Creating Image Datasets in Agricultural Environments using DALL.E: Generative AI-Powered Large Language Model

Ranjan Sapkota, and **Manoj Karkee**

Center for Precision Automated Agricultural Systems, Washington State University, 24106 N. Bunn Rd, Prosser, 99350, Washington, USA

ABSTRACT

This research investigated the role of artificial intelligence (AI), specifically the DALL.E model by OpenAI, in advancing data generation and visualization techniques in agriculture. DALL.E, an advanced AI image generator, works alongside ChatGPT's language processing to transform text descriptions and image clues into realistic visual representations of the content. The study used both approaches of image generation: text-to-image and image-to-image (variation). Two types of datasets depicting fruit crop environment and "crop-vs-weed" environment were generated. These AI-generated images were then compared against ground truth images captured by sensors in real agricultural fields. The comparison was based on Peak Signal-to-Noise Ratio (PSNR) and Feature Similarity Index (FSIM) metrics. For fruit crops, image-to-image generation exhibited a 5.78% increase in average PSNR over text-to-image methods, signifying superior image clarity and quality. However, this method also resulted in a 10.23% decrease in average FSIM, indicating a diminished structural and textural similarity to the original images. Conversely, in crop vs weed scenarios, image-to-image generation showed a 3.77% increase in PSNR, demonstrating enhanced image precision, but experienced a slight 0.76% decrease in FSIM, suggesting a minor reduction in feature similarity. Similar to these measures, human evaluation also showed that images generated using image-to-image-based method were more realistic compared to those generated

https://doi.org/10.32388/V2G1MP.2

with text-to-image approach. The results highlighted DALL.E's potential in generating realistic agricultural image datasets and thus accelerating the development and adoption of precision agricultural solutions.

1 Introduction

In recent years, synthetic images, generated by computer algorithms to resemble real-world entities, have been widely used in various sectors, such as healthcare[1], biomedicine[2], fashion[3], architecture[4], geospatial studies[5], automotive industry[6], and agriculture[7] due to their ability to provide realistic visual representations for observation, and analysis, and driving innovations[8].

Traditional methods of creating these images include parametric techniques such as [9]Bezier Curves used by Chen et al. [9] to develop more accurate synthetic images of cell nuclei for biomedical applications. Similarly, Alberto et al. [10] used Bezier Curves for designing complex mechanical structures using synthetic images. Another classical method is ray tracing, a technique to render realistic images by simulating light paths, improved upon by Ben et al. [11] through Neural Radiance Fields for better low-light image reconstruction. Additionally, Physics-based Rendering (PBR) [12] , a method that mimics real-world light flow, has been used effectively for creating photorealistic images, as demonstrated by Hodan et al. [12] in their AI-based object detection research. These traditional image generation methods highlighted the evolving role of synthetic images in driving technological advancements [13]. However, these traditional models of synthetic image generation come with notable limitations. Parametric models, for instance, rely heavily on the accuracy of parameters and equations, making them less adaptable to complex or irregular shapes[14] and environments that don't align with predefined mathematical structures [15]. The ray tracing technique faces challenges due to high computational intensity and time[16]. This technique also can be limiting in simulating complex lighting effects like indirect light and reflections [17]. PBR, on the other hand, has reduced flexibility and high computational demands[18].

Generative Adversarial Networks (GANs) present a promising alternative to generate synthetic images. These networks, consisting of a generator and a discriminator, efficiently produce realistic synthetic images[19]. GANs provide greater flexibility than parametric and other traditional models by learning from high-dimensional data distributions, enabling them to generate more realistic images, even in complex scenarios [22]. This approach also addresses the computational challenges of ray tracing, as trained GANs can quickly generate new images [17]. Furthermore, GANs balance the realism-flexibility trade-off better than the PBR method, allowing detailed image generations without sacrificing quality[20]. Their ability to create realistic images in complex environments enable them to be adoptable to wider fields and applications[21], making them a crucial tool in synthetic image generation.

In recent years, the application of Generative Adversarial Networks (GANs) in agriculture has gained increasing attention, particularly for tasks like disease detection and image augmentation, yielding promising results. For instance, Abbas et al.[22] demonstrated the effectiveness of GANs, specifically using a Conditional GAN (C-GAN), to generate synthetic images of tomato plant leaves. This technique, combined with a DenseNet121 model and transfer learning, achieved a high accuracy of 99.5% in classifying tomato leaf diseases. It's noteworthy that their approach integrated both synthetic and actual images to enhance classification accuracy, suggesting a blend of novel and traditional methodologies. Furthermore, Lu et al. [23] utilized GANs to create synthetic images of insect pests, thereby augmenting limited actual datasets (collected with sensors). This innovation significantly improved the performance of classifiers for insect pests, highlighting the utility of GANs in scenarios where actual data is scarce. Nazki et al. [24] explored a different aspect of GANs by employing them for image-to-image translation in plant disease datasets, which facilitated more accurate disease classification. These studies indicate a trend towards using GANs not just for dataset augmentation - a role typically filled by conventional techniques like rotation and flipping - but also as a crucial tool in synthesizing and enhancing the quality of agricultural datasets. This shift marks a significant advancement in the application of AI in agriculture, opening new pathways for research and practical applications in the field.

Studies[22,26] have shown that GANs effectively address the challenges of biological variability and the complexity of unstructured agricultural environments by successfully identifying and classifying pest and plant leaf diseases. GANs have been instrumental in several key areas of agricultural image processing. GANs enhance model efficiency by reducing the need for extensive data collection and labeling, particularly in diverse crop scenarios[27]. For instance, Gomaa et al. [28] utilized a combination of Convolutional Neural Networks (CNN) and GANs for disease detection in tomato plants, highlighting the synergistic potential of combining traditional and generative models. Similarly, Madsen et al.[29] applied Wasserstein auxiliary classifier generative adversarial networks (Wac-GAN) to model seedlings of nine different plants, showcasing the versatility of GANs in handling varied crop types. Zhu et al. [30] took a specialized approach with Conditional Deep Convolutional GANs (C-DCGAN) for orchid seedling vigor rating, emphasizing the precision capabilities of GANs. Further, studies like Hartley et al. [31] with wheat for plant head detection using CycleGAN [31], and Bird et al. [32] focusing on lemon quality assessment using C-GAN illustrate the broad applicability of these networks across different crop environment. Table 1 summarizes these recent efforts, showcasing how the integration of GANs in synthetic image generation is revolutionizing agricultural applications and contributing to the advancement of machine vision systems in agriculture.

Table 1: Overview of GAN-based Synthetic Image Generation in Agriculture (2019-2024), Highlighting Image Generation Techniques, Crops, and Key Achievements.

Author Reference	Target Crop	Synthetic Image Generation Technique	Primary Objective
Abbas et. al [22]	Tomato plants	Conditional Generative Adversarial Network (C-GAN)	Disease detection
Gomaa et. al [28]	Tomato plants	Convolutional Neural Network (CNN) and GAN	Disease detection
Madsen et. al[29]	Nine different plant seedlings as: 1. Charlock 2. Cleavers 3. Common Chickweed 4. Fat Hen 5. Maize 6. Scentless Mayweed 7. Shepherd's Purse 8. Small-flowered Cranesbill 9. Sugar Beets	Wasserstein auxiliary classifier generative adversarial network (Wac-GAN)	Modeling plant seedlings
Zhu et. al [30]	Orchid seedlings	Conditional deep convolutional generative adversarial network (C-DCGAN)	Plant Vigor rating
Hartley et. al [31]	Wheat	CycleGAN	Plant head detection

Bird et. al[32]	Lemons	C-GAN	Fruit quality assessment and defect classification
Shete et. al [33]	Maize plants	TasselGAN and deep convolutional generative adversarial networks (DCGAN)	Image generation of maize tassels against sky backgrounds
Guo et. al[34]	Jujubes	DCGAN	Quality grading
Drees et. al [35]	Arabidopsis thaliana and cauliflower plants	C-GAN (Pix2Pix)	Laboratory-grown and field-grown image generation
Kierdorf et. al [36]	Grapevine Berries	C-GAN/CDCGAN	Estimation of occluded fruits
Olatunji et. al [37]	Kiwifruit	C-GAN	Filling in missing fruit surface (Reconstruction)
Bellocchio et. al [38]	Apple orchard	CycleGAN	Unseen fruits counting
Fawakherji et. al [39]	Sugar beet, sunflower	CGAN/CDCGAN	Crop/weed segmentation in precision farming
Zeng et. al[40]	Citrus	DCGAN	Disease severity detection
Kim et. al[41]	Blueberry leaves	DCGAN	Fruit tree disease classification
Tian et. al [42]	Apple canopy	CycleGAN	Disease detection
Cap et. al [43]	Cucumber leaves	CycleGAN	Plant disease diagnosis
Maqsood et. al [44]	Wheat	super-resolution generative adversarial networks (SR-GAN)	Wheat stripe(yellow) rust classification
Bi et. al[45]	Grape, Orange, Potato,	Wasserstein generative	Plant disease classification

	Squash, Tomato	adversarial network with gradient penalty (WGAN-GP)	
Zhao et. al[46]	Apple, Corn, Grape, Potato, Tomato	DoubleGAN	Plant disease detection
Nerkar et. al [47]	Apple, corn, tomato, potato	Reinforced GAN	Leaf disease detection

The recent advancement of Natural Language Processing (NLP) models and Large Language Models (LLMs) has led to capability for handling the complexities of language understanding and generation. Initially, models like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) were foundational in NLP; however, they struggled with capturing long-range dependencies crucial for tasks such as translation and summarization (Hochreiter and Schmidhuber 1997; Radford et al. 2019a). This limitation was substantially mitigated in 2017 with the introduction of the Transformer model by Vaswani et al. (Vaswani et al. 2017). The model's self-attention mechanism allowed it to effectively attend to all tokens in the input sequence, thereby capturing long-range dependencies and significantly enhancing performance across various NLP tasks (Devlin et al. 2018). The historical growth of LLMs starting from development of transformer in 2017 by Vaswani et al. to current state of the art LLMs are presented in Figure 1.

Building on the success of the Transformer, (Devlin et al. 2018) introduced BERT (Bidirectional Encoder Representations from Transformers) in 2018, which utilized deep bidirectional training by conditioning on both left and right contexts simultaneously. This methodology enabled the model to be fine-tuned with just one additional output layer to perform a wide range of NLP tasks effectively, setting a new standard in the field. The development of language models continued to accelerate with the introduction of GPT-2 by Radford et al. in 2019, which employed a transformer-based architecture with 1.5 billion parameters, leveraging self-attention mechanisms to enhance text generation capabilities (Radford et al. 2019b). That same year, the Megatron-LM was developed by Shoeybi et al. (Shoeybi et al. 2019), featuring 8.3 billion parameters which allowed for even more complex pattern recognition and faster training due to its innovative parallelization scheme.

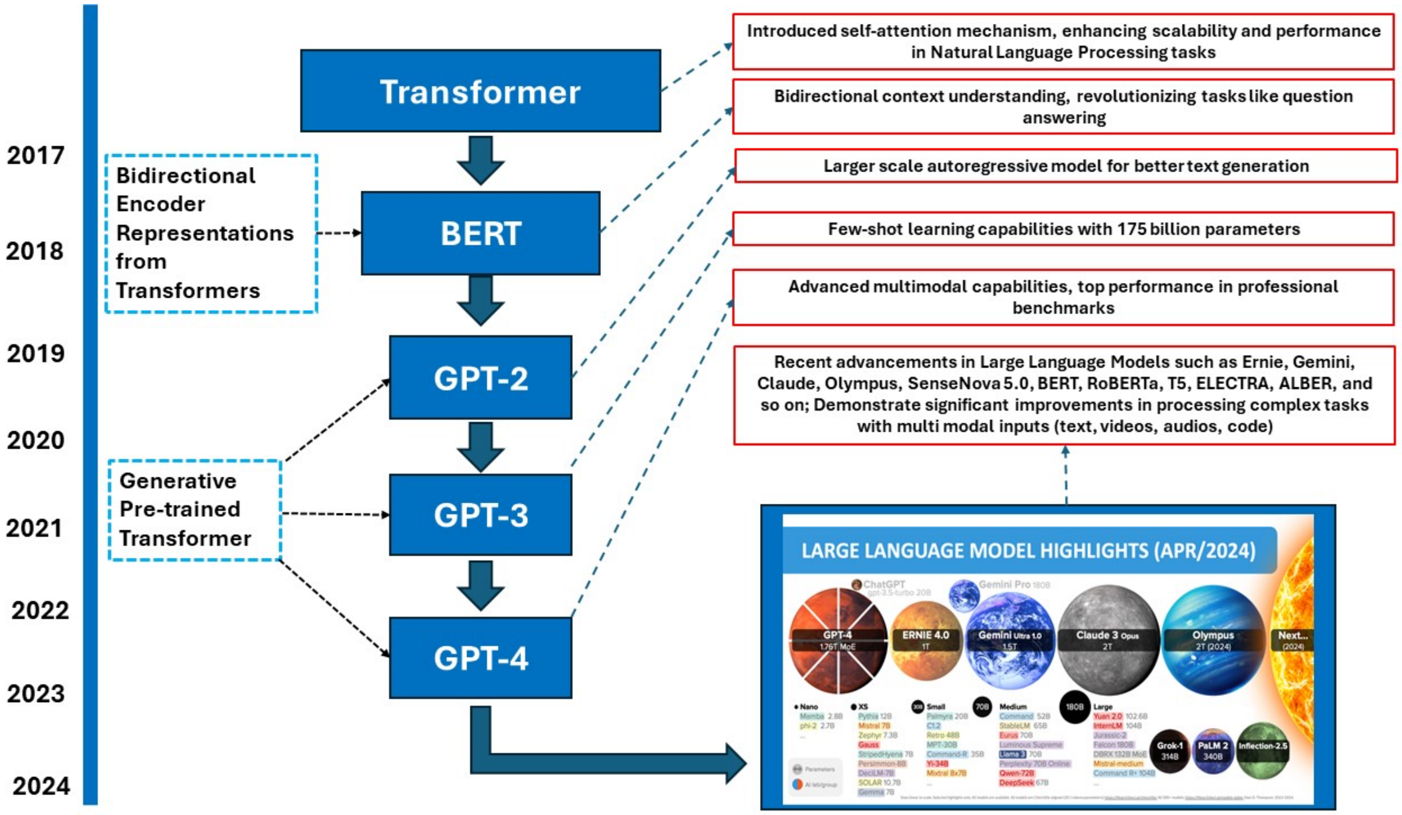
2017
2018
2019
2020
2021
2022
2023
2024
Transformer
BERT
GPT-2
GPT-3
GPT-4
Bidirectional Encoder Representations from Transformers
Generative Pre-trained Transformer
Introduced self-attention mechanism, enhancing scalability and performance in Natural Language Processing tasks
Bidirectional context understanding, revolutionizing tasks like question answering
Larger scale autoregressive model for better text generation
Few-shot learning capabilities with 175 billion parameters
Advanced multimodal capabilities, top performance in professional benchmarks
Recent advancements in Large Language Models such as Ernie, Gemini, Claude, Olympus, SenseNova 5.0, BERT, RoBERTa, T5, ELECTRA, ALBER, and so on; Demonstrate significant improvements in processing complex tasks with multi modal inputs (text, videos, audios, code)
LARGE LANGUAGE MODEL HIGHLIGHTS (APR/2024)
ChatGPT gpt-3.5-turbo 20B
Gemini Pro 180B
GPT-4 1.76T MoE
ERNIE 4.0 1T
Gemini Ultra 1.0 1.5T
Claude 3 Opus 2T
Olympus 2T (2024)
Next... (2024)
Nano
Mamba 2.8B
phi-2 2.7B
XS
Pythia 12B
Mistral 7B
Zephyr 7.3B
Gauss
StripedHyena 7B
Persimmon-8B
DeciLM-7B
SOLAR 10.7B
Gemma 7B
30B
Small
Palmyra 20B
C1.2
Retro 48B
MPT-30B
Command-R 35B
Yi-34B
Mixtral 8x7B
70B
Medium
Command 52B
StableLM 65B
Eurus 70B
Luminous Supreme
Llama 3 70B
Perplexity 70B Online
Qwen-72B
DeepSeek 67B
180B
Large
Yuan 2.0 102.6B
InternLM 104B
Jurassic-2
Falcon 180B
DBRX 132B MoE
Mistral-medium
Command R+ 104B
Grok-1 314B
PaLM 2 340B
Inflection-2.5
Parameters
AI lab/group

Figure 1: Showing the history of Large Language Models starting from 2017 (Animation from Dr Alan D. Thompson, LifeArchitect.ai (June/2024))

In 2020, the release of GPT-3 by Brown et al. demonstrated wider scale and capabilities of Large Language Models (LLMs) with 175 billion parameters, enabling high-quality text generation with minimal fine-tuning (Brown et al. 2020). This model provided the foundation for the next generation of multimodal NLP models. Most recently, in 2023, OpenAI introduced GPT-4, which can process both text and image inputs, demonstrating near-human or superior performance on various professional and academic benchmarks (OpenAI, 2023). Concurrently, Meta released the open-source LLM Llama, which, while smaller in size, offers a valuable resource for researchers worldwide (Meta, 2023).

DALL·E model by OpenAI (OpenAI, California, USA) represents a significant leap forward in the domain of AI-based image generation. Integrating the principles of GANs with innovative technologies such as Compact Language-Image Pretrained (CLIP) embeddings [47], and Principal Component Analysis (PCA) for dimensionality reduction [48], DALL·E transcends the capabilities of traditional image generation methods [49]. One of the major breakthroughs with the DALL·E model is that it can convert textual descriptions into realistic images. Additionally, the model can generate a variation within an image representing similar environments. This capability of the model is achieved using text-conditional hierarchical image generation strategy [48]. Building on the foundation of ChatGPT developed by the same organization (OpenAI, California, USA), this model has been trained on an extensive variety and size of image-text pairs. Both models, stemming from the same OpenAI lineage, manifest exceptional competence in managing intricate, multi-dimensional tasks [50]. For instance, while ChatGPT excels at generating contextually relevant textual responses, DALL·E emerges as a powerhouse in producing images that accurately represent the semantics of the input text[51]. Even though synthetic image generation has become easier and more accessible while providing more realistic images with OpenAI's DALL.E model, there is a need to thoroughly assess and evaluate its capability in representing field environments and its practicality in agricultural applications. To address this need, the following specific objectives were pursued in this study:

Building upon the foundational advancements introduced by Generative Adversarial Networks (GANs) in agricultural image processing, the DALL·E model by OpenAI (OpenAI, California, USA) represents a significant leap forward in the domain of AI-based image generation. Integrating the principles of GANs with innovative technologies such as Compact Language-Image Pretrained (CLIP) embeddings [48], and Principal Component Analysis (PCA) for dimensionality reduction [49], DALL·E transcends the capabilities of traditional image generation methods [50]. One of the major breakthroughs with the DALL·E model is that it can convert textual descriptions into realistic images. Additionally, the model can generate a variation within an image representing similar environments. This capability of the model is achieved using text-conditional hierarchical image generation strategy [49]. Building on the foundation of ChatGPT developed by the same organization (OpenAI, California, USA), this model has been trained on an extensive variety and size of image-text pairs. Both models, stemming from the same OpenAI lineage, manifest exceptional competence in managing intricate, multi-dimensional tasks [51]. For instance, while ChatGPT excels at generating contextually relevant textual responses, DALL·E emerges as a powerhouse in producing images that accurately represent the semantics of the input text[52]. Even though synthetic image generation has become easier and more accessible while providing more realistic images with OpenAI's DALL.E model, there is a need to thoroughly assess and evaluate its capability in representing field environments and its practicality in agricultural applications. To address this need, the following specific objectives were pursued in this study:

- To assess and evaluate the DALL·E model's proficiency in translating detailed textual prompts into accurate and realistic visual representations using text-to-image generation feature of the model.

- To evaluate DALL·E model's ability to accurately transform an image prompt into generating realistic images of the similar environment using image variation feature of the model.

2. Methods

2.1 Data Collection and Compilation

In this study, the focus of image analysis was on two distinct agricultural datasets, as shown in Figure 2. Dataset 1 encompassed a variety of fruit crops, including strawberries, mangoes, apples, avocados, rockmelons, and oranges. These fruits were carefully selected for their distinctive morphological, textural, and color characteristics, as well as their diverse backgrounds. Dataset 2 focused on early-stage crop fields intertwined with weeds, specifically targeting carrot, onion, and corn fields, chosen for their significant relevance in weed management studies. The intention was to assess the DALL·E model's accuracy in depicting agricultural scenarios where the differentiation between crops and weeds is crucial. The original ground truth images for both datasets were obtained from "A Survey of Public Datasets for Computer Vision Tasks in Precision Agriculture" by Lu and Young [53]. Six representative images from the fruit crops dataset and three from the crop versus weed scenarios were randomly selected.

Following this, in the initial image generation step, input text prompts were crafted by carefully examining the original images as depicted in Figure 3. These prompts were then used to generate the first category of images. For the second approach, we directly used the ground truth images as input to the DALL.E model to create variations.

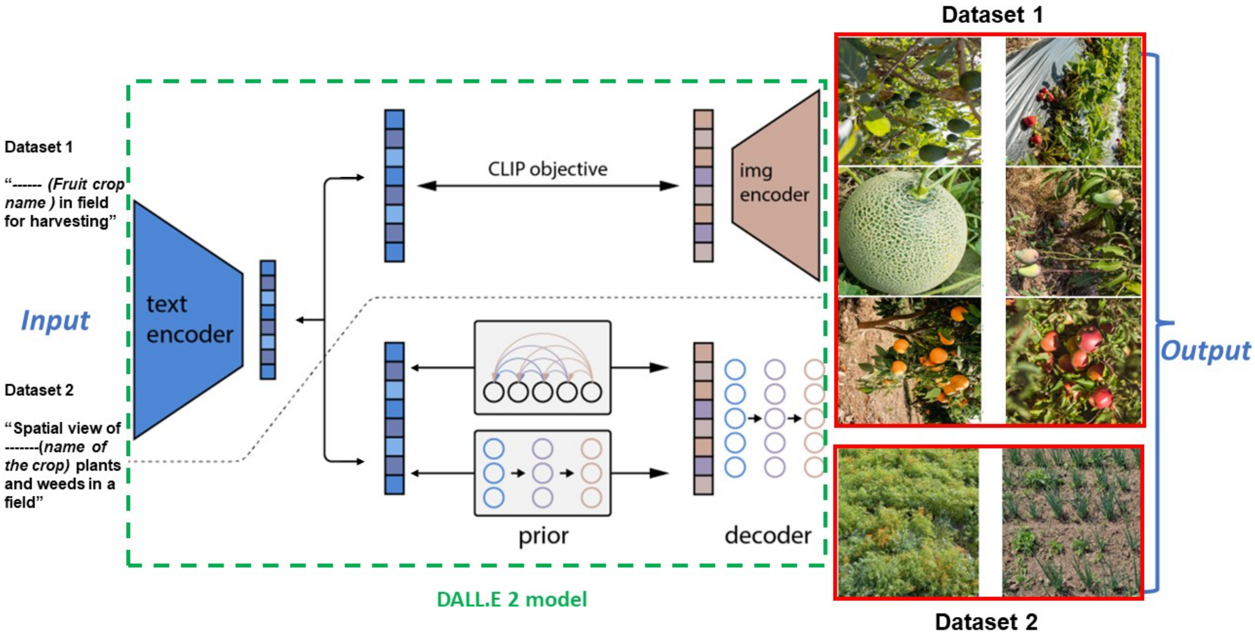
Dataset 1
"------ (Fruit crop name) in field for harvesting"
Input
Dataset 2
"Spatial view of -------(name of the crop) plants and weeds in a field"
text encoder
CLIP objective
img encoder
prior
decoder
DALL.E 2 model
Dataset 1
Output
Dataset 2

Figure 2: Utilizing DALL.E for dataset creation in this research, two distinct sets were employed: Dataset 1 focusing on fruit crops and Dataset 2 on crop vs weed scenarios. Figure 2 demonstrates how textual inputs for text-to-image and image-to-image generation.

After a processing step, images were generated for both categories. The generated images from both approaches were compared against their respective ground truth images. We evaluated the resulting visuals using key metrics: Peak Signal-to-Noise Ratio (PSNR) for image clarity and pixel accuracy, and Feature Similarity Index (FSIM) for structural similarity. Additionally, human assessments were conducted to confirm their realism.

2.2 DALL.E Image Generation Model

In this study, DALL·E 2 (OpenAI, California, USA) image generation model was used, which utilizes hierarchical text-conditional image generation to produce images based on textual descriptions [54]. The hierarchical text-conditional image generation involves a (contrastive model) CLIP image embedding from a given text caption, taking advantage of CLIP's ability to learn robust image representations that encompass both the subject matter and stylistic elements [55]. The

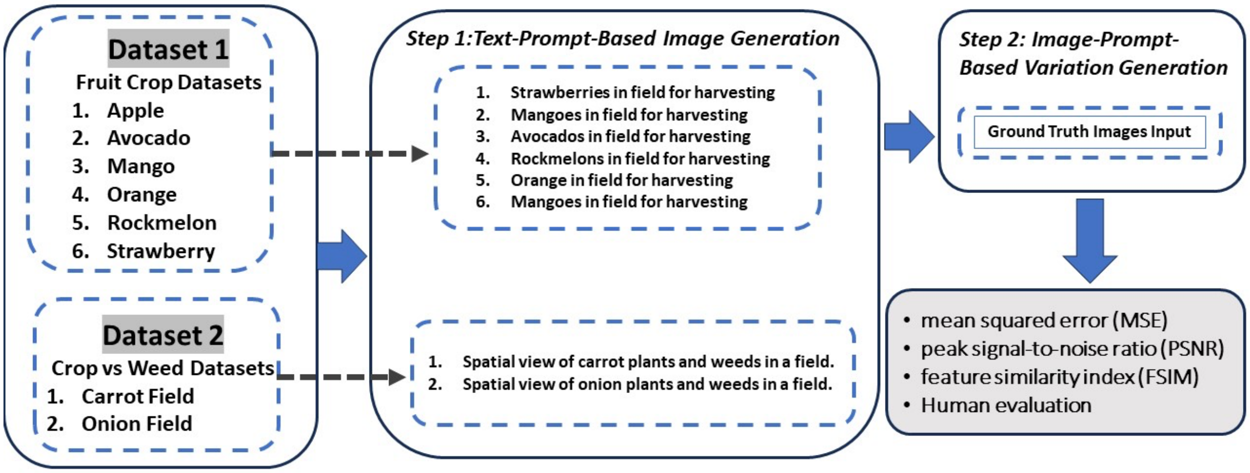
Dataset 1
Fruit Crop Datasets
1. Apple
2. Avocado
3. Mango
4. Orange
5. Rockmelon
6. Strawberry
Dataset 2
Crop vs Weed Datasets
1. Carrot Field
2. Onion Field
Step 1:Text-Prompt-Based Image Generation
1. Strawberries in field for harvesting
2. Mangoes in field for harvesting
3. Avocados in field for harvesting
4. Rockmelons in field for harvesting
5. Orange in field for harvesting
6. Mangoes in field for harvesting
1. Spatial view of carrot plants and weeds in a field.
2. Spatial view of onion plants and weeds in a field.
Step 2: Image-Prompt-Based Variation Generation
Ground Truth Images Input
• mean squared error (MSE)
• peak signal-to-noise ratio (PSNR)
• feature similarity index (FSIM)
• Human evaluation

Figure 3: Flow diagram depicting the Two-Step process utilized to generate agricultural image datasets using by the Generative AI Model DALL.E: The first step involves synthesizing images from textual prompts without any visual input, and the second step generates variations using a ground truth image as a reference.

second stage involves a decoder that creates an image based on this embedding. This method is designed to enhance the variations in the generated images while maintaining their photorealism and relevance to the caption [56]. Additionally, it allows for the generation of image variations that retain the core semantics and style, altering only the incidental details not captured in the image representation. The DALL.E model leverages diffusion models in the decoding phase to discover effective techniques for creating high-quality images. These images can be finely tuned based on textual directions, eliminating the necessity for the model to undergo specialized pre-training for distinct image editing operations.

The model consists of three stage process: encoder, prior and decoder. The model takes a textual input which is then encoded into a Compact Language-Image Pretrained (CLIP) text embedding based on a neural network trained on hundreds of millions of tax-image pairs. Dimensionality of the resulting CLIP text embedding is then reduced using Principal Component Analysis (PCA) before the results are provided the prior stage. In the prior stage, a Transformer model with an attention mechanism transforms the CLIP text embedding into an image embedding. Following the prior stage, the image embedding go through the decoder stage, also known as the unCLIP phase, in which a diffusion model based on Generative Adversarial Network (GAN) is used to convert it into an image. The output is subsequently generated through two Convolutional Neural Networks (CNNs) for upscaling: first from 64x64 resolution to 256x256, and then to a final resolution of 1024x1024. The model utilizes semantic components, handling inpainting tasks, and altering images based on subtle changes in the contextual understanding of the input text to produce the output.

2.3 Text-to-image generation

In this study, text prompts displayed in Figure 3 were created to generate images across the two specified categories. These text prompts were carefully designed to ensure that the synthetic images conveyed significant information, closely representing the real images. Initially, a manual analysis of randomly selected ground truth (actual) images was conducted for six fruit crops and two “crop vs weed” environments. Text prompts were then crafted based on the visual characteristics of the ground truth images, with input text ranging from a minimum of 4 words to a maximum of 10 words, as illustrated in Figure 3. For all fruit crops, the input text prompts were uniform, describing the "*name of the fruit* in the field for harvesting," where "in the field for harvesting" was a common phrase reflecting the harvesting condition, making a 5-word input text prompt used for each fruit crop categories. In the case of “crop vs weed” datasets, considering the spatial nature of the ground truth images for carrot and onion fields (captured from UAV aerial views), the input text was formulated as "Spatial view of *name of the crop* plants and weeds in a field," resulting in a 10-word input text prompt for generating these two sets of “crop vs weed” images. Altogether 32 images were generated using this approach.

2.4 Image-to-image (variations) generation

In this approach, actual images (ground truth images) representing the two specific datasets were provided to the model as input image prompts as shown in Figure 4. The model was then activated to generate four variations of the given input image upon receiving the command "Generate Variations." An illustration of this approach to image generation is depicted in Figure 3. Altogether, 32 images were generated using this approach.

2.5 Analysis of the generated images

In this research, the generated images were analyzed to assess their fidelity and realism. Evaluation metrics like PSNR and FSIM were employed to quantify the similarity between AI-generated and ground-truth images. Additionally, human evaluations by 15 scholars from Washington State University, Irrigated Agriculture Research and Extension Center (IAREC) provided subjective insights into the realistic portrayal of these images. Figure 4 depicts the image analysis procedure used. All generated image datasets, obtained through the two approaches discussed above, underwent a standardized preprocessing procedure. Initially, the images were converted to grayscale and resized to a resolution of 256 by 256 pixels for subsequent pixel-level analysis. For

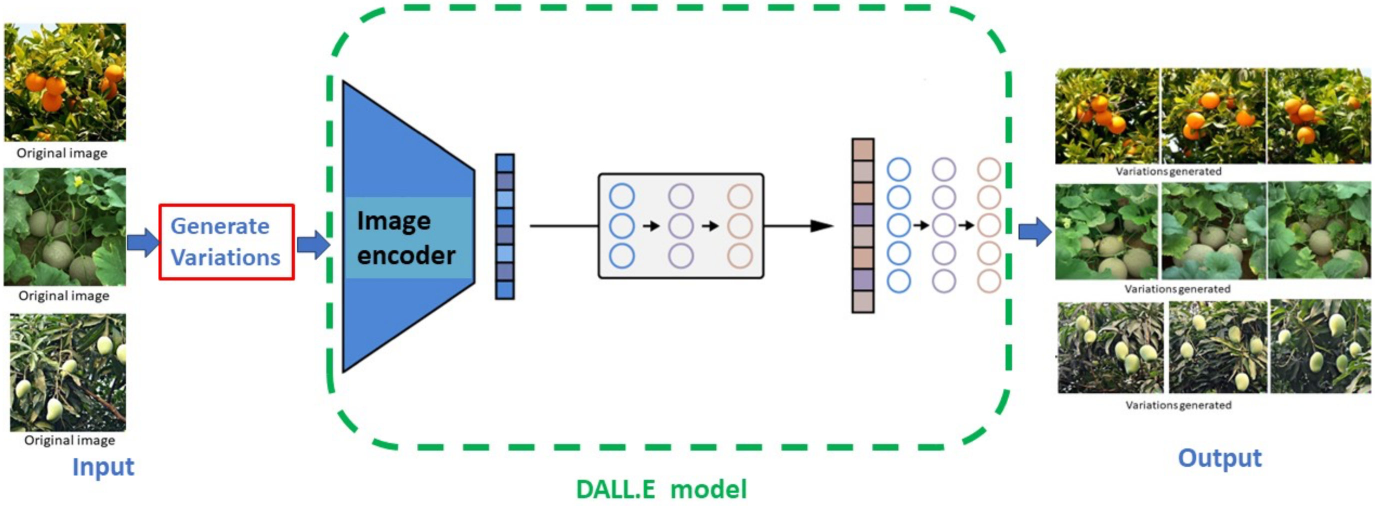

Figure 4: An example showing image-to-image variation generation using the DALL.E model

the statistical comparison of the generated images with the respective ground truth images, the images were resized and converted to grayscale as shown in Figure 5.

2.6 Evaluation Measures

In this study, the images generated in both Step 1 (text-prompt-based image generation) and Step 2 (image-prompt-based image variation generation) were compared against the ground truth images using two standard metrics as follows.

1) Peak Signal to Noise Ratio (PSNR): PSNR served as a metric for assessing image quality by evaluating the ratio of the maximum potential power of the signal (represented by the original image) to the power of disruptive noise (capturing the disparities between the original and the AI-generated image) as given by Equation 1. Mean Squared Error (MSE) estimated using Equation 2 was used to calculate this ratio. A higher PSNR typically indicates that the generated image is closer in quality to the original image and has minimal distortion.

$$PSNR = 10\log_{10}(\frac{MAX_I^2}{MSE})$$

Equation 1

Where, MAX_I denotes the maximum possible pixel value in the image and Mean Squared Error (MSE) is computed as the average of the squared differences between corresponding pixels in the two images. In simple words, MSE helps in understanding how much the generated image (G) deviates from the original image (O) on a pixel-by-pixel basis.

2) Feature Similarity Index (FSIM): FSIM assesses the similarity between the AI-generated images and the original/actual images based on their features. It evaluates both basic and intricate image features, providing a thorough measure of similarity. FSIM considers aspects like structure, luminance, and contrast of the images. Although the exact calculation of FSIM (Equation 3) involves complex comparisons at

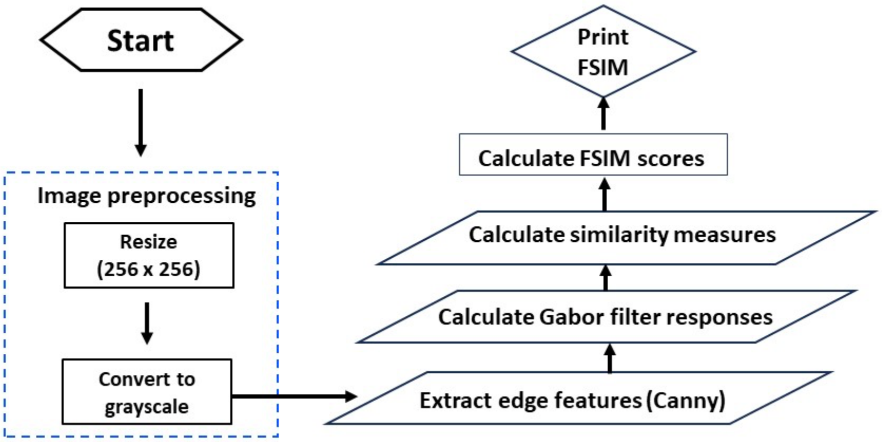
Start
Image preprocessing
Resize
(256 x 256)
Convert to
grayscale
Extract edge features (Canny)
Calculate Gabor filter responses
Calculate similarity measures
Calculate FSIM scores
Print
FSIM

Figure 5: Block diagram showing the process of calculating feature similarity index (FSIM)

multiple scales, the key idea is that it measures how closely the features of the generated image match those of the original image [57]. In this analysis, the images were preprocessed by being resized to 256 by 256 pixels and then converted into grayscale images. Canny edge features were then extracted from the grayscale images to calculate the Gabor filter responses. These responses were used to calculate the similarity measures and evaluate the feature similarity score, as shown in Figure 5.

$$MSE = (\frac{1}{N})\sum_{i=1}^{N}(O - G)^2$$

Equation 2

$$FSIM = \prod_{k=1}^{K}\left(\frac{l_k.c_k.s_k}{L_k.C_k.S_k}\right)^{\alpha_k}$$

Equation 1

Where, l_k , c_k , and s_k are the local similarity, contrast, and structure measurements at scale k, respectively, and α_k is the weight assigned to each scale.

3) Human Evaluation: A group of 15 agricultural scholars including graduate students and professors from Biological Systems Engineering and Horticulture departments at Washington State

University participated in an unbiased survey to evaluate the realism of images generated by the DALL·E model. To ensure independence and minimize bias, the same set of images used for PSNR and FSIM analysis (8 images each for ground truth, text-generated set and image-variation-generated set) was provided to the participants. The participants were unaware of whether the individual images were generated using AI or were acquired in the field using a camera. Each participant was given only the name of the crop environment with no additional information. They used a 5-point likelihood scale to rate the realism, ranging from 'Not at all realistic' (1) to 'Extremely realistic' (5).

3 Results and Discussion

3.1 Image generation for fruit crops

Fruit crop images generated by the DALL·E model from textual inputs are presented in Figure 6. The model accurately depicted strawberries in the field condition with plastic mulch (Figure 6a), mangoes on tree branches (Figure 6b), mature apples in a tree (Figure 6c), avocados in tree canopies with foliage (Figure 6d), a rockmelon with its characteristic netted skin (Figure 6e), and oranges on tree section realistically as shown in Figure 6f. Each image effectively illustrated the distinct morphological features of the fruits and their environments, showcasing the model's ability to create detailed and contextually precise visual representations from text descriptions. Variations of these fruit crops, based on ground truth images, were also generated by the model. These variations were subtly different yet retained the essence of the original images. For strawberries, variations in the ground cover were shown; avocados were depicted hanging in different positions; apples were presented in various cluster formations; and the oranges, rockmelons, and mangoes were characterized by their vibrant colors, unique textures, and distinct shapes.

Likewise, the image generated by the DALL.E model using the ground truth image as an input to generate image-variations are depicted in Figure 7. Each of the six fruit types including strawberries (Figure 7a), avocados (Figure 7b), apples (Figure 7c), mangoes (Figure 7d), rockmelons (Figure 7e), and oranges (Figure 7f) were characterized by subtle yet realistic modifications, maintaining the essence of the original, ground truth images.

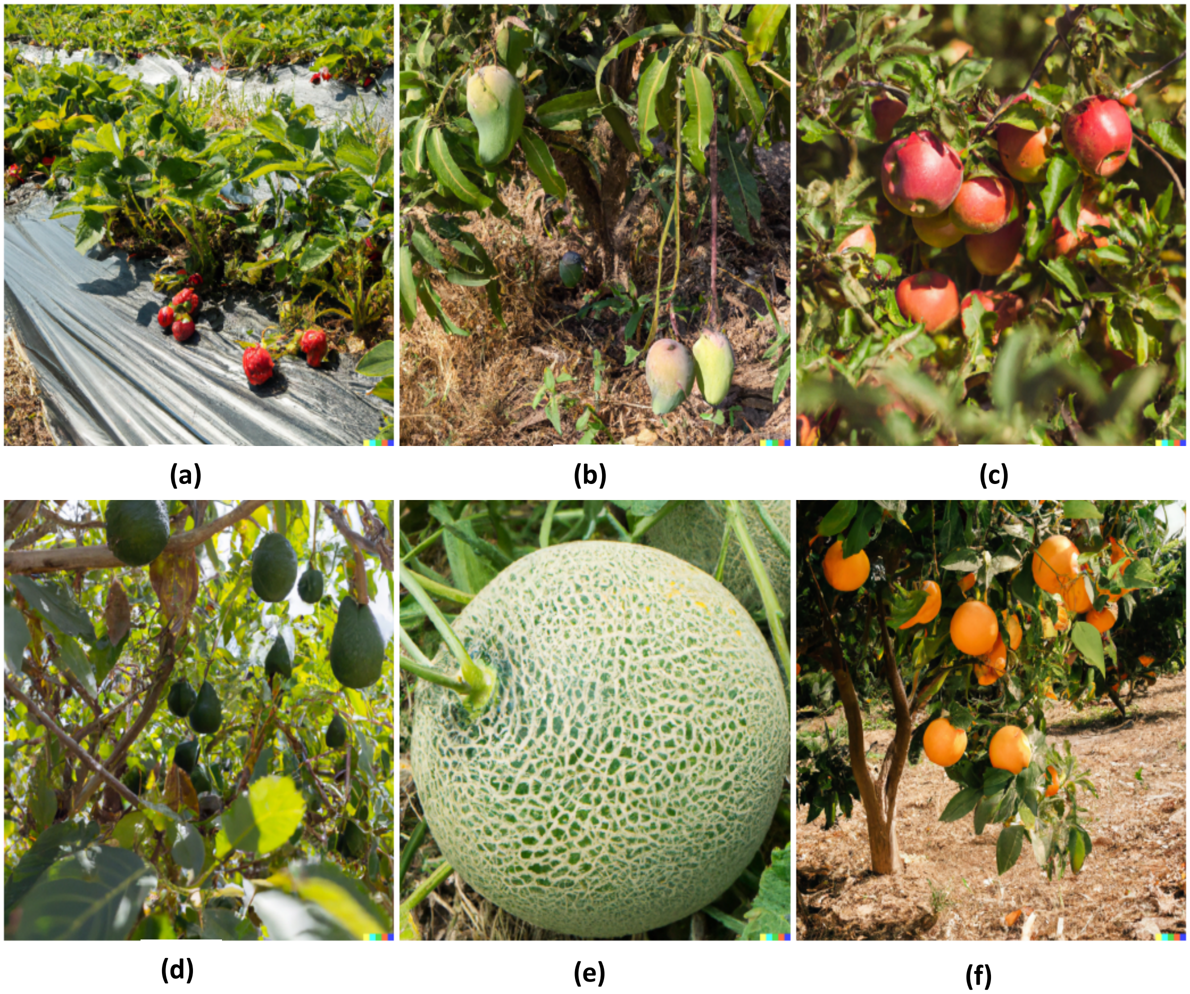

Figure 6: Fruit crop images generated using text-to-image generation approach using the DALL.E model; (a) strawberries; (b) mangoes; (c) apples; (d) avocados; (e) rockmelon; and (f) oranges.

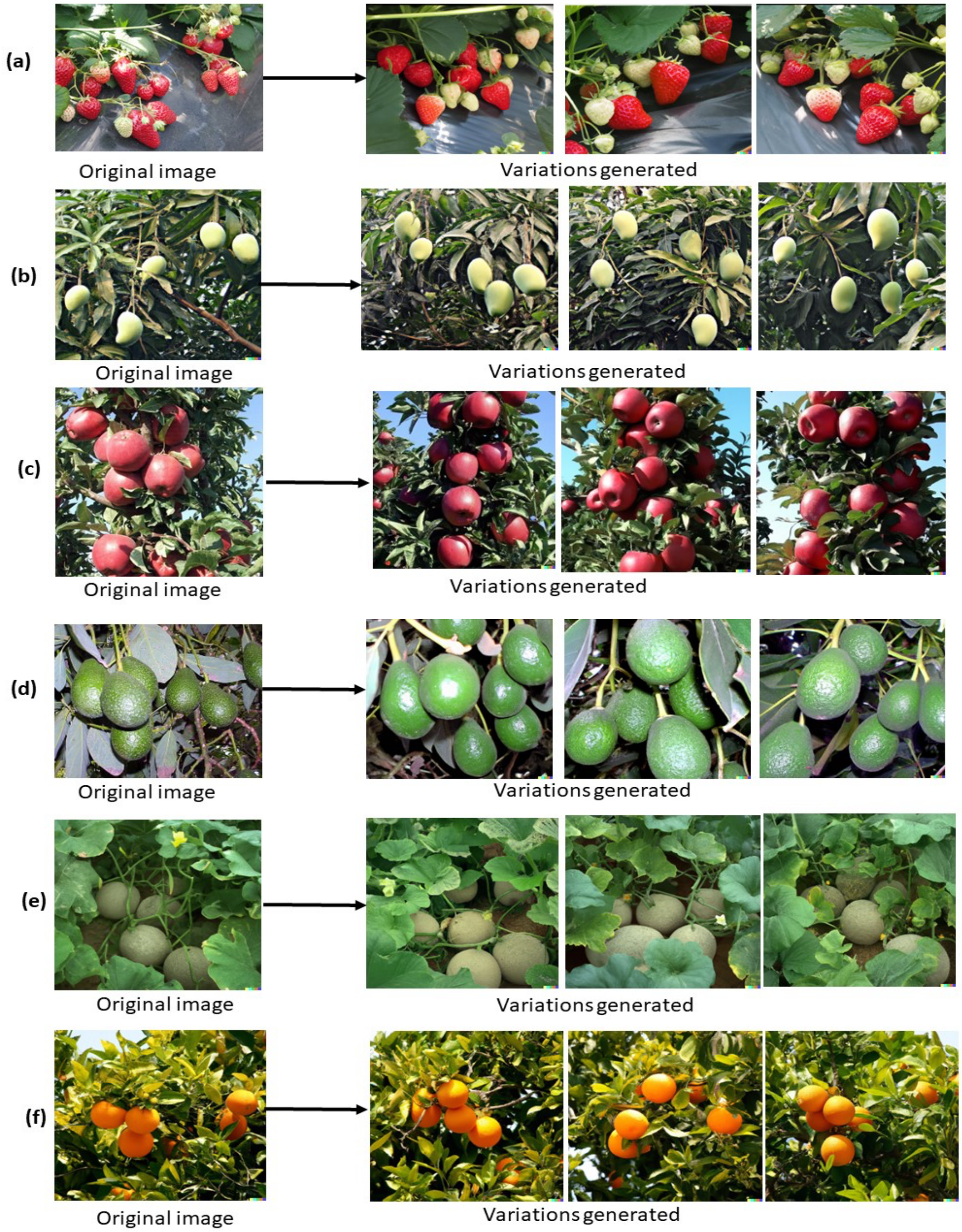

Figure 7: Three variations (right) of fruit crop images generated by DALL·E 2 model using original images (left) as an input; (a) strawberries; (b) mangoes; (c) apples; (d) avocados; (e) rockmelon; and (f) oranges.

3.1.1 Quantitative Similarity Measures

For the text-generated images, as shown in Figure 8a the PSNR values ranged from a low of 8.3 for rockmelons to a high of 10.6 for avocados, indicating a variation in the model's ability to replicate image quality. In contrast, the PSNR for image-generated (variation) images was 14.6 for mangoes, suggesting a potential for superior representation of the reality of fruit crop environments. However, the lowest PSNR in this category was 8.8 for strawberries, highlighting a potential weakness in representing the reality in a wider range of agricultural environments.

In the assessment of FSIM scores for text-generated images, as illustrated in Figure 8b, a range was observed from 0.248 for avocados to 0.308 for rockmelons. This variation was recorded, showcasing an inverse relationship with the PSNR values, which suggests a differential capacity of the model in capturing structural features and textures. Specifically, while avocados achieved the highest realism according to PSNR, indicating minimal distortion from the original in terms of brightness and contrast, they were found to be the least representative in terms of structural similarity according to FSIM. Conversely, rockmelons, which presented the highest FSIM scores, reflecting superior structural and textural fidelity, did not score as highly in PSNR, suggesting possible discrepancies in pixel-level accuracy or contrast. This inverse relationship between FSIM and PSNR scores for avocados and rockmelons indicates that the model's ability to generate realistic images varies significantly depending on the criteria used for evaluation. It implies that while some images may closely match the original in pixel intensity and contrast (as PSNR measures), they may not as effectively capture the structural integrity or texture (as FSIM evaluates), and vice versa.

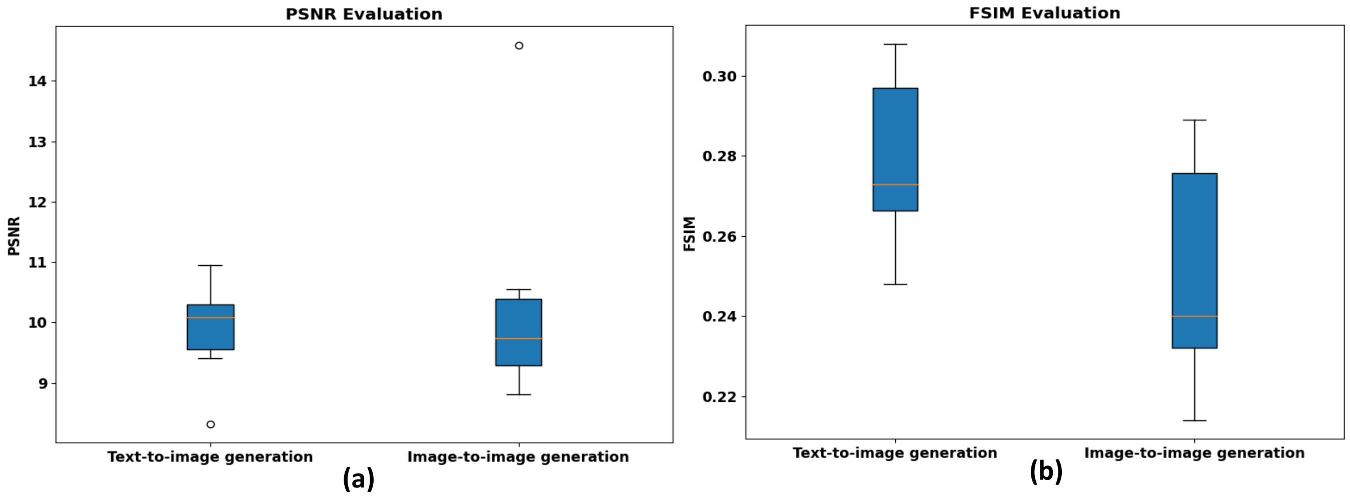

Figure 8: Box plots illustrating the distribution of PSNR and FSIM for all fruit crops tested; a) comparing text-to-image; b) image-to-image generation methods in agricultural AI applications.

The observation that image variation generation underperforms in comparison to text-to-image generation, as measured by PSNR and FSIM, may initially seem counterintuitive. However, this outcome can be attributed to the inherent differences in the model's approach to generating images from textual versus image prompts. Text-to-image generation relies on the model's understanding and interpretation of textual descriptions to create an image from scratch, potentially allowing the model to "idealize" the output, closely matching key features described in the text while maintaining overall coherence and fidelity. In contrast, image variation generation starts with an existing image and attempts to introduce variations within the constraints of the original image's context. This process may inherently limit the extent to which the model can optimize for clarity and structural similarity, as it must balance between preserving the original image's integrity and introducing meaningful variations. As a result, the variations might introduce or exacerbate minor discrepancies in texture or structural details, which could explain the lower PSNR and FSIM scores. This suggests a trade-off in the model's performance between generating novel images from textual descriptions and modifying existing images to create variations, highlighting the challenges in achieving both high fidelity and meaningful diversity in generated images.

These results suggest a generally lower performance in maintaining feature similarity compared to the original images, particularly in the case of avocados (0.2) when compared to text-generated images. While both text and image-prompt approaches displayed strengths in certain aspects, there were notable variations in performance across different fruit types and metrics. This analysis indicates that the model's effectiveness in generating accurate and realistic images in diverse agriculture environment is promising but is dependent on crop types and cropping environments.

3.1.2 Human Evaluation Results

Results of the human assessment of the AI-generated and original images for all six fruit crops are depicted in Figure 9. In the text-to-image category, apples consistently received high ratings, indicating a strong capability of the AI model to interpret textual prompts and generate realistic visual representations. On the other hand, Avocados recorded lower ratings, suggesting challenges in capturing their unique textures, colors and/or other features through text descriptions alone. In generating the image-to-image variations, Mangoes and Rockmelons received notably high ratings, showcasing the model's proficiency in creating realistic variations from existing images. The lower ratings for Strawberries in this category might reflect difficulties in maintaining the fruit's distinct characteristics in generating variations. Ground truth images, as expected, generally received the highest ratings across all categories, affirming their authenticity, which also indicated that there is a huge room for improvement in AI modeling to replicate complexities in the plant canopies and agricultural fields.

Despite those challenges, it is noted that there were instances where text-based or image-based AI generations outperformed the original images in specific fruit crops. For example, image-to-image variations of Mangoes and Rockmelons occasionally surpassed ground truth ratings. This could be attributed to the AI's ability to enhance certain visual aspects, such as color vibrancy or clarity, making them more appealing than the actual photographs to human observers. The success in these instances shows the potential of AI-based image generation to not only replicate but potentially improve upon real-world images of agricultural fields.

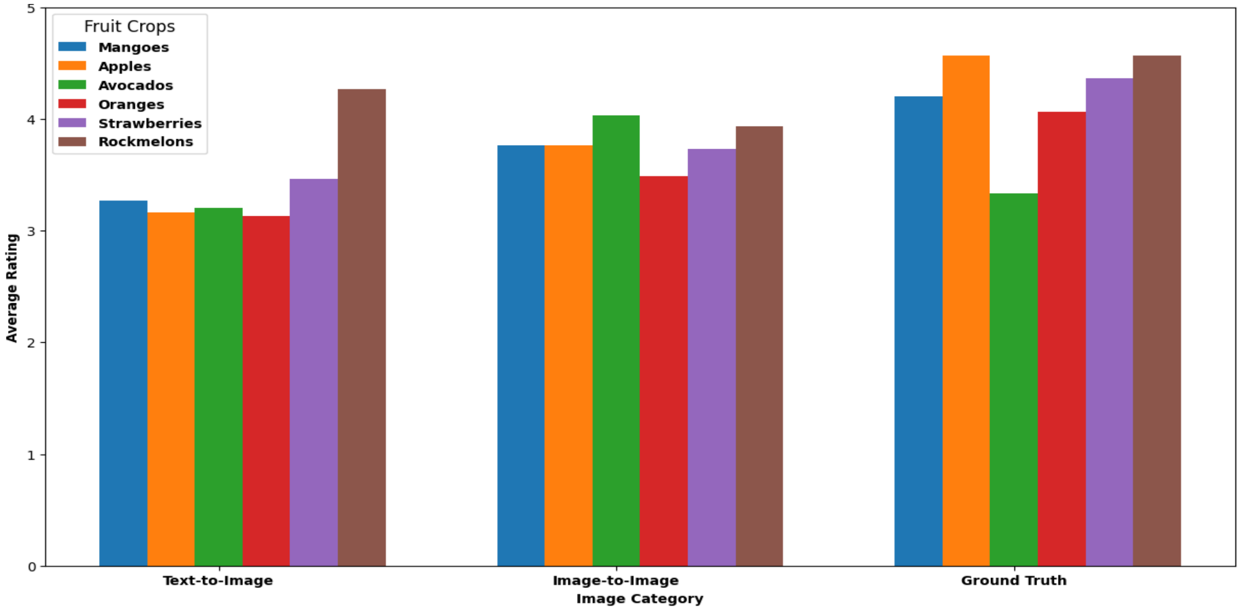

Figure 9: Bar chart illustrating average human evaluation ratings for Text-to-Image, Image-to-Image variations, and Ground Truth across six different fruit crops images in this survey of image generation process using Generative AI

3.2 Image generation results for “crop vs weed” scenario

In Figure 10a, the carrot field image, generated from the textual prompt *"Spatial view of carrot plants and weed in a field"* exhibited a representation where carrot-like colored pixels were randomly distributed across an area predominantly resembling weed patches. This indicated that while the model successfully recognized the color attributes of carrots, it struggled to accurately replicate their geometric and structural

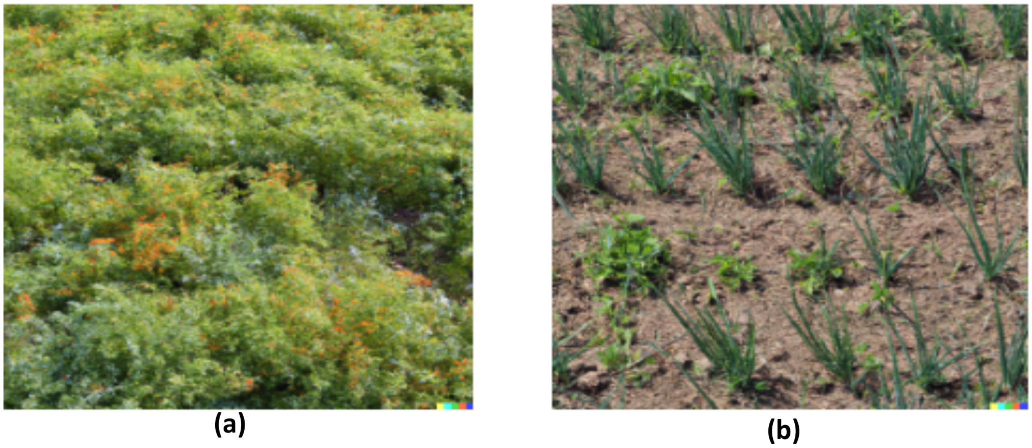

Figure 10: "Crop vs weed" images generated using text-to-image generation approach using the DALL.E model for; (a) Carrot fields; and (b) Onion fields

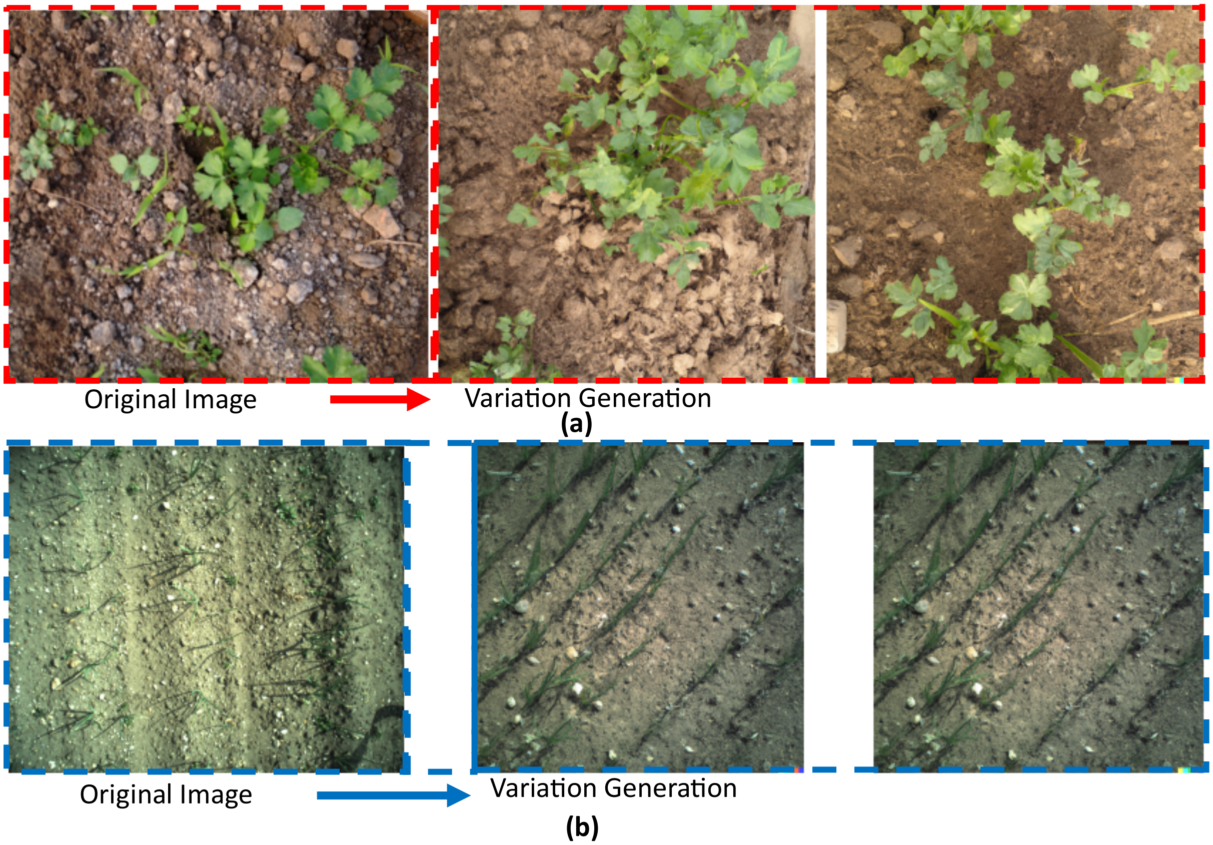

Figure 11: Crop vs weed images created by the DALL.E model by generation of image-to-image variation for a) Carrot plants b) Onion plants

characteristics at the plant level, particularly in a spatial context of crop and weed scenarios. Conversely, as depicted in Figure 10b, the onion field images showcased a higher degree of realism. The text-to-image generation method for onions, using a 10-word prompt, notably surpassed the carrot field in producing realistic and accurate representations. As with fruit crops, this disparity highlighted the model's varying capability in interpreting and visualizing different agricultural scenarios. Here, identical text prompts *"Spatial view of **carrot** plants and weed in a field"* and *"Spatial view of **onion** plants and weed in a field"* were used, differing only the crop name. Despite the similarity in prompts, the resulting images varied significantly. Specifically, no recognizable carrot plants were generated in the carrot field scenario, a stark contrast to the onion field images as shown in Figures 11a and 11b.

3.2.1 Quantitative Similarity Measures

For the images depicting “crop vs weed” scenarios, the PSNR revealed a high difference in performance between the two crop scenarios (Figure 12a and Figure 12b). For carrot fields, the model achieved a PSNR of 13.4, which was the highest among the two, indicating a superior quality in the synthetic images generated for this crop type. In contrast, onion fields exhibited a lower PSNR of 10.4, suggesting a comparatively reduced fidelity in image generation. Similarly, the FSIM scores as shown in Figure 11b, measuring the structural similarity between the generated and original images, showed a comparable result. Carrot fields secured an FSIM score of 0.3 same as that of onion fields (0.3). The AI-generated carrot field images not only exhibited higher image quality but also a marginally better structural resemblance to the ground truth images. This subtle difference indicates that while both crop types were reasonably well-represented in terms of structural features, the carrot fields edged out slightly in replicating the original images' structural details.

Figure 13 presents a heatmap detailing the MSE, PSNR, and FSIM for both Text-to-Image and Image-to-Image (variation) generated outputs. For fruit crops, text-to-image generation yielded PSNR values up to 10.95 (for avocados), indicating a commendable image quality. However, image-based variations surpassed this, with rockmelons achieving a higher PSNR of 14.592, suggesting a closer resemblance to actual images. The Feature Similarity Index (FSIM) echoed this trend; while text-generated images like mangoes scored 0.308, indicating satisfactory structural similarity, image-based generation scored marginally lower at 0.287 for the same fruit. In contrast, the crop vs weed scenario demonstrated a different trend. Text-to-image generation for onions achieved a PSNR of 13.375, closely followed by image-based variations at 13.626. The FSIM scores, measuring structural similarity, remained consistently high across both methods, with text-based generation scoring 0.33 and image-based 0.32 for onions. These results indicate that while text-based generation showed notable proficiency, image-based variations generally offered enhanced clarity and fidelity. The best results was achieved for rockmelons image-based generation, whereas the lowest was in text-based generation for avocados. This pattern suggests that while AI can generate reasonably accurate representations from textual descriptions, providing image prompts leads to more precise and realistic visual outputs, especially in complex agricultural scenarios.

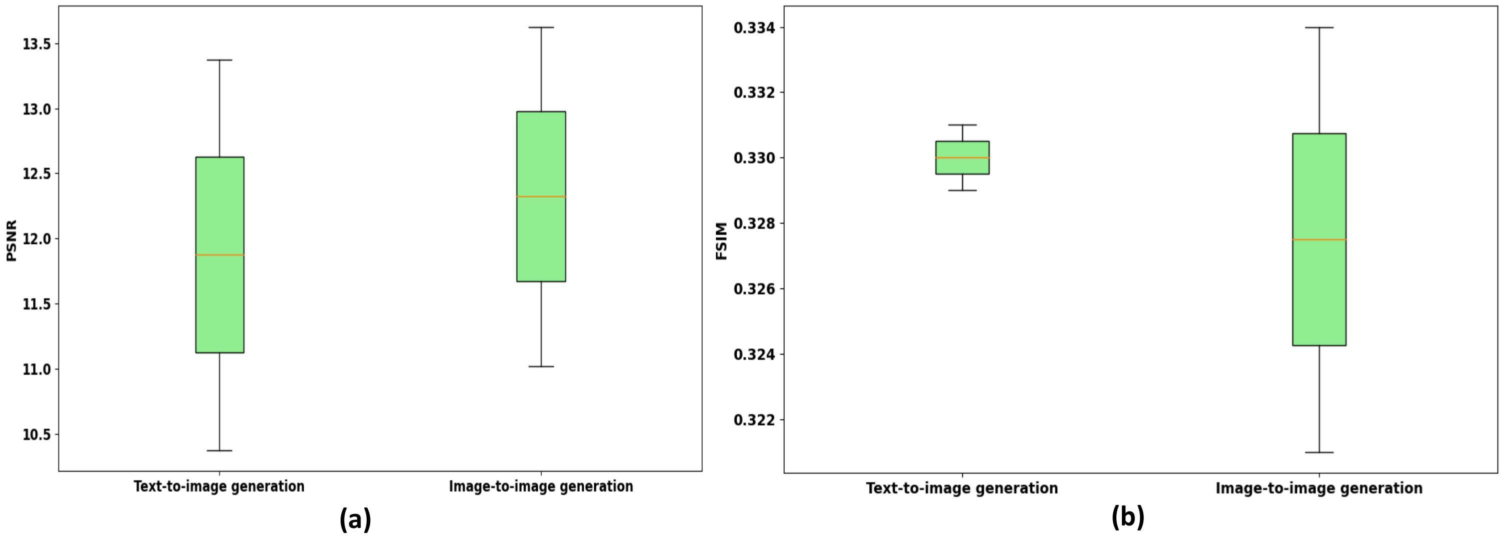

Figure 12: Box Plot for Crop vs Weed Scenarios on carrot and onion fields infested with weeds showing a) Signal to Noise Ration (PSNR) and ; b)Feature similarity measure index (FSIM)

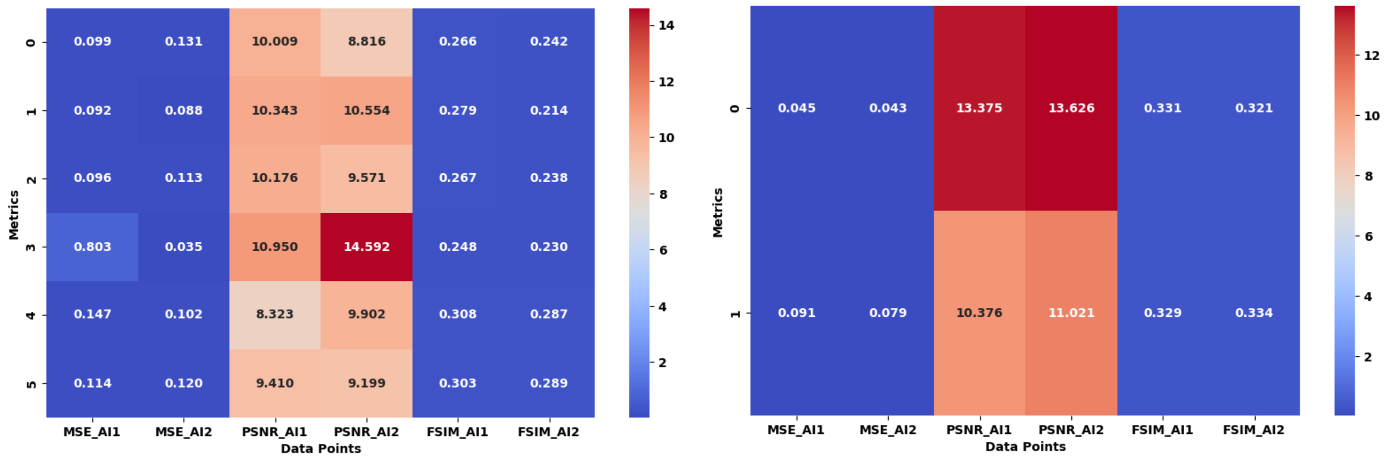

Figure 13: Heatmap for the DALL.E Model Performance: Showcases a detailed comparison of MSE, PSNR, and FSIM metrics across (a) fruit crops and (b)crop vs weed scenario, using text-to-image and image variation generations approaches

The differential performance observed between text-to-image and image-to-image generation methods, particularly in the context of carrot fields, underscores the significance of input modality in influencing the AI's output quality. While text descriptions alone sometimes fell short in capturing the intricate details necessary for a lifelike representation, supplementing the AI with actual image inputs markedly improved its ability to generate images that closely mimic reality. This improvement is attributed to the model's enhanced access to visual context, allowing for a more accurate interpretation and recreation of the subject matter.

For fruit crops, the image-to-image generation method consistently outperformed the text-to-image approach. Notably, rockmelons generated through image-to-image variations exhibited high clarity and detail, as evidenced by superior PSNR score of 14.6. This was indicative of the AI's proficiency in producing clear and detailed images, essential for tasks like crop growth monitoring and yield estimation. Mango crop images, generated through both methods, showcased high structural similarity with the actual images achieving an FSIM score of 0.31 for text-to-image generation method and 0.29 in image-to-image generation method. These results also highlight the AI's capability to maintain structural integrity, crucial for shape-based agricultural tasks such as on fruit crops results.

3.2.2 Human Evaluation Results

As shown in Figure 14, a detailed image analysis was conducted to assess the effectiveness of image-to-image and text-to-image generation techniques across different agricultural scenarios. For carrot fields, the image-to-image variation technique demonstrated superior performance compared to text-to-image based generation, achieving better clarity and details that even surpassed the ground truth images. This result is supported by human evaluators awarding these images the highest possible scores, with a peak at 5, compared to the ground truth images which received a broader score range from 2.0 to 5.0. The finding indicates a capability of the image-to-image variation approach in enhancing the visual quality beyond the original photographs. In contrast, the carrot fields generated using the text-to-image approach were met with more moderate success. These images received scores ranging from 2.5 to 4.0 by human evaluators, suggesting that while the generated images were of acceptable quality, they exhibited certain limitations in capturing the full detail and reality of the actual fields. In the onion field images, however, a different trend was observed. The text-to-image generation technique for onion fields yielded better outcomes, with scores ranging from 4.0 to 5.5. These scores closely mirrored the ratings assigned to the ground truth images, indicating a high degree of accuracy and realism in the images generated from textual descriptions.

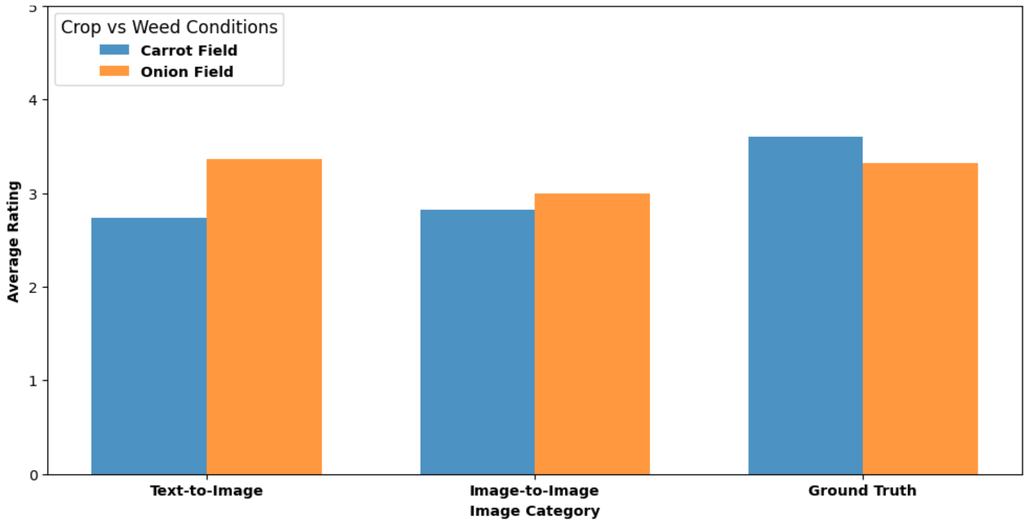

Figure 14: Showing average human evaluation scores for Text-to-Image and Image-to-Image (variation) generations, compared with Ground Truth for crop vs weed scenario images, demonstrating the efficacy of Generative AI enabled model DALL.E in image dataset creation for agricultural applications.

These results demonstrated that DALL.E model can be used to generate large image datasets for agricultural applications with good accuracy and structural integrity. Such AI-generated images can significantly simplify the data generation process, reducing time and costs associated with traditional methods that rely on advanced sensors and intensive field data collection. The findings suggest that DALL.E 2's capabilities in image generation hold potential for advancing machine vision and robotic operations in agriculture, contributing to the development of more efficient and accurate AI-driven agricultural systems and accelerate their adoption.

Previous studies for image generation in the agricultural environments typically depended on labor-intensive and expensive field data collection, often hindering efficiency. However, in this study, we showed an efficient workflow of creating agricultural images using AI that could potentially avoid the reliance on labor-intensive and costly field data collection methods in the near future. Our evaluation of the feature similarity between AI-generated images and real sensor-captured images of crop environments not only validates the practical utility of this technology but also opens up new possibilities for its application in precision agriculture. This shift towards AI-generated imagery could potentially revolutionize the way agricultural studies are conducted, offering a more cost-effective, rapid, and versatile method of data collection.

4 Conclusion and Future Prospects

In modern agriculture, the need for comprehensive image datasets is paramount, especially given the limitations of traditional data collection methods, which are often labor-intensive and time-consuming. Synthetic image generation emerges as a compelling solution, addressing these challenges by creating realistic and diverse datasets efficiently. The utilization of AI-based methods, particularly the DALL.E model developed by OpenAI, exemplifies this approach. Functioning similarly to its counterpart ChatGPT, DALL.E is trained on a vast array of images and textual data, enabling it to generate accurate and diverse images from textual descriptions and existing images. The DALL.E model's potential in agricultural applications is, therefore, substantial. It offers innovative solutions for critical tasks such as fruit quality assessment, automated harvesting, and crop yield estimation. By generating realistic images of various crops and cropping environments, DALL.E aids in the development of smart farming techniques. For instance, the model's ability to create images of fruits in different growth stages can help in training AI models for precise fruit detection, thus improving crop monitoring and harvesting strategies. Similarly, its capacity to depict “crop-versus-weed” scenarios can help enhance weed detection algorithms, facilitating targeted weeding. This study conducted a detailed evaluation of the DALL.E model's efficacy in generating agriculturally relevant images, focusing on its ability to replicate and enhance real-world field conditions through synthetic imagery. We systematically compared the generated images against actual field data, assessing their realism and applicability in supporting advanced agricultural practices and research.

Based on the results, the following specific conclusions can be made from this study:

- Image-to-image generation methods resulted in a 5.78% increase in average PSNR, indicating improved image clarity and quality over text-to-image generation. However, there was a decrease of 10.23% in average FSIM for image-to-image generation, suggesting a reduction in structural and textural similarity to the original images compared to text-to-image generation.

- For image-to-image generation, PSNR saw an increase of 3.77%, reflecting enhanced image precision compared to text-to-image generation. Image-to-image generation also experienced a slight FSIM decrease of 0.76%, indicating a minimal drop in feature similarity with the original images versus text-to-image generation.

This study underscores the potential transformative impact of integrating advanced AI models like DALL.E 2 into advancing agricultural technologies and solutions. The successful application of this model in generating realistic images for various agricultural scenarios opens up new opportunities for enhancing agricultural efficiency and improving crop yield and quality. By leveraging the capabilities of DALL.E 2, a generative AI model based on Large Language Models (LLMs), the agricultural sector could see a significant shift in how data is gathered and analyzed. The traditional reliance on sensors and manual data collection processes, often cumbersome and time-intensive, could be greatly reduced or even be completely replaced in the future. Instead, AI-generated images, as demonstrated in this study, could provide a more efficient and scalable alternative. The ability of models like DALL.E 2 to create accurate depictions of diverse agricultural environments from different crop stages to complex crop vs. weed scenarios offers new potential for smart and precision agricultural practices. In the future, tasks like yield estimation, disease detection, and crop health monitoring could be conducted using datasets generated entirely by AI, streamlining the process and increasing its accuracy and adoptability.

Looking ahead, the advancement in generative AI techniques like DALL.E model holds the promise of automatically creating accurately labeled image datasets. This innovation paves the way for the development of virtual orchards and digital twins, revolutionizing agricultural planning and management with precision and foresight.

References

1. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* **5**, 493–497 (2021).
2. Barrera, K., Merino, A., Molina, A. & Rodellar, J. Automatic generation of artificial images of leukocytes and leukemic cells using generative adversarial networks (syntheticcellgan). *Comput Methods Programs Biomed* **229**, 107314 (2023).
3. Choi, I., Park, S. & Park, J. Generating and modifying high resolution fashion model image using StyleGAN. in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)* 1536–1538 (IEEE, 2022).
4. Bermano, A. H. *et al.* State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. in *Computer Graphics Forum* vol. 41 591–611 (Wiley Online Library, 2022).
5. Luo, C., Wang, Y., Zhang, X., Zhang, W. & Liu, H. Spatial prediction of soil organic matter content using multiyear synthetic images and partitioning algorithms. *Catena (Amst)* **211**, 106023 (2022).
6. Rio-Torto, I., Campaniço, A. T., Pereira, A., Teixeira, L. F. & Filipe, V. Automatic quality inspection in the automotive industry: a hierarchical approach using simulated data. in *2021 IEEE 8th International Conference on Industrial Engineering and Applications (ICIEA)* 342–347 (IEEE, 2021).
7. Sapkota, B. B. *et al.* Use of synthetic images for training a deep learning model for weed detection and biomass estimation in cotton. *Sci Rep* **12**, 19580 (2022).
8. Man, K. & Chahl, J. A Review of Synthetic Image Data and Its Use in Computer Vision. *J Imaging* **8**, 310 (2022).
9. Chen, A. *et al.* Three dimensional synthetic non-ellipsoidal nuclei volume generation using bezier curves. in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 961–965 (IEEE, 2021).
10. Álvarez-Trejo, A., Cuan-Urquizo, E., Roman-Flores, A., Trapaga-Martinez, L. G. & Alvarado-Orozco, J. M. Bézier-based metamaterials: Synthesis, mechanics and additive manufacturing. *Mater Des* **199**, 109412 (2021).

11. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P. P. & Barron, J. T. Nerf in the dark: High dynamic range view synthesis from noisy raw images. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16190–16199 (2022).

12. Hodaň, T. *et al.* Photorealistic image synthesis for object instance detection. in *2019 IEEE international conference on image processing (ICIP)* 66–70 (IEEE, 2019).

13. Zhao, Z. & Bao, G. Artistic Style Analysis of Root Carving Visual Image Based on Texture Synthesis. *Mobile Information Systems* **2022**, (2022).

14. Velikina, J. V, Alexander, A. L. & Samsonov, A. Accelerating MR parameter mapping using sparsity-promoting regularization in parametric dimension. *Magn Reson Med* **70**, 1263–1273 (2013).

15. Araújo, T., Mendonça, A. M. & Campilho, A. Parametric model fitting-based approach for retinal blood vessel caliber estimation in eye fundus images. *PLoS One* **13**, e0194702 (2018).

16. Diolatzis, S., Philip, J. & Drettakis, G. Active exploration for neural global illumination of variable scenes. *ACM Transactions on Graphics (TOG)* **41**, 1–18 (2022).

17. Zhang, Y. *et al.* Modeling indirect illumination for inverse rendering. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 18643–18652 (2022).

18. Eversberg, L. & Lambrecht, J. Generating images with physics-based rendering for an industrial object detection task: Realism versus domain randomization. *Sensors* **21**, 7901 (2021).

19. Wu, X., Xu, K. & Hall, P. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Sci Technol* **22**, 660–674 (2017).

20. Matuszczyk, D., Tschorn, N. & Weichert, F. Deep Learning Based Synthetic Image Generation for Defect Detection in Additive Manufacturing Industrial Environments. in *2022 7th International Conference on Mechanical Engineering and Robotics Research (ICMERR)* 209–218 (IEEE, 2022).

21. Yu, J. *et al.* Generative image inpainting with contextual attention. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 5505–5514 (2018).

22. Abbas, A., Jain, S., Gour, M. & Vankudothu, S. Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Comput Electron Agric* **187**, 106279 (2021).

23. Lu, C.-Y., Rustia, D. J. A. & Lin, T.-T. Generative adversarial network based image augmentation for insect pest classification enhancement. *IFAC-PapersOnLine* **52**, 1–5 (2019).

24. Nazki, H., Lee, J., Yoon, S. & Park, D. S. Image-to-image translation with GAN for synthetic data augmentation in plant disease datasets. *Smart Media Journal* **8**, 46–57 (2019).

25. Lu, Y., Chen, D., Olaniyi, E. & Huang, Y. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Comput Electron Agric* **200**, 107208 (2022).

26. Liu, B., Tan, C., Li, S., He, J. & Wang, H. A data augmentation method based on generative adversarial networks for grape leaf disease identification. *IEEE Access* **8**, 102188–102198 (2020).

27. De, S., Bhakta, I., Phadikar, S. & Majumder, K. Agricultural Image Augmentation with Generative Adversarial Networks GANs. in *International Conference on Computational Intelligence in Pattern Recognition* 335–344 (Springer, 2022).

28. Gomaa, A. A. & Abd El-Latif, Y. M. Early prediction of plant diseases using cnn and gans. *International Journal of Advanced Computer Science and Applications* **12**, (2021).

29. Madsen, S. L., Dyrmann, M., Jørgensen, R. N. & Karstoft, H. Generating artificial images of plant seedlings using generative adversarial networks. *Biosyst Eng* **187**, 147–159 (2019).

30. Zhu, F., He, M. & Zheng, Z. Data augmentation using improved cDCGAN for plant vigor rating. *Comput Electron Agric* **175**, 105603 (2020).

31. Hartley, Z. K. J. & French, A. P. Domain adaptation of synthetic images for wheat head detection. *Plants* **10**, 2633 (2021).

32. Bird, J. J., Barnes, C. M., Manso, L. J., Ekárt, A. & Faria, D. R. Fruit quality and defect image classification with conditional GAN data augmentation. *Sci Hortic* **293**, 110684 (2022).

33. Shete, S., Srinivasan, S. & Gonsalves, T. A. TasselGAN: An Application of the generative adversarial model for creating field-based maize tassel data. *Plant Phenomics* (2020).

34. Guo, Z. *et al.* Quality grading of jujubes using composite convolutional neural networks in combination with RGB color space segmentation and deep convolutional generative adversarial networks. *J Food Process Eng* **44**, e13620 (2021).

35. Drees, L., Junker-Frohn, L. V., Kierdorf, J. & Roscher, R. Temporal prediction and evaluation of Brassica growth in the field using conditional generative adversarial networks. *Comput Electron Agric* **190**, 106415 (2021).

36. Kierdorf, J. *et al.* Behind the leaves: estimation of occluded grapevine berries with conditional generative adversarial networks. *Front Artif Intell* **5**, 830026 (2022).

37. Olatunji, J. R., Redding, G. P., Rowe, C. L. & East, A. R. Reconstruction of kiwifruit fruit geometry using a CGAN trained on a synthetic dataset. *Comput Electron Agric* **177**, 105699 (2020).

38. Bellocchio, E., Costante, G., Cascianelli, S., Fravolini, M. L. & Valigi, P. Combining domain adaptation and spatial consistency for unseen fruits counting: a quasi-unsupervised approach. *IEEE Robot Autom Lett* **5**, 1079–1086 (2020).

39. Fawakherji, M., Potena, C., Pretto, A., Bloisi, D. D. & Nardi, D. Multi-spectral image synthesis for crop/weed segmentation in precision farming. *Rob Auton Syst* **146**, 103861 (2021).

40. Zeng, Q., Ma, X., Cheng, B., Zhou, E. & Pang, W. Gans-based data augmentation for citrus disease severity detection using deep learning. *IEEE Access* **8**, 172882–172891 (2020).

41. Kim, C., Lee, H. & Jung, H. Fruit tree disease classification system using generative adversarial networks. *International Journal of Electrical and Computer Engineering (IJECE)* **11**, 2508–2515 (2021).

42. Tian, Y., Yang, G., Wang, Z., Li, E. & Liang, Z. Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense. *J Sens* **2019**, (2019).

43. Cap, Q. H., Uga, H., Kagiwada, S. & Iyatomi, H. Leafgan: An effective data augmentation method for practical plant disease diagnosis. *IEEE Transactions on Automation Science and Engineering* **19**, 1258–1267 (2020).

44. Maqsood, M. H. *et al.* Super resolution generative adversarial network (Srgans) for wheat stripe rust classification. *Sensors* **21**, 7903 (2021).

45. Bi, L. & Hu, G. Improving image-based plant disease classification with generative adversarial network under limited training set. *Front Plant Sci* **11**, 583438 (2020).

46. Zhao, Y. *et al.* Plant disease detection using generated leaves based on DoubleGAN. *IEEE/ACM Trans Comput Biol Bioinform* **19**, 1817–1826 (2021).

47. Nerkar, B. & Talbar, S. Cross-dataset learning for performance improvement of leaf disease detection using reinforced generative adversarial networks. *International Journal of Information Technology* **13**, 2305–2312 (2021).

48. Zhao, L., Zheng, K., Zheng, Y., Zhao, D. & Zhou, J. RLEG: vision-language representation learning with diffusion-based embedding generation. in *International Conference on Machine Learning* 42247–42258 (PMLR, 2023).

49. Schuhmann, C. *et al.* Laion-5b: An open large-scale dataset for training next generation image-text models. *Adv Neural Inf Process Syst* **35**, 25278–25294 (2022).

50. Van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. ChatGPT: five priorities for research. *Nature* **614**, 224–226 (2023).

51. McLean, S. *et al.* The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence* **35**, 649–663 (2023).

52. Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D. & Smith, A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Health* **5**, e105–e106 (2023).

53. Lu, Y. & Young, S. A survey of public datasets for computer vision tasks in precision agriculture. *Comput Electron Agric* **178**, 105760 (2020).

54. Ding, M., Zheng, W., Hong, W. & Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Adv Neural Inf Process Syst* **35**, 16890–16902 (2022).

55. Conde, M. V & Turgutlu, K. CLIP-Art: Contrastive pre-training for fine-grained art classification. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3956–3960 (2021).

56. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* **1**, 3 (2022).

57. Zhang, L., Zhang, L., Mou, X. & Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* **20**, 2378–2386 (2011).