AV4GAInsp: An Efficient Dual-Camera System for Identifying Defective Kernels of Cereal Grains

Lei Fan[1,2,†], Dongdong Fan[1], Yiwen Ding[1], Yong Wu[1], Hongxia Chu[1], Maurice Pagnucco[2], Yang Song[2,†]

Abstract— Grain Appearance Inspection (GAI) is a prerequisite for grain quality determination, providing guidance for grain processing, storage and trade. GAI is routinely performed by trained inspectors who are required to visually inspect cereal grains for each individual kernel. Since grain kernels (e.g., wheat, rice) are tiny with heterogeneous shapes and appearance, manually performing GAI is time-consuming and error-prone. This paper presents a machine vision-based customization of an automated system for grain appearance inspection, called AV4GAInsp, which consists of a device and an analysis framework. The device is equipped with an elaborate feeding module and a capturing module for automatically pre-processing grain kernels and efficiently acquiring high-quality images for these kernels. The framework employs deep convolutional neural networks to process these captured images to classify the kernels as normal or defective. We also built and released a large-scale dataset, named GrainDet, that includes over $140K$ images for three types of grains: wheat, sorghum and rice. Comprehensive experiments are conducted to validate the efficacy and performance of our AV4GAInsp system, achieving an average F1-score of 98.4% and excelling at inspection efficiency by over 20× speedup. Kappa statistic tests are performed to confirm the consistency between our system and human experts. It is expected that AV4GAInsp will alleviate inspectors' workloads and inspire further research in smart agriculture. The project can be found at https://github.com/hellodfan/GrainDet.

I. Introduction

Cereal grains play a fundamental role in human nutrition and development. The assurance of grain quality is crucial for safeguarding human health, ensuring food security, facilitating international trade and contributing to "good health and well-being" [1]. Grain Appearance Inspection (GAI) is one of the prerequisites for high-throughput grain quality determination. The aim of GAI is to classify normal or defective kernels according to cereal vocabulary standards [2]. Kernels can be categorized as normal, impurities or six types of defective grains (see Table I). GAI thus provides crucial guidance for grain stratification [3] and processing [4], and serves as a pivotal determinant of grain pricing [5].

Typically, GAI is performed manually based on the visual attributes of grain kernels, such as color and shape [6]. This process serves as an inspection step that mainly involves the inspection of a representative volume of grain kernels without the removal of defective kernels. Once harvested, raw kernels are delivered to grain depots or processing plants, in which most defective kernels are filtered out with color sorter machines. Representative samples are extracted from granaries or freighters with specialized sampling probes. These kernels are then examined by experienced inspectors with a variety of hand tools, such as wide and fine-mesh sieves to isolate extra matter and tweezers to pick out impurities (see Fig. 1). Take a shipment of wheat grains as an example. The process involves taking a laboratory sample (60 grams (g), near 1500 kernels) and inspecting each of them manually. Therefore, even qualified inspectors (with over 5 years of experience) require upwards of 25 minutes to complete this inspection. This time-consuming inspection poses three significant challenges: 1) during the harvesting season, tons of cereal grains are transported into supply centers and these grains must be stored immediately due to various weather conditions and trading markets, which means that inspectors are required to expedite the inspection; 2) as grain kernels are small in physical size with subtle differences between normal and defective kernels, performing GAI is difficult and labor-intensive, imposing huge demands on inspectors' concentration; 3) since inspectors have varying work experience, manual inspections conducted by different inspectors can be subjective and error-prone.

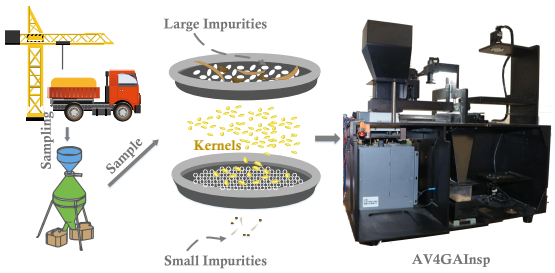

Fig. 1. The overview of grain appearance inspection. Harvested grains are conveyed to the supply centers, where representative samples are selected and pre-processed. These samples are fed into our AV4GAInsp for inspection.

Recently, deep learning techniques have revolutionized vision-based systems in multiple fields including medical applications [7], [8], fruit harvest [9], [10] and crop monitoring [11], [12]. In the crop-grain-process streamline,

† Corresponding Author.
[1]Lei Fan, Dongdong Fan, Yiwen Ding, Yong Wu and Hongxia Chu are with Gaozhe Technology, China
[2]Lei Fan, Maurice Pagnucco and Yang Song are with School of Computer Science and Engineering, University of New South Wales, Australia {lei.fan1,yang.song1}@unsw.edu.au

many studies have been dedicated to intelligent planting and harvest [13], monitoring growth [14], food processing [15] and food safety [16]. However, less attention [17], [18], [19], [20] has been focused on the inspection of grain quality. Therefore, it is beneficial to develop an automated system for GAI with the aid of deep learning techniques.

We aim to develop an automated, streamlined and cost-effective system for GAI that surpasses human inspectors in terms of efficiency (i.e., inspection time cost <2 minutes). However, implementing such a system is challenging. Due to the large number and small physical size of cereal grains, kernels are easily stacked with occlusion, whereas each individual grain kernel must be inspected carefully and efficiently. On the other hand, the differences between normal and defective kernels are subtle. For example, some crucial attributes (e.g., mouldy points) are even smaller than $1{\times}1{\times}1mm^3$. In addition, cereal grains manifest in heterogeneous shapes and varying appearances but belong to the same type being normal or defective. These challenges impose difficult requirements for the recognition algorithms.

To address these challenges, in this work, we introduce an Automated Vision-based system for Grain Appearance Inspection named AV4GAInsp, consisting of a prototype device and a deep learning-based analysis framework, as illustrated in Fig. 3. Specifically, our device is designed to capture high-quality visual appearance information from various types of grain kernels. To prevent occlusion between kernels, our device utilizes a feeding module equipped with a vibration belt, enabling simultaneous processing of multiple kernels within a single batch. The device also incorporates a capturing module comprising two industrial camera units, enabling the acquisition of high-resolution and wide field-of-view images from dual perspectives for a batch of kernels. The analysis framework integrates advanced deep learning techniques to analyze the captured images, which can precisely identify whether kernels are normal, have impurities, or are one of six types of defective kernels. We conducted experiments to explore the most time-efficient strategy for our device, verifying the effectiveness and efficiency of our AV4GAInsp system. Moreover, we conducted Kappa statistical tests to assess the consistency between our system and human experts, thereby demonstrating the feasibility and scalability of our system. The main contributions can be summarized as follows:

- We developed an automated vision-based system for GAI, named AV4GAInsp. It comprises a device for capturing visual information of grain kernels efficiently and a deep learning-based framework for analyzing captured images.
- The prototype device can accommodate a variety of grain kernels and capture high-quality visual images for various types of cereal grains in an efficient manner.
- Our extensive experiments demonstrate the advantage of our system compared to human experts.
- We released a dataset, named GrainDet, including over $140K$ images for wheat, sorghum and rice.

II. Automated system design

A. Mechatronic Design of Our Prototype Device

Our device consists of three primary parts: a feeding module, a capturing module and a transparent plate, as shown in Fig. 3. Considering the small physical size of kernels and their subtle visual attributes, the core concept revolves around our device: 1) the feeding module for separating kernels, 2) the capturing module for acquiring visual information, and 3) the transparent plate for delivering and recycling kernels.

The feeding module is designed to pre-process grain kernels, separating them from each other. This module includes four essential components: 1) material entrance, 2) vibration plate, 3) feeding gate and 4) flow camera, as shown in Fig. 2. The shape of the material entrance resembles a square funnel and there is a rectangular cabin below which can hold all grain kernels conveniently and store them. The feeding gate is controlled by a 24V DC stepper motor. It controls the adjustment of the gate height according to the type of cereal grain being processed which effectively prevents the kernels from stacking up. The vibration plate consists of a vibrator and a rectangular groove with an area of $160{\times}90$ mm^2 from a top-down view. The vibrator is powered by a 220V AC high-frequency inductance coil, allowing for precise control over frequency f_v and duration t_v. Kernels are separated and oscillated along the plate toward the subsequent modules.

Furthermore, the feeding module incorporates a flow camera to monitor the feeding status and provide feedback signals to adjust both the vibration plate and feeding gate. The camera resolution is $1276{\times}1016$ pixels, the focal length is 4 mm, and the object length is 110 mm. The field-of-view (FoV) spans $141{\times}131$ mm^2. The capturing module is built to acquire high-quality visual information for grain kernels efficiently. Considering the trade-off among the visible regions, manufacturing cost and complexity in the mechanical structure, we select a dual-camera strategy (see Fig. 3). Two cameras are

module.pdf module.bb

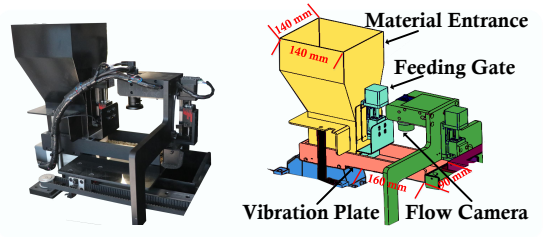

Fig. 2. The mechanical configuration of the feeding module.

TABLE I

Wheat and sorghum examples of normal, defective grains and impurities (abbreviations used in the subsequent content).

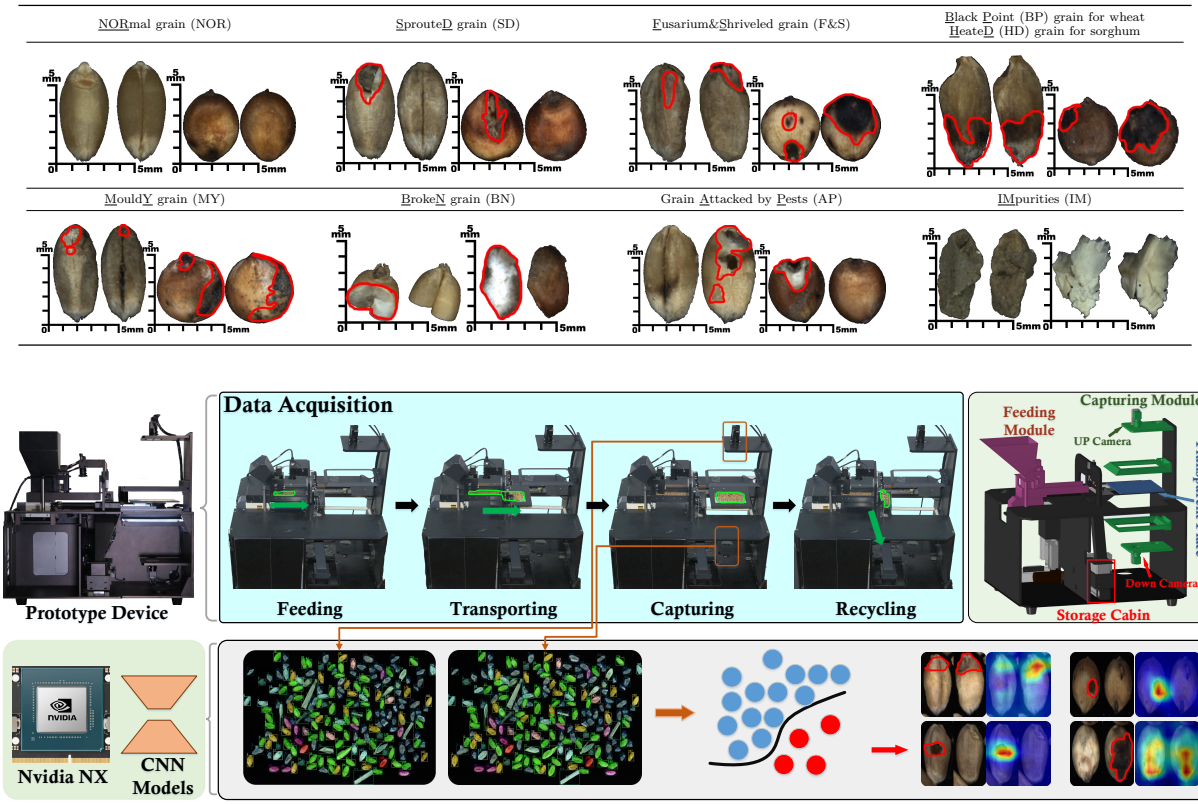

Fig. 3. The AV4GAInsp system includes a prototype device and an analysis framework. Data acquisition includes: feeding, transporting, capturing and recycling phases. Two high-resolution images I_{up} and I_{do} are obtained for a batch of grain kernels, and these images are analyzed by our CNN models.

positioned in a vertical arrangement, with one placed above and the other below the transparent plate. They are referred to as UP and DOWN cameras respectively.

Each camera has high specifications, including a resolution of 5480×3648 pixels, a focal length of 25 mm, an object length of 270 mm, an effective camera FoV reaching 131×87 mm^2 and close to 860 dots-per-inch (DPI). The flow camera and UP and DOWN cameras are calibrated using color cards to fine-tune the exposure time and RGB gains of imaging. In addition, to capture high-quality images and mitigate the influence of environmental factors (e.g., dust), each camera is equipped with a rectangular ring light to provide uniform, non-shadow illumination for optimal image clarity. Each ring light has a color temperature of 7000 K (Kelvin) and powered by a 12V DC controller.

B. Data Acquisition

According to the flow of grain kernels, data acquisition can be divided into four phases, as illustrated in Fig. 4:

Feeding: The laboratory samples are sent into the feeding module and all kernels are stored within the cabin of the material entrance. The feeding gate and vibrating plate collaborate to shake the grains to propel them toward subsequent modules. By performing image processing techniques on images captured by the flow camera, the approximate number of feeding volumes N_a in the vibration plate can be obtained. N_a can be fine-tuned by employing a Proportional-Integral (PI) controller among the frequency f_v and duration t_v of the vibrator. Finally, the feeding module can extract and divide grain kernels into multiple batches.

Transporting: For each batch, the transparent plate is positioned below the vibration plate, serving as a vessel to hold the kernels and move to the capturing module. This plate is mounted on a circular conveyor belt. While the vibration plate shakes kernels, the transparent plate steadily moves in the direction of grain flow until it holds a single batch of grains. Subsequently, the transparent plate swiftly moves to the middle of the capturing module.

Capturing: After the transparent plate remains fixed in the middle of the capturing module, the UP ring light is activated for 0.3 seconds (s), allowing the UP camera to capture an image I_{up}. Following that, the DOWN ring light is activated for 0.3s, enabling the DOWN camera to capture an image I_{do}. These two images are paired for a batch of kernels and will be used in the data analysis framework.

Recycling: The transparent plate cycles back to the

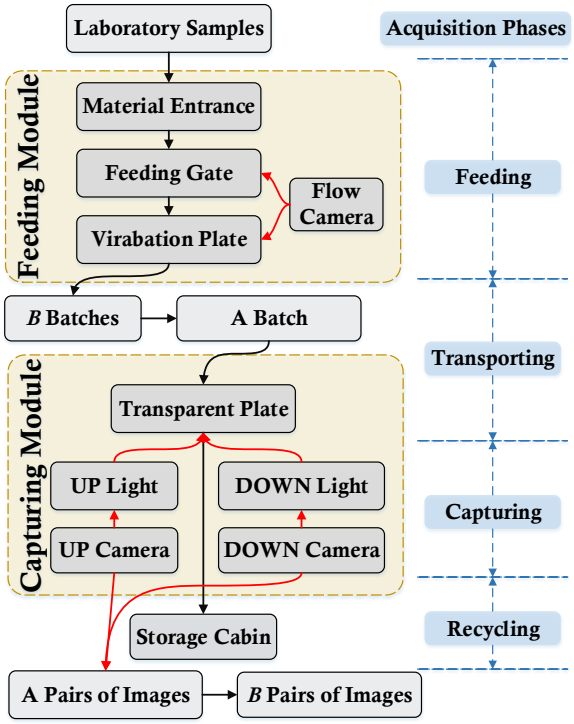

Fig. 4. Workflow of our device. Black arrows indicate the kernel flow, reds and blues arrows denote the control signals and acquisition phases respectively.

base of the vibration plate. A brush is placed beneath the vibrating plate to gather all grain kernels on the transparent plate and guides them into the storage cabin.

For a typical laboratory sample (e.g., 60 g), all grain kernels are input into the feeding module where they are sorted and divided into B batches. For each batch $b \in \{1, \ldots, B\}$, we cyclically perform the transporting, capturing and recycling phases to capture images. As a result, we can capture B pair of UP (I_{up}) and DOWN (I_{do}) images.

C. Deep Learning-based Analysis Framework

Deep learning, especially Convolutional Neural Networks (CNNs), has achieved remarkable progress in many computer vision tasks [21], [22], [23], [24]. When applying these models to analyze captured images, it may seem straightforward to employ off-the-shelf detection models but our experiments indicate that these models yielded moderate performance. We consider that there are three main challenges: 1) the raw I_{up} and I_{do} images possess very high resolutions of 5480×3648 pixels, and every individual kernel has two views from UP and DOWN angles; 2) the differences between normal and defective kernels are subtle; and, 3) cereal grains of the same normal or defective type may exhibit heterogeneous shapes and varying appearance. To tackle these issues, we propose a three-stage (detection - pairing&rotation - recognition) analysis framework (see Fig. 5). During

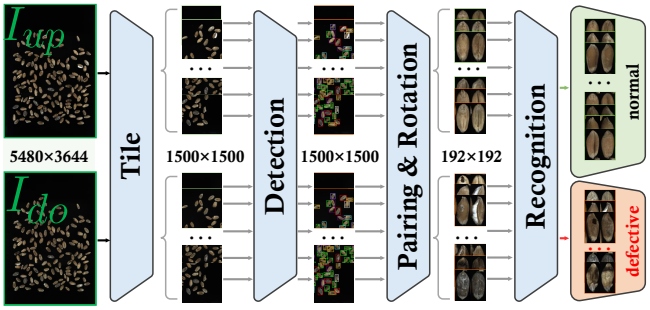

Fig. 5. Our three-stage framework for a batch of kernels. A pair of raw images, I_{up} and I_{do}, are tiled into M patches. Each patch is fed into a detection model to crop individual kernels. Then, two corresponding sides of each kernel in I_{up} and I_{do} are paired based on their centroid positions, and cropped images are adjusted to be vertically oriented. Finally, a recognition model is employed to determine whether each kernel is normal or defective.

image pre-processing, each raw image is tiled into M (empirically set to 15) overlapping patches, with each patch ($\{p^1, \cdots, p^M\}$) of 1500×1500 pixels. Then:

Detection: All tiled patches are fed into a detection model ϕ_d to detect every individual kernel. The detection results are gathered to form two sets **UpDetRes** and **DoDetRes** derived from I_{up} and I_{do} respectively. We fine-tune the advanced YoloX [23] to perform the detection task, due to the superiority in both performance and efficiency.

Pairing&Rotation: As the UP and DOWN images symmetrically depict two sides of kernels, we pair them from **UpDetRes** and **DoDetRes**. Two individual images of the same kernel are using the Hungarian matching algorithm [25] according to their centroid position and spatial distance, generating a set of paired single-kernel images $\{I_s^1, \cdots, I_s^K\}$ for K individual kernels. Each I_s includes both UP and DOWN views.

Following the pairing process, we adjust the orientation of each kernel using a series of image morphological operations, as shown in Fig. 6. This involves basic binarization of cropped images to identify potential kernel contours. Then, the minimum bounding rectangle surrounding these contours can be obtained, after which we can roughly determine the angle θ of vertical orientation for the grain. We perform this rotation for two main reasons: it enhances the visual presentation of each grain, especially facilitating comparisons across different categories. When compared to a non-oriented approach, this method slightly improves classification performance by mitigating the influence of adjacent kernels.

Recognition: We utilize the widely-used ResNet-50 [22] as our backbone ϕ_r to extract features $\mathbf{f}_k \in \mathbb{R}^{e \times hw}$ (where e, h, w are channel, height and width respectively) for the input I_s^k where two views are merged horizontally. Due to the subtle attributes of defective kernels, inspired by fine-grained recognition [26], we employ the trilinear product to explicitly build the inter-channel relationships, which encourages the model to learn discriminative parts. The

trilinear product is computed as:

$$\mathbf{f}_k := (\mathbf{f}_k \mathbf{f}_k^T)\mathbf{f}_k, \tag{1}$$

where T denotes the matrix transpose. The trilinear feature $\mathbf{f}_k$ is fed into the softmax operation to obtain the final prediction outcome. Considering that grain kernels from the same category have varying appearance, we employ the prototypical contrastive loss $\mathcal{L}_{pc}$ [27] to learn a prototype for each category and regularize the similarity among samples from the same category in a training mini-batch:

$$\mathcal{L}_{pc}(\mathbf{f}_k) = -\log \frac{\exp(\mathbf{f}_k \cdot \mathbf{p}_{c_k}/\tau)}{\sum_{j=k,j\neq c_i}^{C} \exp(\mathbf{f}_k \cdot \mathbf{p}_j/\tau)}, \tag{2}$$

where $\mathbf{p}_{c_k}$ is the prototype features for class c_k and C is the number of categories, and τ is the temperature hyper-parameter. All prototype features are stored and initialized as dictionary entries. They are updated by averaging features from the corresponding classes in each training epoch.

Moreover, due to the nature of defective kernels, the distribution of different types of kernels is imbalanced. We employ the class-balanced softmax cross-entropy ($\mathcal{L}_{CB\_CE}$) [28] and focal loss ($\mathcal{L}_{focal}$) [29] together, in which heavy penalties are applied when rare categories are identified incorrectly. Finally, the training objective is:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{pc} + (1-\alpha)(\mathcal{L}_{CB\_CE} + \mathcal{L}_{focal}). \tag{3}$$

where α is used to balance these losses. α linearly decreases from 0.5 to 0.1 as the training epochs increase. Thereby the model is encouraged to learn prototypical features in the early training phase, and then is forced to learn hard examples.

To train the model, we utilize our proposed framework to generate a vast number of candidate single-kernel images. These images are then annotated by four experienced inspectors and experts. Then, we built and released a large-scale dataset, named GrainDet. It includes over $140K$ single-kernel images, consisting of over $80K$, $20K$ and $40K$ kernels for wheat, rice and sorghum respectively, as shown in Table II. It is noted that the black point (BP), unripe (UN) and heated (HD) grains are unique to wheat, rice and sorghum respectively.

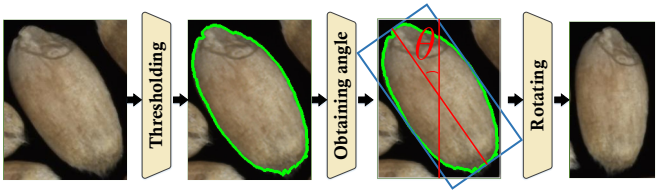

Fig. 6. Illustration of the rotation process.

TABLE II

Distributions of normal, defective and impurities in GrainDet.

Species	NOR	Defective						IM	Total
		SD	FS	- -	MY	BN	AP		
Wheat	$48.1K$	$5.1K$	$5.1K$	$1.4K$(BP)	$5.2K$	$5.8K$	$5.2K$	$4K$	$80K$
Rice	$16K$	$1.1K$	$1.2K$	$1.2K$(UN)	$1.2K$	$1.2K$	$1.2K$	$1.6K$	$24.7K$
Sorghum	$24K$	$3.6K$	$1.4K$	$0.2K$(HD)	$3.6K$	$3.6K$	$3.6K$	$1.2K$	$41K$

III. Experiments and Results

A. Experimental Setup

AV4GAInsp: We utilized an MCU (STM32F103) to regulate our prototype device and employed an embedded computational platform (Nvidia Jetson Xavier NX) to execute our CNN-based analysis framework. All models are trained on a multi-GPU workstation and models are then converted and deployed on the TensorRT framework.

Metrics: We reported the inspection time according to the physical measurement. For recognition, we reported results for accuracy, recall, F1-score and confusion matrix to provide a comprehensive evaluation of our approach. For the AV4GAInsp vs. inspectors experiments, we adopted the widely recognized Kappa statistic [30] to measure the level of agreement. Kappa results span from -1 to 1, with a score between 0.8 and 1 signifying near perfect agreement.

B. Efficiency Analysis for Our Acquisition Device

As shown in Fig. 7, we analyzed four phases of the data acquisition process to evaluate the efficiency of our device.

Feeding: This stage is particularly time-consuming as it involves vibrating kernels from the cabin at the material entrance to the end of the vibration plate. We observed that high vibration speeds lead to over-dispersion of the kernels, thereby reducing the efficiency of the inspection process. On the other hand, slower speeds result in the occlusion of kernels which must be avoided. Therefore, we incorporated a PI controller that uses monitoring signals from a flow camera to adjust the vibration frequency f_v and duration time t_v. This optimization ensures that grain kernels spread evenly across the entire plate. Consequently, the average vibration time for a batch of kernels is optimized to 3.8 seconds (s).

Transporting & Recycling: A slow speed of the transparent plate affects the efficiency while a high speed may cause some kernels to fall off the plate. To ensure the stability of all kernels on the plate, we adopted a trapezoidal trajectory during both transporting and recycling phases. The speed initially linearly ramps up to 80 mm/s, then remains constant and finally decelerates linearly. After optimization, the average time per batch is reduced to 2.3s.

Capturing: Underexposure occurs if the exposure time is overly shortened, resulting in dark images. If both

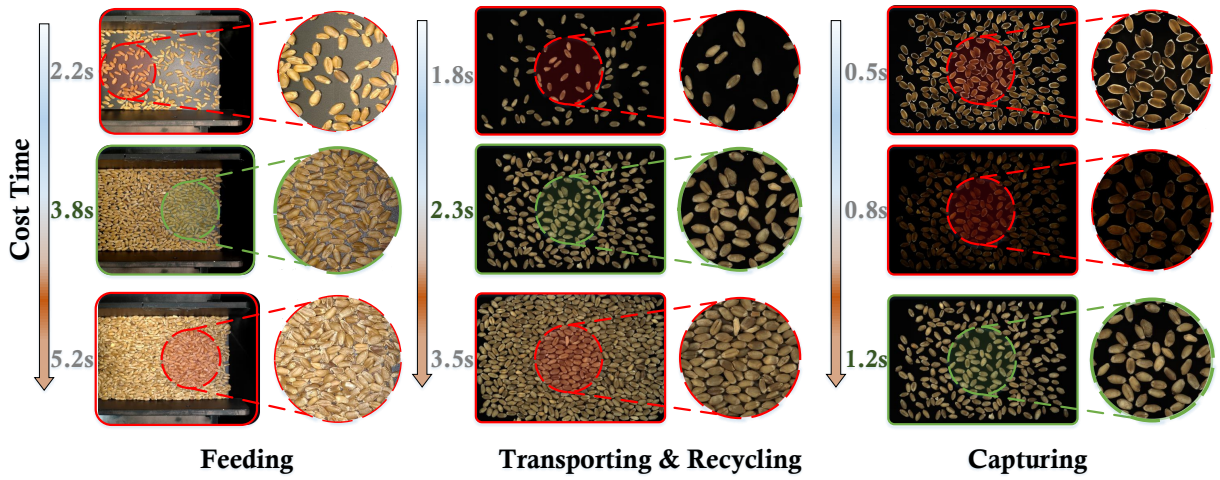

Fig. 7. Efficiency analysis for data acquisition. We optimized the feeding, transporting & recycling, and capturing phases to 3.8s, 2.3s and 1.2s per batch.

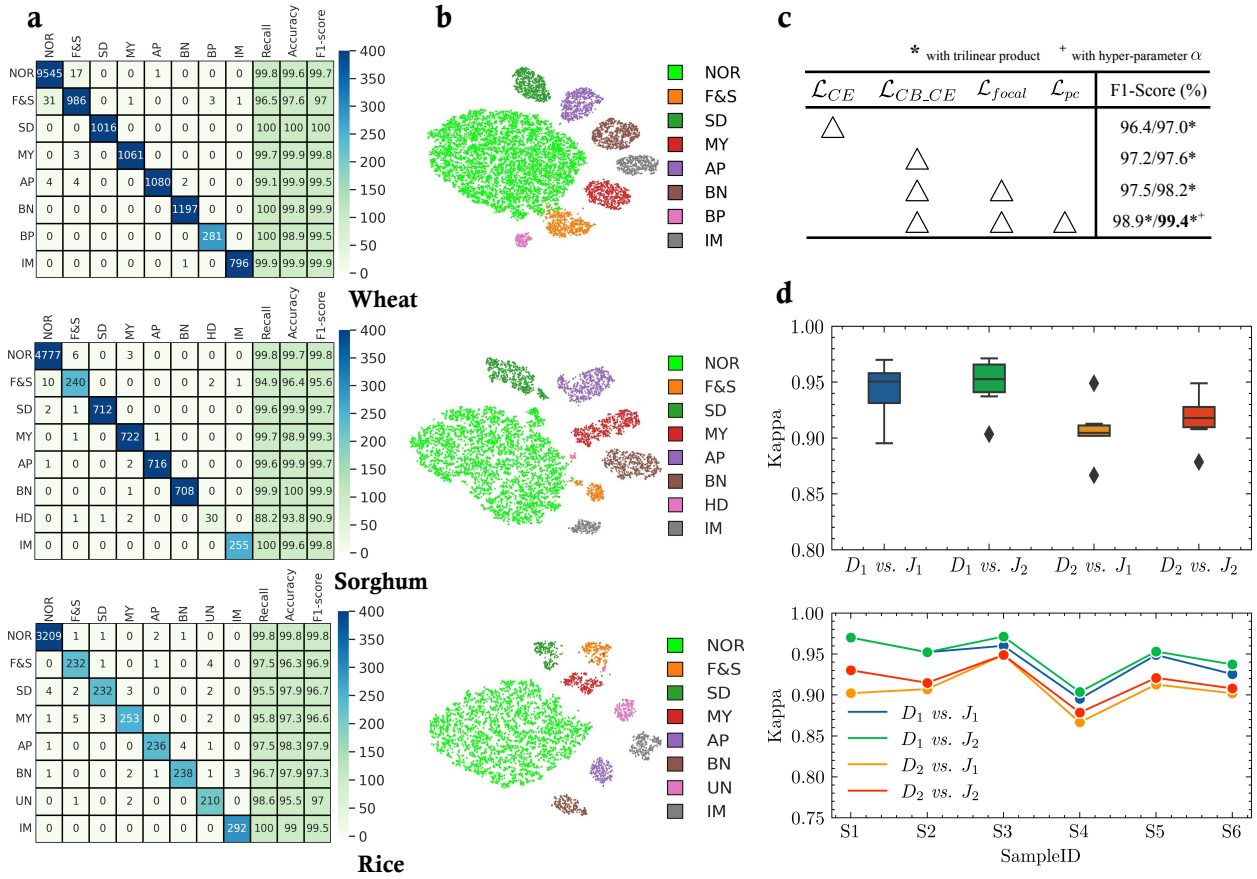

Fig. 8. a) The performance including recall, accuracy and F1-score. b) The t-SNE features visualization on test sets. c) Ablation studies of the training settings on the wheat data. d) The Kappa statistics between AV4GAInsp (D_1, D_2) and inspectors ($J_1, J2$) on six wheat samples (S1,$\cdots$, S6).

camera units are synchronized to capture images simultaneously, the captured images are over-exposed and unsuitable for analysis. Therefore, we chose to activate the UP and DOWN camera units sequentially, which optimizes this phase to a streamlined duration of 1.2s per batch.

Based on the above optimizations, an entire inspection for a batch was reduced to 9.5s. We proceeded to measure the time costs of our system in comparison to inspectors. For three types of grains, we tested samples

of three regular weights, with each weight including 3 individual samples. The results are presented as average and standard deviations in Table III, For example, our system completes the inspection of a 60g wheat sample in a mere 114s (i.e., approximately 2 minutes) while human inspectors require approximately 2118s (i.e., approximately 35 minutes), demonstrating an over 18-fold efficiency improvement. In particular, when handling sorghum with a smaller physical size, our system shows remarkable efficiency with over 20× speedup, completing a 60g sorghum analysis in around 204s compared to the inspectors' 4220s. The high efficiency and stability of our device provide a huge potential for application during harvesting seasons.

TABLE III
Comparison in inspection time cost for three types of grain.

Weights	Cereal Grains (AI4GAInsp / Inspectors)		
	Wheat	Sorghum	Rice
50g	94±7s / 1592±124s	178±8s / 3520±143s	106±9s / 1787±128s
60g	114±8s / 2118±182s	204±15s / 4220±161s	137±7s / 2224±189s
120g	219±9s / 3920±211s	431±20s / 9512±223s	289±11s / 5150±291s

C. Recognition

To verify the efficacy of our AV4GAInsp system, we trained our models on our released GrainDet. The dataset is divided into training, validation and test subsets by randomly sampling 70%, 10% and 20% proportions respectively. We trained the models with standard training parameters. For example, all input images are resized into 192×192 pixels. Following the training settings [22], the training epoch was set to 90, the learning schedule employed the step strategy, and the initial learning rate was set to 0.0001. We followed the original papers [27], [28], [29] and set the corresponding hyper-parameters for each loss, e.g., τ to 0.1. The initial weights of our model are determined using models pre-trained on ImageNet [22]. We reported the recognition results on the test sets for wheat, sorghum, and rice.

As shown in Fig. 8a, we observed that our system demonstrates promising performance across all three types of cereal grains. For wheat, our system achieved an average recall of 99.4%, an accuracy of 99.5% and an F1-score of 99.4%. It also attained competitive average accuracy, recall and F1-scores of over 98.5%, 97.7% and 98.1% for sorghum respectively, and over 97.7%, 97.6% and 97.7% for rice respectively. These results demonstrated the efficacy and superiority of our system. Moreover, to provide a qualitative analysis, we utilized the t-SNE technique [31] to visualize the features extracted by our trained recognition models. As illustrated in Fig. 8b, we observed that samples of different categories are distinctly separate from each other in the feature spaces, which verifies that our model effectively learns clear classification boundaries in classifying normal or defective kernels.

We also conducted ablation studies on the wheat data to verify the training settings (see Fig. 8c). Compared to the cross-entropy loss $\mathcal{L}_{CE}$, the model with only $\mathcal{L}_{CB\_CE}$ obtains an F1-score of 97.2%. By incorporating the focal loss $\mathcal{L}_{focal}$, the model produces a slight improvement of 0.3%. We observed that consistent improvements can be obtained by introducing the trilinear product. When introducing the prototypical contrastive loss $\mathcal{L}_{pc}$ without α, our model achieves improvements of 0.7%. After introducing α to balance losses, the model produces the best performance of 99.4%. These experiments demonstrate the effectiveness of our training designs.

D. Comparisons with Existing Methods

We extended to compare our three-stage framework with a variety of deep learning methods. Specifically, for object detection methods, we considered Faster R-CNN [32] and YoloX [23]; for instance segmentation methods, we included Mask R-CNN [33] and RTMDet [34], and we additionally tested the popular SAM model [35].

In order to fairly compare our framework with existing methods, we selected more than 1K, 0.8K, and 0.5K raw high-resolution images for wheat, rice and sorghum, and these images are divided into 70%, 10% and 20% for training, validation and testing. A subset of raw images is released publicly for validation. For training, considering that existing methods are designed for 3-channel images, we concatenated a pair of UP and DOWN images (I_{up} and I_{do}) along the channel dimension to obtain a 6-channel image as the input. The training objectives are to predict the bounding boxes of two sides and defective categories. For the evaluation, we followed the previous studies [32], [33] and adopted AP_{50} as the metric. AP_{50} refers to Average Precision at a 50% Intersection over Union (IoU), meaning the IoU between the predicted box and the ground truth box (or mask) must exceed 50%.

As shown in Table IV, we can observe that our framework outperforms all existing methods, achieving AP_{50} of 58.7%, 41.2% and 33.6% for wheat, rice and sorghum respectively. We consider that the existing methods are typically designed for natural images, where the target objects and their relationships are highly relevant and complex both spatially and semantically. These models usually employ a shared backbone to concurrently perform both detection and classification tasks, thereby enhancing overall performance. In contrast, the images captured by our device have a simplistic background and spatial relationships. Our framework decouples detection and classification into separate tasks and employs different backbones. Such strategy enables more efficient training for individual tasks. In addition, our framework has lower parameters and computational complexity, and it requires no pixel-level annotations during the training process.

TABLE IV

Comparisons between our framework and existing methods.

Methods		Params (M ↓)	Flops (G ↓)	Wheat (↑)	Rice (↑)	Sorghum (↑)
Object Detection	Faster R-CNN [32]	41.4	63.8	38.1%	20.8%	25.8%
	YoloX [23]	25.3	37.3	27.7%	21.3%	35.5%
Instance Segmentation	Mask R-CNN [33]	44.0	115.1	32.5%	23.0%	29.7%
	RTMDet [34]	27.5	34.7	18.8%	14.1%	24.2%
SAM [35]		91.1	219.6	4.9%	3.9%	3.2%
Our		34.9	18.7	58.7%	41.2%	33.6%

E. AV4GAInsp vs. Human Experts

The goal of developing the AV4GAInsp system is to support inspectors by, not only enhancing efficiency, but also providing consistent inspection results comparable to human experts. In practice, inspectors are required to assess a specific weight of kernels [6], [36]. To compare our system's performance against human expertise, we conducted experiments involving AV4GAInsp vs. human experts. For this, we constructed two individual prototype devices, named D_1 & D_2, and employed two experts (named J_1 & J_2) possessing over 5 years of inspection experience and working in quality inspection centers. We prepared six wheat samples (each of 60g) containing various proportions of defective kernels ranging from about 1% to 20%. These samples were shuffled and tested blindly during the testing phase.

As shown in Fig. 8d, the box chart illustrates that both devices D_1 and D_2 show high consistency with Kappa scores exceeding 0.85 across all testing samples. Our AV4GAInsp system displays strong consistency with human experts, achieving average Kappa statistics of 94.19, 94.79, 90.65 and 91.68 compared to human experts J_1 and J_2, respectively. Furthermore, in the line chart, our AV4GAInsp system aligns with J_1 and J_2 across all samples from S1 to S6 with increase of the proportions of defective kernels.

Compared to human experts, our AV4GAInsp system possesses overwhelming advantages in inspection time. On the other hand, inspectors require extensive training and working experience but our AV4GAInsp is only trained using $80K$ images yet shows remarkable performance and efficiency. It can be easily adapted and extended to various types of cereal grains, showcasing its versatility and scalability.

IV. conclusions and future work

In this paper, we presented an automated vision-based system AV4GAInsp for GAI tasks. Our system consists of a prototype device and a deep learning-based analysis framework. The device is designed to capture high-quality images of grain kernels from different views efficiently, providing rich visual information for inspection. The analysis framework is developed to analyze the captured images to perform the GAI tasks. We conducted comprehensive experiments across three types of cereal grains to demonstrate the feasibility, superiority and consistency of our AV4GAInsp system compared to human experts.

Limitations and future work: Our system exhibits proficiency in processing grains of a spherical shape. However, when it comes to polyhedral-shaped grains like maize, capturing comprehensive appearance information presents a challenge. This can potentially be addressed by incorporating more camera units. In the analysis framework, we primarily focus on recognition tasks where each grain kernel can be classified into a pre-defined category. However, in real-world applications, numerous kernels cannot be identified into known classes. To tackle this, the anomaly detection task may prove beneficial. While AV4GAInsp is highly specialized, it also holds the potential for processing more types of cereal grains and other tiny objects.

Beyond the recognition task, the system could be utilized for other tasks such as counting, detection and segmentation. We believe that our system and its fundamental design can serve as a prototype for other projects aimed at the automated analysis of a large number of tiny objects. We hope that our work will stimulate interest and inspire further research in the field of GAI and smart agriculture.

References

[1] J. D. Sachs, "From millennium development goals to sustainable development goals," The Lancet, vol. 379, no. 9832, pp. 2206–2211, 2012.

[2] "ISO 5527: Cereals – Vocabulary," International Organization for Standardization, Standard, Feb. 2015.

[3] Y.-N. Wan, C.-M. Lin, and J.-F. Chiou, "Rice quality classification using an automatic grain quality inspection system," Transactions of the ASAE, vol. 45, no. 2, p. 379, 2002.

[4] T. Brosnan and D.-W. Sun, "Improving quality inspection of food products by computer vision—-a review," Journal of food engineering, vol. 61, no. 1, pp. 3–16, 2004.

[5] P. R. Johnson, T. Grennes, and M. Thursby, "Devaluation, foreign trade controls, and domestic wheat prices," American Journal of Agricultural Economics, vol. 59, no. 4, pp. 619–627, 1977.

[6] "ISO 24333: Cereals and cereal products — Sampling," International Organization for Standardization, Standard, Dec. 2009.

[7] H. Phalen, P. Vagdargi, M. L. Schrum, S. Chakravarty et al., "A mosquito pick-and-place system for PFSPZ-based malaria vaccine production," IEEE Transactions on Automation Science and Engineering, vol. 18, no. 1, pp. 299–310, 2020.

[8] L. Fan, A. Sowmya, E. Meijering, and Y. Song, "Cancer survival prediction from whole slide images with self-supervised learning and slide consistency," IEEE Transactions on Medical Imaging, vol. 42, no. 5, pp. 1401–1412, 2023.

[9] N. Häni, P. Roy, and V. Isler, "Minneapple: A benchmark dataset for apple detection and segmentation," IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 852–858, 2020.

[10] Y. Xiong, Y. Ge, and P. J. From, "Push and drag: An active obstacle separation method for fruit harvesting robots," in ICRA, 2020, pp. 4957–4962.

[11] M. Bakken, V. R. Ponnambalam, R. J. Moore, J. G. O. Gjevestad, and P. J. From, "Robot-supervised learning of crop row segmentation," in ICRA, 2021, pp. 2185–2191.

[12] J. Weyler, A. Milioto, T. Falck, J. Behley, and C. Stachniss, "Joint plant instance detection and leaf count estimation for in-field plant phenotyping," IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 3599–3606, 2021.

[13] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, "Deep-plant: Plant identification with convolutional neural networks," in ICIP. IEEE, 2015, pp. 452–456.

[14] S. Rasti, C. J. Bleakley, N. Holden, R. Whetton, D. Langton, and G. O'Hare, "A survey of high resolution image processing techniques for cereal crop growth monitoring," Information Processing in Agriculture, vol. 9, no. 2, pp. 300–315, 2022.

[15] V. Kakani, V. H. Nguyen, B. P. Kumar, H. Kim, and V. R. Pasupuleti, "A critical review on computer vision and artificial intelligence in food industry," Journal of Agriculture and Food Research, vol. 2, p. 100033, 2020.

[16] H. Chen, A. Chen, L. Xu, H. Xie, H. Qiao, Q. Lin, and K. Cai, "A deep learning cnn architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources," Agricultural Water Management, vol. 240, p. 106303, 2020.

[17] L. Fan, Y. Ding, D. Fan, D. Di, M. Pagnucco, and Y. Song, "Grainspace: A large-scale dataset for fine-grained and domain-adaptive recognition of cereal grains," in CVPR, 2022, pp. 21 116–21 125.

[18] X. Yang, L. Shu, J. Chen, M. A. Ferrag, J. Wu, E. Nurellari, and K. Huang, "A survey on smart agriculture: Development modes, technologies, and security and privacy challenges," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 2, pp. 273–302, 2021.

[19] L. Fan et al., "Identifying the defective: Detecting damaged grains for cereal appearance inspection," in ECAI, 2023, pp. "660–667".

[20] ——, "An annotated grain kernel image database for visual quality inspection," Scientific Data, 2023.

[21] L. Fan, A. Sowmya, E. Meijering, and Y. Song, "Fast FF-to-FFPE Whole Slide Image Translation via Laplacian Pyramid and Contrastive Learning," in MICCAI, 2022, pp. 409–419.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[24] L. Fan, A. Sowmya, E. Meijering, and Y. Song, "Learning visual features by colorization for slide-consistent survival prediction from whole slide images," in MICCAI, 2021, pp. 592–601.

[25] H. W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.

[26] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in CVPR, 2019, pp. 5012–5021.

[27] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in CVPR, 2021, pp. 943–952.

[28] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in CVPR, 2019, pp. 9268–9277.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in ICCV, 2017, pp. 2980–2988.

[30] M. L. McHugh, "Interrater reliability: the kappa statistic," Biochemia medica, vol. 22, no. 3, pp. 276–282, 2012.

[31] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," JMLR, vol. 9, no. 86, pp. 2579–2605, 2008.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE T-PAMI, vol. 39, no. 06, pp. 1137–1149, 2017.

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in ICCV, 2017, pp. 2961–2969.

[34] C. Lyu et al., "Rtmdet: An empirical study of designing real-time object detectors," arXiv preprint arXiv:2212.07784, 2022.

[35] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," arXiv preprint arXiv:2304.02643, 2023.

[36] "ISO 7970: Wheat," International Organization for Standardization, Standard, Jan. 2021.