, this work investigates Common Feature Representations (CFRs) to bridge multi-crop data, significantly expanding data volume and knowledge domain beyond single-crop constraints. By leveraging CFRs, related crops can share their data,

*corresponding author
gilgamesh@mail.ustc.edu.cn (Z. Zhang); yjdjdjd@stu.ahau.edu.cn (J. Yuan); pfzhang@aiofm.ac.cn (P. Zhang); xuzp@iim.ac.cn (Z. Xu); wangqi@ipp.ac.cn (Q. Wang)
ORCID(s): 0009-0006-7849-3222 (Z. Zhang); 0009-0001-0713-5536 (J. Yuan); 0000-0001-5415-9592 (P. Zhang); 0000-0002-1629-5988 (Z. Xu); 0000-0002-5810-9223 (Q. Wang)

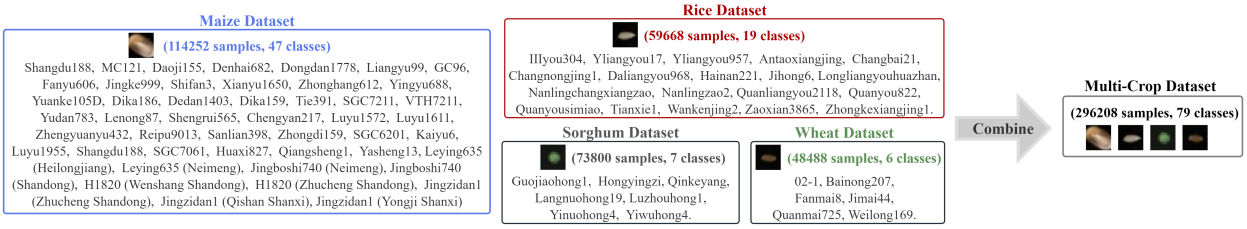

Figure 1: HSI seed datasets. The maize, rice, sorghum, and wheat datasets are combined to construct the multi-crop dataset.

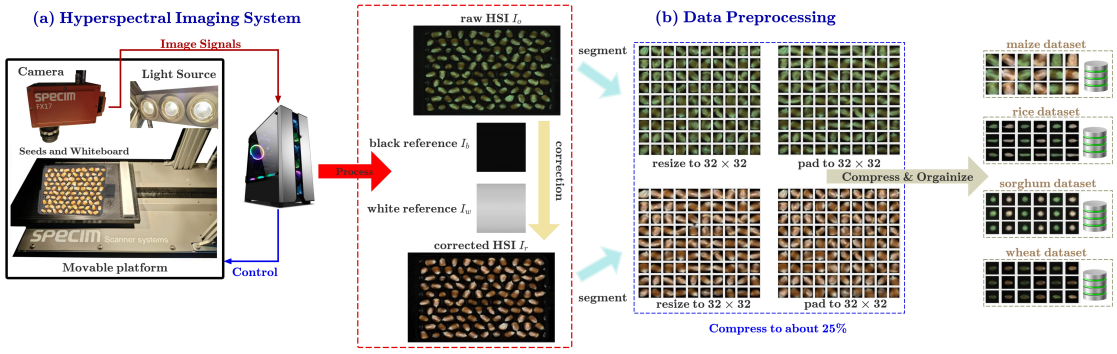

Figure 2: HSI dataset construction. **(a)** acquisition: A computer-controlled system captures raw HSI I_o, black reference I_b, and white reference I_w, which are then processed to generate corrected HSI I_c. **(b)** preprocessing: I_o and I_c are segmented and standardized to generate HSI samples, which are compressed and organized into datasets. The acquisition and preprocessing are similar for rice, wheat, and sorghum. Hyperspectral images are visualized by extracting the 13-th, 80-th, and 132-rd channels to generate RGB images.

and existing data can be used for modeling unseen crops, promoting efficient resource utilization. We propose a transfer-learning framework that exploits CFRs for HSI-based seed classification, comprising three key components:

- **Multi-Crop Dataset:** Four datasets, consisting of 47 maize, 19 rice, 7 sorghum, and 6 wheat classes, are combined into a multi-crop dataset containing 296208 samples. This results in 2.6× to 6.1× data-size increase over single-crop datasets, facilitating the training of high-capacity models.
- **Larger HSI Models:** Deep convolution-attention models, with over 30 million parameters, are designed to process full hyperspectral images end-to-end. Inspired by advanced CV models, the architectures are tailored for stale and effective HSI processing, including downsampling adjustment, and replacing BatchNorm with LayerNorm + LayerScale.
- **CFR-Driven Training:** Models are pre-trained on the multi-crop dataset to learn transferable CFRs, and then fine-tuned on target-crop data to specialize the learned knowledge. To further enhance performance, we leverage hardware acceleration, and modern training techniques, including sophisticated optimizers and HSI data augmentation.

Finally, this work analyzes how learning CFRs leads to higher accuracy and more robust convergence.

2. Dataset Construction

As detailed in Figure 1, this work develops four HSI seed datasets: a **maize** dataset contains 114252 samples across 47 classes, a **rice** dataset contains 59668 samples across 19 classes, a **sorghum** dataset contains 73800 samples across 7 classes, and a **wheat** dataset contains 48488 samples across 6 classes. These individual datasets are each larger than those used in many previous studies. Furthermore, to create a more comprehensive resource, all four datasets are combined into a **multi-crop** dataset, containing 296208 samples across 79 classes. Each unique seed variety is treated as an independent class, with the exception of several maize varieties (Leying635, Jingboshi740, H1820, and Jingzidan1) from different origins, which are considered as distinct classes.

The process of dataset construction, consisting of HSI acquisition and data preprocessing, is illustrated in Figure 2.

2.1. HSI Acquisition

As shown in Figure 2 (a), the hyperspectral images are acquired by a line-scan hyperspectral imaging system, consisting of a Specim FX17e camera, a 150W incandescent light source, a diffuse whiteboard, and a motorized linear translation platform. This system is connected to a computer, which controls HSI acquisition and processes signals to generate hyperspectral images.

The acquisition is carried out in a dark box, to avoid ambient light and ensure consistent illumination. Before acquisition, the light source and camera are activated for 30 minutes, to allow for temporal and thermal stabilization. Seeds and whiteboard are placed on the platform, at a 26 cm vertical distance to the camera lens. During scanning, the platform moves at a constant speed of 3 cm/s, while the camera acquires data with a frame rate of 80 and an exposure time of 4.9 ms.

The dark reference I_b is acquired by closing the camera shutter, and the white reference I_w is acquired by scanning the whiteboard.

The camera scans seeds to acquire the raw HSI I_o, which is subsequently normalized to calculate the corrected HSI $I_c = \frac{I_o - I_b}{I_w - I_b}$. Each raw or corrected HSI has a spatial resolution of 720×640 pixels, and contains 224 spectral bands ranging from 935.6 to 1720.2 nm wavelength.

2.2. Data Preprocessing

As shown in Figure 2 (b), each raw or corrected HSI is segmented based on spatial coordinates, with each segment containing only one seed. The segments are spatially resized or padded to a resolution of 32×32 pixels, where the pixel values are linearly normalized into the range $[0, 255]$ and converted into uint8 datatype. Each processed segment is designated as an HSI sample.

HSI samples are formatted in HWC, where H, W, and C denote the height, width, and channel dimensions, respectively. Each channel corresponds to a specific spectral band, with the pixels within each channel forming a 2D image $\in \mathbb{R}^{H \times W}$. HSI samples are compressed to approximately 25% of their original size using the deflate algorithm, significantly reducing memory usage and bandwidth demands.

The HSI samples for each crop are construct the corresponding single-crop dataset. Each dataset is split into training, validation, and test sets with an 8:1:1 ratio. To prevent data leakage, HSI samples originating from the same seed are arranged in the same set. The training set includes both raw and corrected HSI samples for data augmentation. To ensure unbiased evaluation, the validation and test sets retain only the corrected HSI samples. Finally, all individual training, validation, and test sets are combined to create the corresponding sets of the multi-crop dataset.

3. Architecture Design

Leveraging the substantial scale of our multi-crop dataset, we employ convolution-attention DNNs to establish end-to-end mappings from hyperspectral images to seed classes. Our models comprise 52 to 119 convolution and fully-connected layers, representing in a notable increase in depth and capacity than many previous models [12–15, 17, 17–20, 20, 21]. Inspired by advanced CV models [26, 27, 32, 38, 39], the architectures are further optimized for hyperspectral images. while classical spectral-preprocessing (e.g. SNV, MSC) and feature-extraction (e.g. UVE, SPA, PCA) techniques address the high correlation and redundancy in hyperspectral images, they inevitably lose fine-grained features that are crucial for accurate classification. In contrast, our high-capacity DNNs are able to process full spectral and spatial information, enabling a more comprehensively understanding [37].

To facilitate both CFR learning and crop-specific optimization, each model comprises three components: **stem**, **backbone**, and **classifier**, with the backbone accounting for over 95% of the total parameters. The stem extracts low-level features from hyperspectral images; then the backbone processes these low-level features to generate high-level representations; finally, the classifier encodes the high-level representations into class probabilities, and predicts the class with the highest probability. The stem and backbone can be pre-trained on the multi-crop dataset to learn generalizable CFRs, and all three components can be fine-tuned on single-crop datasets for better adaptation.

3.1. Stem and Classifier

All models share the same stem and classifier architectures, as illustrated in Figure 3.

The stem extracts low-level features, by downsampling 224 input channels to 128. This 128-channel configuration was experimentally determined, to achieve the best trade-off between accuracy and memory usage. Given the relatively low spatial resolution (32×32), the stem utilizes a 3×3 stride-1 convolution to better preserve fine-grained spatial details, rather than the 7×7 stride-2 convolution commonly used for ImageNet dataset.

The classifier first reduces feature spatial dimensions into 1×1, by adaptive average pooling. The resulting vector is then processed by two fully-connected layers. Finally, a Softmax layer encodes the features into class probabilities. To mitigate overfitting, two Dropout [40] layers with 0.4 dropout-rate are incorporated.

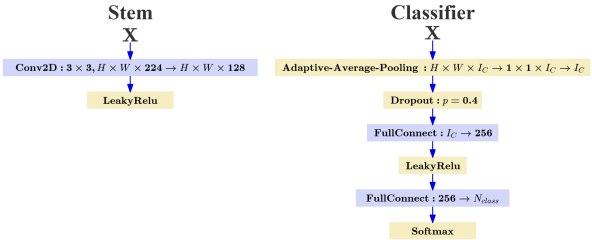

Figure 3: Stem and Classifier Architectures. H, W, and I_C denote the input height, width, and channel size, respectively. N_{class} is the number of classes.

3.2. Normalization Layer

Normalization layers are crucial for DNNs, providing faster convergence, greater stability, and enhanced generalization.

Batch Normalization (BatchNorm) [41] has proven effective in computer version, particularly on RGB datasets. However, when applied to HSI datasets, it can lead to the inconsistency between training loss and inference accuracy, severely degrading model performance on unseen data. As exemplified in Figure 4 (a), when training ResNet48 (with BatchNorm) on our maize dataset using 512 batch size, 5e-4 learning rate, and Adam optimizer [43], the inference accuracy violently fluctuates, inconsistent with the near-zero training loss. This suggests that the model encounters convergence problems, rendering the training loss an unreliable indicator of progress. Similar problems [42] have been observed, when training Transformers with BatchNorm for Natural Language Processing (NLP) tasks.

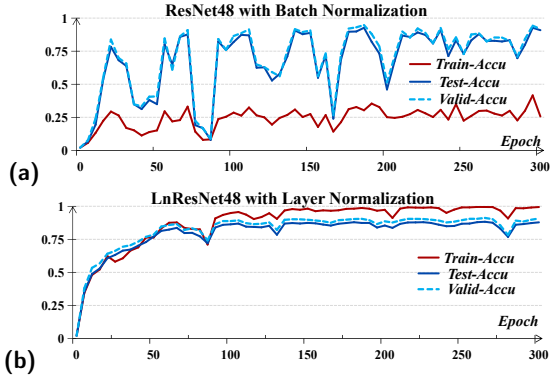

Figure 4: BatchNorm and LayerNorm training on HSI datasets. 'Train-Accu', 'Test-Accu', and 'Valid-Accu' represent the inference accuracy on train, test, and validation sets, respectively.

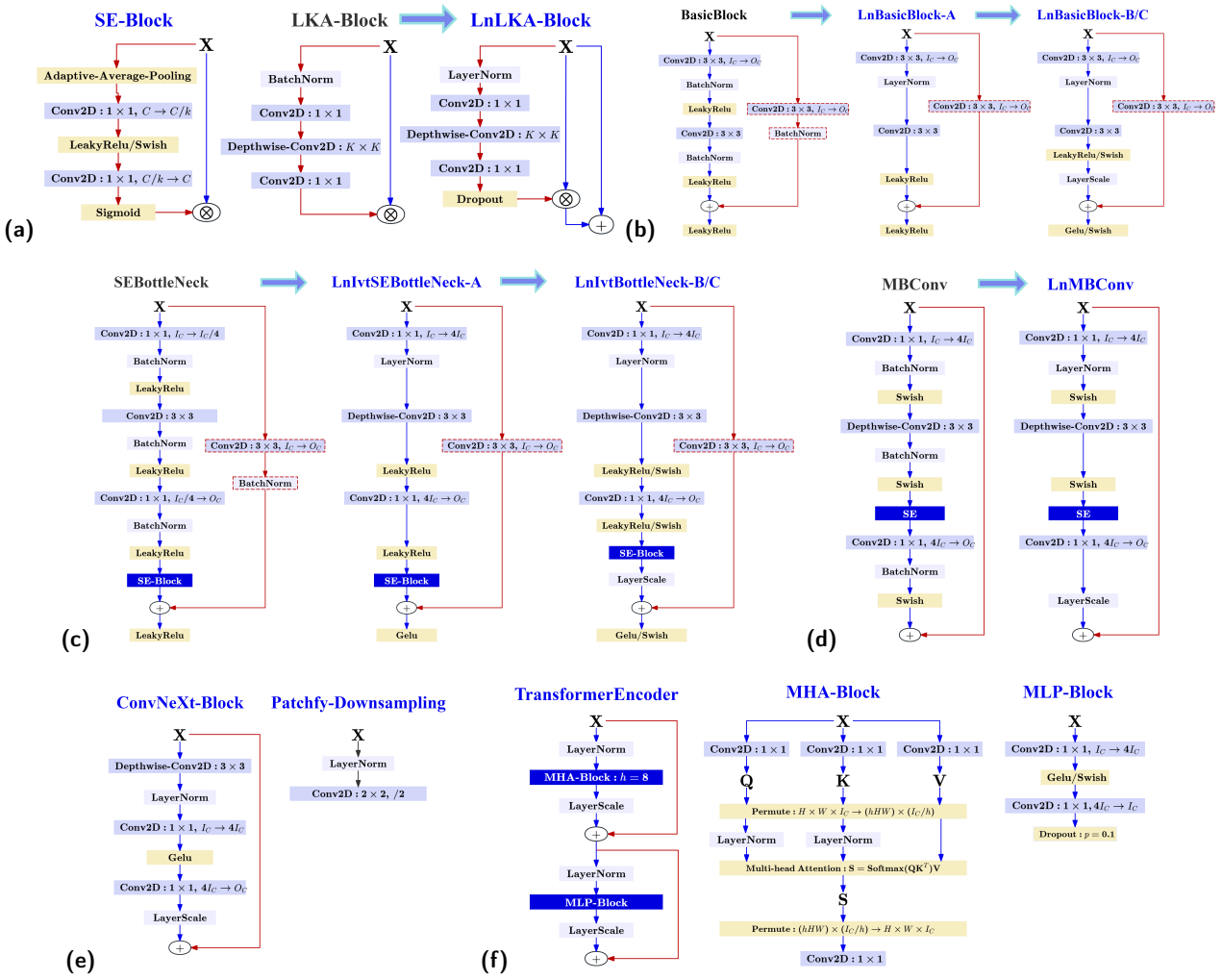

Figure 5: Backbone Block Architectures. H, W, and I_C denote the input height, width, and channel size, respectively. The main path and shortcut connection of the residual structure are in blue and red, respectively. Downsampling is performed on the shortcut connection using Conv2D layers, when the input and output dimensions differ.

This inconsistency arises because hyperspectral images interfere with BatchNorm's statistical estimation. During training, BatchNorm computes batch statistics (mean and variance) to update the running statistics for inference. Compared to RGB images, the huge channel sizes of hyperspectral images introduce much greater inter-sample variability, resulting in substantial differences in the statistics of different batches. Consequently, the accumulated running statistics fluctuate sharply, and can diverge from the testing-data statistics, causing the observed inconsistency.

To address this issue, we replace BatchNorm with Layer Normalization (LayerNorm), which is widely used in NLP and gains traction in CV models [28, 30, 31]. LayerNorm operates consistently in both training and inference, thus avoiding the statistical divergence in BatchNorm. Additionally, LayerNorm normalizes across the channel dimension, preserving the integrity of spectral features; whereas BatchNorm normalizes across spatial dimensions, potentially distorting spectral features. As shown in Figure 4 (b), we replace BatchNorm in ResNet48 with LayerNorm to construct LnResNet48-A, whose inference accuracy converges and aligns with the training loss.

3.3. Backbone Blocks

The backbone comprises many building blocks, designed with following principles to enhance performance:

(1) Residual Learning with LayerScale: Each block has a shortcut connection to facilitate convergence and prevent degradation. LayerScale [44] is integrated at the main path's output, to regularize feature magnitudes and suppress overfitting.

(2) Minimal Normalization and Activation Inspired by ConvNeXt [28] and Transformers [30, 31, 45], we simplify the main path to a single LayerNorm and activation functions, reducing information loss. Beyond LeakyRelu, we adopt Gelu [46] and Swish [47] for their smoother nonlinearity and input-adaptive gradients.

(3) Attention-Augmented Representation. To complement convolutions' local focus, we integrate multi-scale attention mechanisms. Squeeze-Excitation (SE) for channel-wise modeling, Large-Kernel-Attention (LKA) [39, 48–50] to capture long-range spatial interactions, and Multi-Head-Attention (MHA) [30, 31, 45] for global context modeling.

Figure 5 presents the block architectures:

(a) SE-Block and LnLKA-Block. The squeeze factor of SE-Block is reduced from 16 to 4 to enhance channel-wise attention.

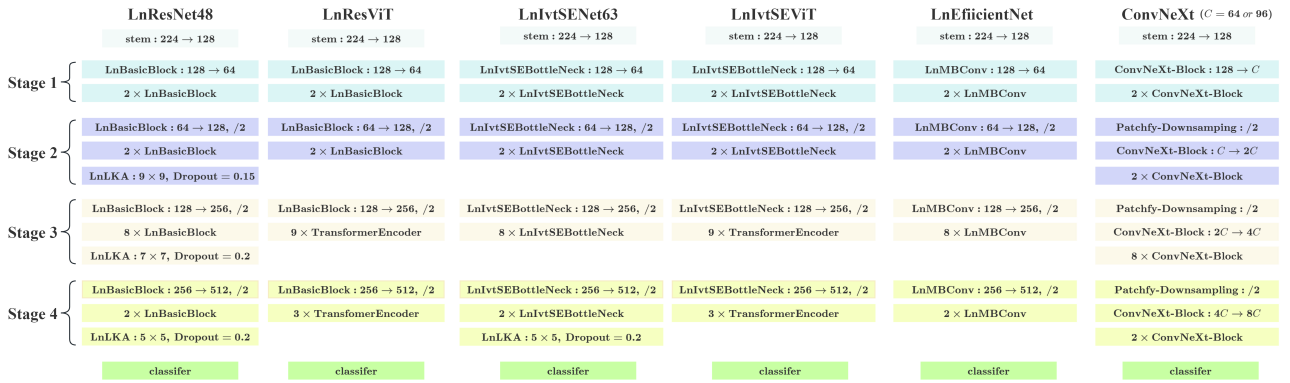

Figure 6: Complete Model Architectures. The backbone comprises four stages, respectively stacked by 3, 3, 9, and 3 fundamental blocks. '$div = 2$' represents halving the spatial height and width. $I_C \rightarrow O_C$ denotes modifying channel size from I_C to O_C.

LnLKA-block builds upon LKA by replacing BatchNorm with LayerNorm, and adding residual connection and Dropout.

(b) LnBasicBlocks. Variants of BasicBlock in ResNet [26]. LnBasicBlock-A replaces 2 BatchNorm layers with 1 LayerNorm layer, and reduces 3 LeakyRelu functions to 2. LnBasicBlock-B further integrates LayerScale, and replaces the last LeakyRelu with Gelu. LnBasicBlock-C replaces all activation functions with Swish.

(c) LnIvtBottleNecks. Variants of SEBottleNeck in SENet [32]. On the basis of SEBottleNeck, LnIvtBottleNeck-A replaces 3 BatchNorm layers with 1 LayerNorm layer, and reduces 4 activation functions to 3. Following the design of MobileNet [38, 51], the input-channel sizes of the main-path convolutions are inverted from ($I_C \rightarrow I_C/4 \rightarrow I_C$) to ($I_C \rightarrow 4I_C \rightarrow I_C$). LnIvtBottleNeck-B further adds LayerScale, and LnIvtBottleNeck-C replaces all activation functions with Swish.

(d) LnMBConv. A variant of MBConv in EfficientNet [27]. It replaces 3 BatchNorm layers with 1 LayerNorm layer, and introduces a LayerScale when the input and output shapes are the same to enable shortcut connections.

(e) ConvNeXt-Block and Patchfy-Downsampling.

(f) TransformerEncoder. The basic architecture of Transformers [30, 31, 45]. It includes a MHA-block and a MLP-Block. To prevent numerical overflow and stabilize training, We add 2 LayerNorm layers in MHA-block. We also add 2 LayerScale layers, one after the MHA-block and one after the MLP-block.

3.4. Vertical Layout Design

Figure 6 illustrates the vertical layout of complete models, with the stem and classifier positioned at the beginning and end. Inspired by ConvNeXt [28] and SwinTransformer [31], the backbone comprises four Stages with a fundamental-block depth of $(3, 3, 9, 3)$. In each Stage, the first block alters spatial and channel dimensions, while the subsequent blocks maintain identical input and output dimensions. Specifically, Stage 1 reduces channel size, but preserves spatial height and width to retain fine-grained features; whereas Stage 2-4 halve spatial height and width, but double the channel size. To balance generalization and capacity [33], TransformerEncoders and LnLKA-Blocks are integrated into Stage 2-4.

LnBasicBlocks, LnIvtBottleNecks, LnMBConv, and ConvNeXt-Block construct the backbones of LnResNet48, LnIvtSENet63, LnEfficientNet, and ConvNeXt, respectively. ConvNeXt-C64 and -C96 are two ConvNext configurations, with channel sizes in multiples of 64 and 96. For the first block in Stage 1 of LnIvtSENet63 and LnEfficientNet, the input-channel sizes of main-path convolutions are adjusted to ($128 \rightarrow 64 \rightarrow 64$), which reduces memory usage and enables 512 batch size. Based on LnResNet48 and LnIvtSENet63, LnResViT and LnIvtSEViT replace certain blocks in Stage 3 and 4 with 9 and 3 TransformerEncoders, to build a simplified Vision Transformer (ViT).

4. Training Framework

The training pipeline comprises two phases:

(1) Multi-crop Pre-training: The model learns transferable CFRs from the multi-crop dataset with a wider knowledge domain, capturing cross-crop underlying patterns to enhance generalization.

(2) Single-crop Fine-tuning: The pre-trained stem and backbone are transferred to a single-crop dataset, and recombined with a new classifier for fine-tuning. The freezing-unfreezing strategy preserves CFRs while adapting models to crop-specific nuances.

To improve training efficiency, we use sophisticated optimizers, address data imbalance, augment HSI data, carefully design hyperparameters, and accelerate model execution on GPUs.

4.1. Configurations

The models were trained on two RTX4090 GPUs, each with 24 GB global memory. The training setups are as follows:

Data Preparation. Pre-training utilizes both raw and corrected HSI samples for data augmentation. Fine-tuning only uses corrected samples to better approximate the target domain. HSI samples are augmented and then linearly scaled into interval $[-1, 1]$. Labels are in one-hot formats with 0.1 label-smoothing [52].

Class Balancing. Within a single-crop dataset or across multiple datasets, seed classes exhibit imbalanced sample counts: We solve this problem via: (1) Resampling equalizes per-class occurrence frequency. Pre-training selects randomly 4096 samples per class; Fine-tuning oversamples all classes to match the sample count of the largest class. This resampling is repeated every 5 epochs, with minimized data duplication. (2) Model performance is evaluated via Class Average Accuracy, which averages Top-1 accuracy across all classes. This metric is more representative than aggregate accuracy, in the presence of data imbalance.

Optimization Setup. Models are optimized using AdamW [53] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0.01$), with AMSGrad [54] to improve convergence. Except during warmup epochs, the

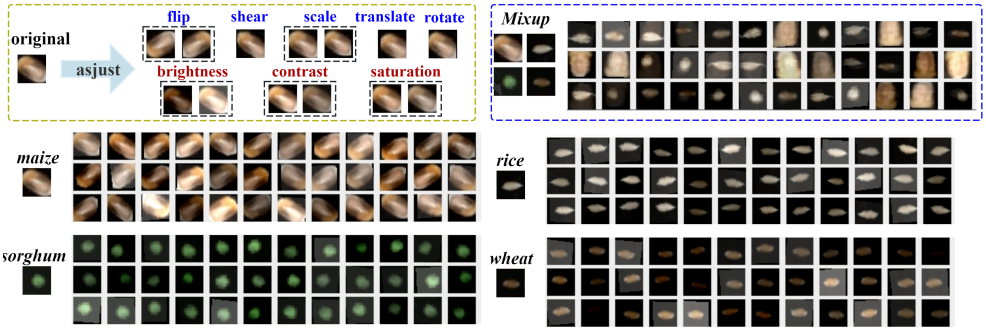

Figure 7: HSI Data augmentation. Four distributions represent augmented samples from maize, rice, sorghum, and wheat, using random spatial and color adjustment. Mixup generates a combined distribution for further augmentation. Hyperspectral images are visualized by extracting the 13-th, 80-th, and 132-rd channels to generate RGB images.

learning rate is scheduled using Cosine Annealing [55], with a minimum learning rate of 10^{-4} and a maximum iteration number of 20 epochs. For regularization, Stochastic Depth [56] is applied to blocks with identical input and output dimensions, linearly increasing the drop probability from 0 to 0.5.

Training Schedule. Models are saved at the epoch that achieves the best validation accuracy. Pre-training involves 570 epochs with a 5-epoch warmup, as detailed in Table 1. Fine-tuning setups are similar to those of Pre-training, To fully exploit model potential, fine-tuning aims for approximately 100% training accuracy in the final epochs. The backbone or stem is frozen during the first 50 epochs. This freezing strategy prevents model degradation, and typically results in higher validation accuracy than that in pre-training. For transformer-based models and sorghum datasets, we freeze both backbone and stem; for other cases, freezing backbone alone is sufficient to maintain adaptability and preserve CFRs.

Table 1
Pre-training Setups. The batch size is 1024 for LnResNet48, and 512 for the other models.

epoch count	initial learning rate	batch size	Mixup λ
50	1e-4	1024 / 512	1.00
5-epoch warmup: learning rate 1e-4 to 5e-4			
160	5e-4	1024 / 512	0.80
160	4e-4	1024 / 512	0.90
100	3e-4	1024 / 512	0.95
50	2e-4	1024 / 512	1.00
50	1e-4	1024 / 512	1.00

4.2. HSI Data Augmentation

To enhance model robustness, we extend RGB augmentation techniques to hyperspectral images, as illustrated in Figure 7.

Spatial Transformation. The magnitudes of random shearing, translation, scaling, and rotation are 0.1, 0.1, 0.05, and 0.05π, respectively. The probability of horizontal and vertical flips is 0.5.

Color Jitter. For an HSI sample $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, its brightness, contrast, and saturation are randomly adjusted as follows:

$$\mathbf{X} \leftarrow \mathbf{X} + 255k_b$$

$$\mathbf{X} \leftarrow \mathbf{X} + k_s(\mathbf{X} - \mathbf{X}_{\text{gray}})$$

$$\mathbf{X} \leftarrow (1 + k_c)\mathbf{X} - (1 - \sqrt{1 + k_c})\mathbf{X}_{\text{mean}}$$

Here, k_b, k_s, and k_c follow a uniform distribution in region $[0, 0.3]$, and they determine the degree of brightness, saturation, and contrast adjustment, respectively. $\mathbf{X}_{\text{mean}} \in \mathbb{R}^C$ is the channel-wise mean of $\mathbf{X}$, and $\mathbf{X}_{\text{gray}} \in \mathbb{R}^{H\times W}$ is the grayscale image produced by averaging $\mathbf{X}$ across spectral channels.

Mixup. We employ Mixup [57] to train models beyond empirical risk minimization. Mixup is applied to each batch of data, by linearly interpolating the shuffled batch with the original batch. While mixup can effectively improve generalization, it can impede model convergence. To achieve perfect convergence, Mixup is only applied during pre-training, with interpolation coefficient λ gradually increased from 0.8 to 1.0.

4.3. Hardware Acceleration

The Im2col-Winograd [58] and C-K-S [59] algorithms are employed to accelerate convolution layers. We optimize the depthwise convolutions to speedup LnLKA-Block. LayerScale, Dropout, and LayerNorm are fused with activation functions, reducing memory usage and bandwidth requirement by up to 33%.

Due to the much larger size of hyperspectral images over RGB images, loading HSI data from hard discs can be a significant bottleneck for training throughput. To fasten HSI loading, we utilize multiple threads to load and unpack batches of compressed HSI samples. Prior to loading, compressing HSI data reduces bandwidth requirement by about 75%. Besides, the loading and unpacking are executed in parallel, effectively hiding each other's latency. Consequently, the loading speed increases by nearly 4×, allowing GPUs to each over 90% of their peak power.

Although training the model requires tens of GPU hours, the resulting model enables efficient inference. On a single RTX 4090 GPU, the largest LnIvtSEViT model achieves can process about 2438 samples per second in FP32 precision, The throughput can be doubled using FP16 precision. This efficiency makes the model practical for real-time large-scale classification tasks.

5. Results and Analysis

To evaluate the effectiveness of CFR learning, we compare **CFR-driven models**, against **baseline models** trained on single-crop data. To isolate the impact of CFRs, both model types use identical training protocols, with two key differences:

(1) Pre-training data: A CFR-driven model is pre-trained on the multi-crop dataset to learn CFRs; a baseline model is pre-trained on a single-crop dataset.

Table 2
Model Accuracy. 'base' and 'CFR' denote baseline and CFR-driven models, respectively; 'test' and 'valid' represent the class average accuracy (%) on test and validation sets.

Model	Maize				Rice				Sorghum				Wheat			
	test		valid		test		valid		test		valid		test		valid	
	base	CFR	base	CFR	base	CFR	base	CFR	base	CFR	base	CFR	base	CFR	base	CFR
LnResNet48-A	91.0	**92.2**	94.0	**94.8**	85.1	**91.1**	87.0	**91.3**	90.6	**91.1**	89.4	**90.2**	97.2	**97.8**	96.5	**97.8**
LnResNet48-B	93.2	**95.1**	94.9	**96.1**	88.1	**92.6**	88.4	**92.5**	91.0	**91.5**	90.2	**91.0**	97.3	**98.5**	97.0	**98.0**
LnResNet48-C	93.6	**94.8**	95.1	**96.2**	87.7	**92.0**	88.7	**92.3**	91.2	**91.8**	90.4	**91.3**	97.2	**98.0**	96.9	**98.0**
LnIvtSENet63-A	91.0	**92.6**	93.4	**94.0**	84.8	**90.8**	85.2	**91.5**	89.3	**89.9**	89.0	**90.2**	96.5	**97.5**	96.9	**97.7**
LnIvtSENet63-B	92.7	**94.3**	94.7	**96.1**	89.4	**91.0**	90.3	**91.5**	91.3	**91.4**	90.7	**91.2**	97.7	**97.9**	97.7	**97.9**
LnIvtSENet63-C	92.7	**94.2**	95.1	**95.9**	88.9	**91.0**	90.3	**91.1**	91.8	**92.3**	90.5	**91.2**	97.2	**98.2**	97.7	**98.2**
ConvNeXt-C64	91.4	**92.5**	93.6	**94.0**	87.5	**89.0**	88.4	**89.7**	90.3	90.3	**90.3**	89.2	96.9	**98.0**	97.1	**97.5**
ConvNeXt-C96	90.3	**92.2**	92.6	**94.0**	87.3	**89.5**	88.2	**90.9**	90.5	**91.5**	89.7	**89.9**	97.1	**98.0**	97.0	**97.6**
LnEfficientNet	93.2	**93.9**	95.0	**95.4**	84.2	**89.1**	85.0	**90.5**	89.5	**90.4**	88.7	**90.0**	96.5	**98.1**	96.9	**97.9**
LnResViT-B	94.3	**95.7**	96.2	**97.0**	85.9	**91.6**	87.0	**90.7**	91.5	**91.7**	**91.0**	90.7	97.1	**98.4**	97.5	**98.0**
LnResViT-C	94.4	**95.4**	96.2	**97.1**	87.1	**91.2**	88.1	**92.1**	91.9	**92.6**	**91.4**	91.2	97.5	**98.6**	97.2	**97.8**
LnIvtSEViT-B	94.1	**96.3**	96.4	**97.3**	88.1	**91.3**	88.9	**91.9**	91.8	**91.9**	91.0	**91.6**	97.7	**98.1**	97.6	**97.8**
LnIvtSEViT-C	94.4	**95.9**	96.0	**97.2**	86.2	**91.3**	88.6	**91.9**	91.3	**92.1**	90.7	**91.9**	97.3	**98.4**	97.5	**97.9**
average	92.8	**94.2**	94.9	**95.8**	86.9	**90.9**	88.0	**91.4**	90.9	**91.4**	90.3	**90.7**	97.2	**98.1**	97.2	**97.9**
maximum	94.4	**96.3**	96.4	**97.3**	89.4	**92.6**	90.3	**92.5**	91.9	**92.6**	91.4	**91.9**	97.7	**98.6**	97.7	**98.2**
minimum	90.3	**92.2**	92.6	**94.0**	84.2	**89.0**	85.0	**89.7**	89.3	**89.9**	88.7	**89.2**	96.5	**97.5**	96.9	**97.5**

(2) Classifier handling: In fine-tuning, CFR-driven models use a new classifier; baseline models retain the pre-trained classifier.

We also compare our DNNs with traditional ML models, to demonstrate their high performance.

5.1. Accuracy Improvement

As shown in Table 2, CFR-driven models consistently outperform baseline models, demonstrating that CFR learning can enhance universal feature extraction in an architecture-agnostic manner. For maize, rice, sorghum, and wheat, CFR-driven models achieve the highest accuracy at 96.3%, 92.6%, 92.6%, and 98.6%, respectively. This phenomenon parallels NLP researches [60, 61], where cross-lingual pre-training similarly enhances downstream performance.

Table 3
Accuracy Improvement across all model architectures. 'Max-Accu Impr', 'Min-Accu Impr', and 'Avg-Accu Impr' denote the maximum, minimum, and average accuracy improvement of CFR-driven models over baseline models; 'test' and 'valid' represent the test set and validation set, respectively.

Crop	Max-Accu Impr		Min-Accu Impr		Avg-Accu Impr	
	test	valid	test	valid	test	valid
Maize	**1.9**	0.9	**1.9**	1.4	**1.4**	0.9
Rice	**3.2**	2.2	**4.8**	4.7	**4.0**	3.4
Sorghum	**0.7**	0.5	**0.6**	0.5	**0.5**	0.4
Wheat	**0.9**	0.5	**1.0**	0.6	**0.9**	0.7
average	**1.7**	1.0	**2.1**	1.8	**1.7**	1.4

This systematic improvement, summarized in Table 3, reveals two advantages of CFR-driven models:

(1) Superior Generalization. CFR-driven models achieve higher improvement on test set (1.7%) than validation set (1.4%). This indicates that they have a smaller bias across different data partitions, potentially more generalizable to unseen data.

(2) Enhanced Stability. CFR-driven models have more consistent performance. First, their minimum accuracy improvement (2.1%) exceeds the maximum improvement (1.7%), indicating reliable performance floors. Second, their accuracy has a lower standard deviation for most crops (rice: 1.08% vs. 1.6%; sorghum: 0.81% vs. 0.84%; wheat: 0.31% vs. 0.38%), with only a negligible exception for maize (1.46% vs. 1.44%).

This cross-domain effectiveness of CFRs leads us to examine the crop-specific behavior. While CFR-driven models show strong accuracy improvements for maize and rice, their improvements for wheat and sorghum are relatively modest due to following reasons.

(3) Wheat: baseline ceiling effects. The already-high baseline (96.5% to 97.7%) leaves limited optimization space. Besides, the trace contaminants in samples and spectral saturation effects physically constraints accuracy improvement. Nevertheless, CFRs reduce misclassification rate by 0.9% on average, a notable 32.1% relative reduction,

(4) Sorghum: feature scarcity. As the smallest seeds, sorghum contains fewer discriminative features, which physically limits the transfer potential of CFRs. Further enhancements could incorporate more similar crops (e.g. millet, Quinoa) to enlarge common feature intersections.

5.2. Convergence Improvement

CFR-driven models exhibit faster convergence over baseline models, as presented in Figure 8, where baseline accuracy is calculated by averaging per-class accuracy across all four crops.

This faster convergence is attributed to the wider knowledge domain of multi-crop dataset, which improves gradient quality. Even though multi-crop data presents more intricate decision boundaries, high-capacity DNNs effectively navigate this complexity. Unlike similar and harder-to-distinguish single-crop data, cross-crop differences provide clearer signals for discrimination, especially in the early training stages. With a greater number of classes, multi-crop data promotes more competitive probabilities, sharpening gradients of Cross Entropy loss; besides, each batch tends to contain more devise data, reducing the potential for biased updating that favors specific classes. Additionally, the larger volume of multi-crop dataset enables CFR-driven models to explore a wider range of parameter space within an epoch.

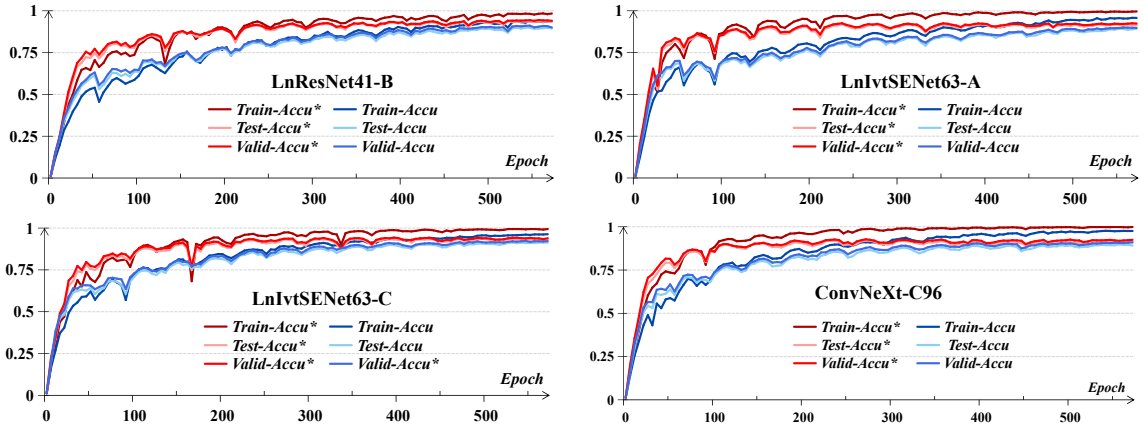

Figure 8: Pre-training convergence behavior. 'Train-Accu', 'Test-Accu', and 'Valid-Accu' denote the accuracy (%) on training, test, and validation sets. Markers with * indicate CFR-driven models; unmarked indicate baseline models.

For larger model architectures like LnIvtSEViT, we even need to triple the epoch count to train baseline models on simpler wheat or sorghum datasets, to achieve nearly 100% training accuracy. Despite using fewer iterations than baseline models, CFR-driven models not only achieve nearly 100% training accuracy, but also provides higher accuracy on test set. This demonstrates CFR learning as an efficient method to accelerate cross-crop modeling.

Moreover, multi-crop pre-training enhances training robustness. Specifically, certain baseline models (without LayerScale) may suffer from one-class dominance (100% accuracy for one class, 0% for the others), when pre-trained on simpler wheat and sorghum datasets. However, this problem never occurs with CFR-driven models, suggesting that increased data complexity prevents models from converging to bad local minima.

5.3. Fine-Tuning Impact

After fine-tuning, CFR-driven models show higher accuracy on both test and validation sets, with slightly higher validation improvements as summarized in Table 4. This confirms enhanced crop-specific performance.

Sorghum shows more notable improvements (0.8%), compared to maize (0.2%), rice (0.1%), and wheat (0.1%), This suggests a greater degree of shared features among maize, rice, and wheat, whereas sorghum features are more crop-specific. This aligns with observations in Section 5.1, where sorghum accuracy benefits less from CFR learning.

Table 4
Fine-tuning improvement for CFR-driven models. 'Avg-Accu', 'Max-Accu', and 'Min-Accu' represent the average, maximum, and minimum accuracy (%); 'test' and 'valid' indicate test and validation sets, respectively.

	Maize		Rice		Sorghum		Wheat	
	test	valid	test	valid	test	valid	test	valid
Avg-Accu	94.0 + 0.2	95.5 + **0.3**	90.8 + 0.1	90.9 + **0.5**	90.6 + 0.8	89.9 + 0.8	98.0 + 0.1	97.4 + **0.5**
Max-Accu	96.3 + 0.0	97.3 + 0.0	92.5 + 0.1	92.3 + **0.2**	92.4 + 0.2	91.6 + **0.3**	98.4 + 0.2	97.9 + **0.3**
Min-Accu	91.5 + **0.7**	93.8 + 0.2	88.9 + 0.1	89.6 + 0.1	87.4 + 2.5	86.5 + **2.7**	97.5 + 0.0	96.7 + **0.8**

We observed that CFR-driven models degrade without freezing the backbone or stem during the initial 50 epochs. As exemplified in Figure 9 (a), the frozen LnResNet48-B achieves higher accuracy stability, compared to the unfrozen counterpart. This initial freezing period allows the model to optimize the new classier, while preventing the pre-trained components from losing learned CFRs. Subsequently, unfreezing enables finer-grained adaptation to the target crop. As depicted in Figure 9 (b), LnResNet48-B accuracy ultimately peaks at 92.6% after a temporary dip from 92%. The final 92.6% is higher than the 92% achieved in freezing period.

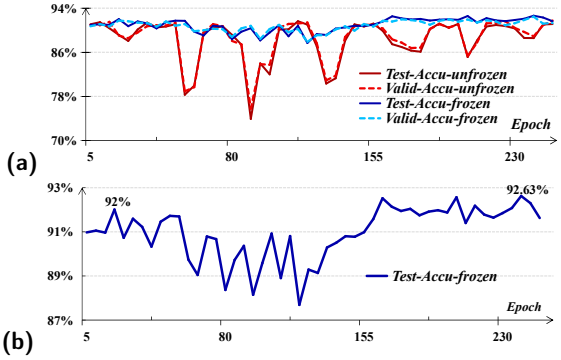

Figure 9: Accuracy variation of LnResNet48-B in fine-tuning. 'Test-Accu' and 'Valid-Accu' denote the accuracy on test and validation sets, respectively.

5.4. Comparison to Machine Learning Models

We further compare CFR-driven DNNs against traditional ML models, that rely on spectral preprocessing and feature selection:

ML Models: SVM, RF, KNN, Logistic regression (LR), and eXtreme Gradient Boosting (XGB).

Spectral Preprocessing: Savitzky-Golay Smoothing, multiplicative scatter correction, and standard normal variate.

Feature Selection: Competitive adaptive reweighted sampling, uninformative variable elimination, successive projections algorithm, random frog, and genetic algorithm.

Table 5 lists the maximum accuracy achieved by each ML model across all preprocessing and feature selection techniques.

CFR-driven DNNs significantly outperform ML models, achieving 7.1% to 26.6% accuracy gains across all crops. For simpler sorghum (7 class) and wheat (6 class) classifications, ML models

Table 5
Maximum Accuracy of Machine Learning Models. 'ML max' and 'DL max' denote the maximum accuracy (%) among all ML models and our CFR-driven DL models, respectively.

Model	Maize	Rice	Sorghum	Wheat
KNN	53.2	58.3	72.8	74.3
LR	67.1	62.8	79.4	76.9
RF	65.8	59.2	77.6	83.5
SVM	25.3	41.4	61.1	65.3
XGB	69.7	68.4	85.5	88.2
ML max	69.7	68.4	85.5	88.2
DL max	**96.3**	**92.6**	**92.6**	**98.6**

attain moderate accuracy of 85.5% and 88.2%, but CFR-driven DNNs still relatively reduce misclassification rates by 49% and 88%. For more complex maize (47 class) and rice (19 class) classifications, the accuracy of ML models crops considerably to 69.7% and 68.4%, while CFR-driven DNNs maintain high performance of 96.3% and 92.6%, relatively reducing misclassification rates by 88.8% and 76.6%. Therefore, the high capacity of DNNs is a key factor to improve accuracy, especially for complex tasks.

5.5. Extended Analysis

Why CFR Learning works. Training a model via gradient descent can be conceptualized as searching for suitable local minima within the parameter space. Similar to how adding more equations reduces the solution space, richer data imposes more constraints on the model's parameters. Consequently, combining single-crop datasets into the multi-crop dataset implicitly prunes the parameter space, restricting model optimization to a more promising region. This leads to more informed navigation and smoothed loss landscape, shortening the path towards favorable local minima. This pre-training provides effective starting points for fine-tuning, enabling it to discover superior solutions for a specific crop.

Scaling Law for HSI Deep Learning. This work provides empirical validation of the scaling law in two aspects. First, CFRs bridge multi-crop data together to expand data size and knowledge domain, demonstrably improving model accuracy, robustness, and convergence. Second, given sufficient training data, large DNNs with substantial parameters consistently outperform small ML models, especially in complex tasks. This highlights that while high-capacity models are essential for capturing intricate patterns in HSI data, their effectiveness relies on sufficient data volume.

6. Conclusion

In this work, we exploit CFRs to integrate multi-crop and enlarge knowledge domain, facilitating the training of high-capacity models. To efficiently process full HSI data end-to-end, we build deeper convolution-attention DNNs with HSI-specific modifications. We also design a sophisticated CFR-driven training frame with hardware acceleration, enabling effective HSI modeling. Finally, we analyze how CFR learning enhances model performance. Our results demonstrate that CFR learning is an effective approach to scaling up data richness and model capacity, which are important for accurate and efficient crop identification.

Future researches include exploring state-of-the-art architecture, optimizing training frame, more innovative data-integration methods, and novel upstream pre-training techniques.

References

[1] Yufei Ge, Shaozhong Song, Shuang Yu, Xiaoli Zhang, and Xiongfei Li. Rice seed classification by hyperspectral imaging system: A real-world dataset and a credible algorithm. *Computers and Electronics in Agriculture*, 219:108776, 2024.

[2] Wenwen Kong, Chu Zhang, Fei Liu, Pengcheng Nie, and Yong He. Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors*, 13(7):8916–8927, 2013.

[3] Beyza Çiftci, Necati Çetin, Seda Günaydın, and Mahmut Kaplan. Machine learning approaches for binary classification of sorghum (sorghum bicolor l.) seeds from image color features. *Journal of Food Composition and Analysis*, 140:107208, 2025.

[4] Zhiyong Zou, Jiangbo Zhen, Qianlong Wang, Qingsong Wu, Menghua Li, Dongyu Yuan, Qiang Cui, Man Zhou, and Lijia Xu. Research on nondestructive detection of sweet-waxy corn seed varieties and mildew based on stacked ensemble learning and hyperspectral feature fusion technology. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 322:124816, 2024.

[5] Kai Wu, Tingyu Zhu, Zhiqiang Wang, Xuerong Zhao, Ming Yuan, Du Liang, and Zhiwei Li. Identification of varieties of sorghum based on a competitive adaptive reweighted sampling-random forest process. *European Food Research and Technology*, 2023.

[6] Seda Günaydın, Ewa Ropelewska, Kamil Sacilik, and Necati Çetin. Exploration of machine learning models based on the image texture of dried carrot slices for classification. *Journal of Food Composition and Analysis*, 129:106063, 2024.

[7] Qingyun Liu, Zuchao Wang, Yuan Long, Chi Zhang, Shuxiang Fan, and Wenqian Huang. Variety classification of coated maize seeds based on raman hyperspectral imaging. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 270:120772, 2022.

[8] Dong Yang, Yuxing Zhou, Yu Jie, Qianqian Li, and Tianyu Shi. Non-destructive detection of defective maize kernels using hyperspectral imaging and convolutional neural network with attention module. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 313:124166, 2024.

[9] Tian Hu, Yineng Chen, Di Li, Chenfeng Long, Zhidong Wen, Rong Hu, and Guanghui Chen. Rice variety identification based on the leaf hyperspectral feature via lpp-svm. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(15):2350001, 2022.

[10] Yanlin Wei, Xiaofeng Li, Xin Pan, and Lei Li. Nondestructive classification of soybean seed varieties by hyperspectral imaging and ensemble machine learning algorithms. *Sensors*, 20(23), 2020.

[11] Gözde Özdoğan and Aoife Gowen. Wheat grain classification using hyperspectral imaging: Concatenating vis-nir and swir data for single and bulk grains. *Food Control*, 168:110953, 2025.

[12] Shouguo Zheng, Chaohui Guo, Debao Tu, Jianpeng Xu, Shizhuang Weng, and Gongqin Zhu. Spectral super-resolution for high-accuracy rice variety classification using hybrid cnn-transformer model. *Journal of Food Composition and Analysis*, 137:106891, 2025.

[13] Xin Zhao, Haotian Que, Xiulan Sun, Qibing Zhu, and Min Huang. Hybrid convolutional network based on hyperspectral imaging for wheat seed varieties classification. *Infrared Physics and Technology*, 125:104270, 2022.

[14] Haotian Que, Xin Zhao, Xiulan Sun, Qibing Zhu, and Min Huang. Identification of wheat kernel varieties based on hyperspectral imaging technology and grouped convolutional neural network with feature intervals. *Infrared Physics and Technology*, 131:104653, 2023.

[15] Jinliang An, Chen Zhang, Ling Zhou, Songlin Jin, Ziyang Zhang, Wenyi Zhao, Xipeng Pan, and Weidong Zhang. Tensor based low rank representation of hyperspectral images for wheat seeds varieties identification. *Computers and Electrical Engineering*, 110:108890, 2023.

[16] Jingwu Zhu, Hao Li, Zhenhong Rao, and Haiyan Ji. Identification of slightly sprouted wheat kernels using hyperspectral imaging technology and different deep convolutional neural networks. *Food Control*, 143:109291, 2023.

[17] Liu Zhang, Jinze Huang, Yaoguang Wei, Jincun Liu, Dong An, and Jianwei Wu. Open set maize seed variety classification using hyperspectral imaging coupled with a dual deep svdd-based incremental learning framework. *Expert Systems with Applications*, 234:121043, 2023.

[18] Liu Zhang, Shubin Zhang, Jincun Liu, Yaoguang Wei, Dong An, and Jianwei Wu. Maize seed variety identification using hyperspectral imaging and self-supervised learning: A two-stage training approach without spectral preprocessing. *Expert Systems with Applications*, 238:122113, 2024.

[19] Zhihua Diao, Jiaonan Yan, Zhendong He, Suna Zhao, and Peiliang Guo. Corn seedling recognition algorithm based on hyperspectral image and lightweight-3d-cnn. *Computers and Electronics in Agriculture*, page 107343, 2022.

[20] Zhihua Diao, Peiliang Guo, Baohua Zhang, Jiaonan Yan, Zhendong He, Suna Zhao, Chunjiang Zhao, and Jingcheng Zhang. Spatial-spectral attention-enhanced res-3d-octconv for corn and weed identification utilizing hyperspectral imaging and deep learning. *Computers and Electronics in Agriculture*, 212:108092, 2023.

[21] Xinjun Hu, Mingkui Dai, Jianheng Peng, Jiahao Zeng, Jianping Tian, and Manjiao Chen. Rapid sorghum variety identification by hyperspectral imaging combined with super-depth-of-field microscopy. *Journal of Food Composition and Analysis*, 137:106930, 2025.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[23] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[24] Shujia Li, Laijun Sun, Yujie Tian, Xiaoli Lu, Zhongyu Fu, Guijun Lv, Lingyu Zhang, Yuantong Xu, and Wenkai Che. Research on non-destructive identification technology of rice varieties based on hsi and gbdt. *Infrared Physics and Technology*, 142:105511, 2024.

[25] Haoping Huang, Xinjun Hu, Jianping Tian, Xinna Jiang, Ting Sun, Huibo Luo, and Dan Huang. Rapid and nondestructive prediction of amylose and amylopectin contents in sorghum based on hyperspectral imaging. *Food Chemistry*, 359:129954, 2021.

[26] Kaing He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[27] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022.

[29] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of 9th International Conference on Learning Representations (ICLR)*, 2021.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[32] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[33] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977, 2021.

[34] Cheng Cui, Jingzhu Wu, Qian Zhang, Le Yu, Xiaorong Sun, Cuiling Liu, and Yi Yang. Maturity detection of single maize seeds based on hyperspectral imaging and transfer learning. *Infrared Physics and Technology*, 138:105242, 2024.

[35] Na Wu, Shizhuang Weng, Qinlin Xiao, Hubiao Jiang, Yun Zhao, and Yong He. Rapid and accurate identification of bakanae pathogens carried by rice seeds based on hyperspectral imaging and deep transfer learning. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 311:123889, 2024.

[36] Hongfei Zhu, Yifan Zhao, Lianhe Yang, Longgang Zhao, and Zhongzhi Han. Pixel-level deep spectral features and unsupervised learning for detecting aflatoxin b1 on peanut kernels. *Postharvest Biology and Technology*, 202:112376, 2023.

[37] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[39] Menghao Guo, Chengze, Zhengning Liu, Mingming Cheng, and Shimin Hu. Visual attention network. 2023.

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. 15(1), 2014.

[41] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 448–456. ACM, 2015.

[42] Sheng Shen, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. PowerNorm: Rethinking Batch Normalization in Transformers. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

[43] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[44] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve Jegou. Going deeper with image transformers. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[46] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023.

[47] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017.

[48] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11963–11975, June 2022.

[49] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Karkkainen, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Boqian Wu. More Convnets in the 2020s: Scaling

up Kernels Beyond 51x51 Using Sparsity. In *Proceedings of 11th International Conference on Learning Representations (ICLR)*, 2023.

[50] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5513–5524, 2024.

[51] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[52] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[53] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[54] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[55] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[56] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 646–661, Cham, 2016. Springer International Publishing.

[57] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[58] Zhiyi Zhang, Pengfei Zhang, Zhuopin Xu, Bingjie Yan, and Qi Wang. Im2col-winograd: An efficient and flexible fused-winograd convolution for nhwc format on gpus. In *Proceedings of the 53rd International Conference on Parallel Processing*, ICPP '24, page 1072–1081, 2024.

[59] Zhiyi Zhang, Pengfei Zhang, Zhuopin Xu, and Qi Wang. Reduce computational complexity for convolutional layers by skipping zeros. In *2023 IEEE 30th International Conference on High Performance Computing, Data, and Analytics (HiPC)*, pages 347–356, 2023.

[60] Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[61] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021.