Impact of Synthetic Data from Diffusion Models on Weed Detection Performance

José D. G. Ramos[1], Tatiana T. Schein[1], Larissa S. Gomes[1], Stephanie L. Brião[1], Kristofer S. Kappel[1], Paulo V. K. Borges[3], Felipe G. Oliveira[1,2], Rodrigo da S. Guerra[1] and Paulo L. J. Drews-Jr[1]

***Abstract*—Weeds represent a significant challenge to agriculture, making their accurate detection essential for minimizing crop losses and ensuring sustainable farming practices. Traditional weed detection methods often fail to adapt to changing conditions, making integrating advanced technologies like robotics, computer vision, and deep learning crucial. AI-driven robotics in precision agriculture enables real-time weed identification and targeting, reducing herbicide use while enhancing farming efficiency and promoting more sustainable practices. However, these methods depend on diverse and high-quality data. Taking the aforementioned into account, in this work, we propose i) a novel approach to generate synthetic data for weed detection using a combination of diffusion models and clustering techniques and ii) an impact analysis of the use of synthetic data during the training process of learning models for weed detection operation. The methodology integrates IP-Adapter and SeeCoder models with a DINOv2-based clustering algorithm, enabling efficient synthetic data generation without requiring network retraining or complex prompt engineering. Experiments conducted on the CottonWeedYolo dataset demonstrated the effectiveness of the proposed method. Our synthetic images achieved a CLIP Mean Maximum Discrepancy (CMMD) score of 1.317, very close to the 1.153 images generated with DreamBooth, while requiring significantly less computational resources. Incorporating synthetic images using YOLOv8 improved mAP across all species, with the best performance observed when combined with a balanced real dataset. The results demonstrate that synthetic data, although valuable as a complement, do not replace the need for real data, highlighting the importance of quality over quantity in developing robust detection networks.**

***Index Terms*—Synthetic Data Generation, Diffusion Models, Weed Detection, Computer Vision, Agricultural Robotics.**

I. INTRODUCTION

In the face of global climate change and rapid population growth, agriculture ensures food supplies. The sector is under increasing pressure due to rising temperatures, unpredictable weather patterns, and extreme events such as droughts and floods, further exacerbated by plant diseases, pests, and weed infestations [1]. To address these issues, adopting cutting-edge technologies has become essential to optimize resource use and ensure food security [2].

Weeds are unwanted plants that grow in areas where they are not cultivated, usually competing with crops or ornamental plants for essential resources such as light, water, nutrients, and space. Weeds can hinder the growth of cultivated plants, reducing productivity and, in some cases, even destroying crops. In addition, many weeds can be difficult to control or remove. They can emerge naturally in any environment and are generally adapted to grow quickly and survive under various conditions [3].

Traditionally, weed detection has relied on manual inspection and mechanical or chemical control methods, often labor-intensive, time-consuming, and costly. Farmers and agricultural workers visually identify weeds in the field, which becomes increasingly challenging in large-scale farms where precision and efficiency are critical. Additionally, manual detection is prone to human error, as weeds can sometimes resemble crops, making accurate identification difficult [4], [5], as we can observe in Figure 1.

Fig. 1. Example of a plantation with the presence of weeds, emphasizing the degree of difficulty in detecting the infestation. In the left figure, a plantation with weeds is presented, while in the right figure, the weeds are highlighted.

These challenges highlight the importance of integrating robotics, computer vision, and machine learning to improve the accuracy and efficiency of weed detection, enabling more precise and scalable solutions for modern precision agriculture [6].

Considering the abovementioned, we propose a study of a Dreambooth training method with an image clustering tool

*The authors would also like to thank the PRH-ANP, CNPQ, FINEP, FAURG, FAPERGS and FAPEAM organizations for their research support and financial assistance. This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2024/10523-5.

[1]Centro de Ciências Computacionais (C3). Universidade Federal do Rio Grande, Rio Grande - RS, Brasil. {jgarcia,tatischein,larissaesilva,stephanie.loi,kris,rodrigo.guerra,paulodrews}@furg.com

[2]Instituto de Ciências Exatas e Tecnologia – ICET. Universidade Federal do Amazonas – UFAM. Itacoatiara, AM, Brazil. felipeoliveira@ufam.edu.br

[3]Data 61. Commonwealth Scientific and Industrial Research Organization (CSIRO). Pullenvale - QLD, Australia. paulo.borges@data61.csiro.au

979-8-3315-5288-6/25/$31.00©2025 IEEE

based on DINOv2. Additionally, we will introduce a new pipeline to generate data without the need to retrain the network, without using prompts, and without relying solely on the training data of the diffusion network. To validate our approach, we conducted extensive experiments to evaluate hybrid training strategies, combining real and synthetic datasets.

The main contributions of our work are summarized as follows:

- We propose a novel approach to generate synthetic data for weed detection using a combination of diffusion models and clustering techniques.
- We present an impact analysis of the use of synthetic data during the training process of learning models for weed detection tasks.

II. RELATED WORK

Synthetic data generation in agriculture is necessary to overcome collection limitations and poor-quality data. Diffusion models can create this data for applications such as pest detection. This section reviews advances, applications, and challenges in this context.

A. Deep Learning Applications and Challenges in Agriculture

Artificial intelligence is a tool that can optimize processes in various fields, particularly agronomy. In this domain, reviews such as [7] and [4] highlight both classical solutions and those based on Deep Learning, which have been widely adopted in the current industry. These technologies enhance precision in crop monitoring, yield prediction, and detecting fertilizers, diseases, and pests [3], [8].

However, a recurring challenge in agriculture is data acquisition for training neural networks. The lack of data due to high costs and complexity of acquisition, poor quality of open-source data, highly specific scenarios, or variable resolutions that negatively affect the network architecture makes it difficult for robust model development [9]. Furthermore, the generalization capability of networks is impacted by the limited amount of available data [7].

For this reason, several studies seek ways to augment these datasets. Early approaches involve complex data augmentation methods, such as the one used in [10], which applies a transparent image method over grass. However, these approaches generally result in a significant amount of time spent and require specialized personnel to perform these augmentations.

B. Diffusion Models for Synthetic Data Generation in Agriculture

Diffusion models have made significant strides in generating high-quality synthetic images, becoming valuable tools for database augmentation. For instance, DiffuseMix [11] focuses on modifying tones, contrasts, and effects in base images, while Effective Data Augmentation [12] generates diverse variations within the same class. However, these models face limitations due to their reliance on training datasets. Although these datasets are extensive, they often lack specificity for certain species [13], which restricts their practical applicability.

One prominent approach is Dreambooth [14], which generates images based on a set of example inputs. Studies such as [15] and [6] have utilized this model but with varying outcomes. While [15] reported failures in generating plant images, [6] achieved better results using a similar database [16]. This variability can be attributed to Dreambooth's numerous adjustable parameters, which can lead to overfitting [13] and the critical role of input image distribution in model performance. Selecting a limited data set can reduce diversity and create imbalances in the generated outputs [17].

To address these challenges, newer models like RIVAL [18] and GDA [19] have been proposed. These models adapt latent distributions to create more precise databases. However, their applicability is limited by their reliance on specialized architectures and pre-trained networks, which constrain their flexibility. Another notable challenge is the use of prompts for image generation. Slight prompt modifications can lead to vastly different results [11], [20], highlighting the need for more robust control mechanisms. Recent research, such as [21], has explored replacing the CLIP architecture [22] with visual inputs to improve control over generated features. Similarly, [23] introduces an attention layer to precisely guide the generation process, balancing control and visual quality. These advancements underscore the need for more efficient and accessible approaches to synthetic data generation.

C. Evaluation Metrics for Synthetic Data in Generative Models

The evaluation of synthetic data often relies on the Frechet Inception Distance (FID) [24], which measures the discrepancy between the feature distributions of real and generated images. However, FID has several limitations. It assumes that Inception-v3 embeddings follow a multivariate normal distribution, an assumption that does not hold for the complex outputs of modern generative models [25]. Additionally, FID requires large sample sizes to produce reliable estimates and fails to consistently reflect incremental improvements or degradations in image quality [26].

To overcome these limitations, the CLIP-MMD Distance (CMMD) has been proposed [25]. Using an RBF Gaussian kernel, this metric combines CLIP embeddings with Maximum Mean Discrepancy (MMD). Unlike FID, CMMD does not rely on assumptions about data distribution, is computationally efficient, and effectively captures incremental changes in image quality. Empirical studies have shown that CMMD aligns more closely with human perception, making it a robust alternative for evaluating modern generative models.

III. METHODOLOGY

A key question arises from the state-of-the-art analysis: Can synthetic data effectively enhance deep learning training? Studies like [15] and [6] yield contrasting results despite applying similar methodologies to the same dataset, underscoring the need for a deeper investigation into methodological influences and a standardized baseline for future research.

To address this, we propose a series of experiments to assess method variability and establish a unified framework. For consistency in comparisons, we adopt the [16] dataset, previously used in related studies.

A. Synthetic Images Generation by DreamBooth Model

Synthetic image generation supports data augmentation strategies. Here, we use diffusion models, which iteratively denoise random noise based on learned distributions to produce high-quality images.

For our experiments, we adopt a diffusion-based approach inspired by DreamBooth, which enables the generation of class-specific images using a small set of reference images. DreamBooth fine-tunes a pre-trained diffusion model to learn the distinctive characteristics of each class, ensuring the production of diverse and accurate synthetic samples. According to [14], three to eight reference images per class are sufficient to capture key features, making this an efficient method for synthetic data generation.

However, one of the challenges in the synthetic image generation process is the preparation of a high-quality reference dataset. A carefully curated and diverse set of images is necessary to ensure the generation of accurate and varied samples by the model.

B. Clustering based on the DINOv2 Algorithm

One of the main aspects of building an appropriate reference dataset for training the image generation model is the clustering of images. Since training a Dreambooth model requires between three and eight reference images per class, efficiently clustering the images into manageable groups is a priority. Typically, this process involves selecting images based on their ”similarity” [6].

However, this clustering has two main limitations: subjectivity in determining image similarity and a loss of efficiency in large datasets. These limitations can affect the quality and utility of the generated data.

The DINOv2 proposal [27], designed for feature extraction, combined with the K-nearest neighbors (KNN) for clustering, addresses the aforementioned limitations and clusters the dataset into distinct groups. This approach promises higher quality in forming groups, which is relevant for the subsequent pipeline stages.

C. Enhanced Synthetic Data Generation

Based on the state-of-the-art analysis, we propose a methodology for automatic synthetic data generation that addresses key challenges in data augmentation while reducing computational costs. Instead of training a diffusion model from scratch, which is time-consuming and requires significant Video Random Access Memory (VRAM), our approach leverages two complementary methods: IP-Adapter [23] and SeeCoder [21].

IP-Adapter is a modification of the original diffusion model architecture that enhances the attention calculation. This modification introduces an additional step where the model compares the embeddings of guide images, improving the model's ability to generate images based on reference data rather than textual prompts [23]. SeeCoder, on the other hand, extracts visual features from an image, overcoming the vocabulary limitations of text-to-image models by using images as prompts instead of text [21].

The reason for combining these two approaches lies in their complementary strengths. While SeeCoder extracts the visual features of an image, it still suffers from certain limitations in image generation, particularly in terms of alignment. Although the visual prompt serves as a guide to generate unseen images, there is often a gap between the generated result and the original prompt. The images generated using SeeCoder may lack visual realism and deviate significantly from the reference image in terms of content and quality.

On the other hand, the IP-Adapter still faces challenges related to prompt engineering, requiring significant time and effort to identify the optimal prompt for effective data generation. The trial-and-error process of selecting the right words to guide the model can make it difficult to create diverse classes, potentially breaking synthetic data generation.

We hypothesize that both models improve the generation process by addressing different limitations. SeeCoder enhances visual feature extraction, enabling more diverse image generation. At the same time, the improved attention mechanism in IP-Adapter helps align the generated images with the desired output, enhancing realism and relevance. Our generation model leverages both and can achieve better results without requiring extensive, prompt engineering or additional training.

Figure 2 demonstrates how both IP-Adapter and SeeCoder are used together with guide images to enhance synthetic data generation. The figure illustrates how the embeddings from the guide images are processed through the modified attention mechanism of IP-Adapter and the visual feature extraction of SeeCoder, resulting in more accurate and realistic image generation.

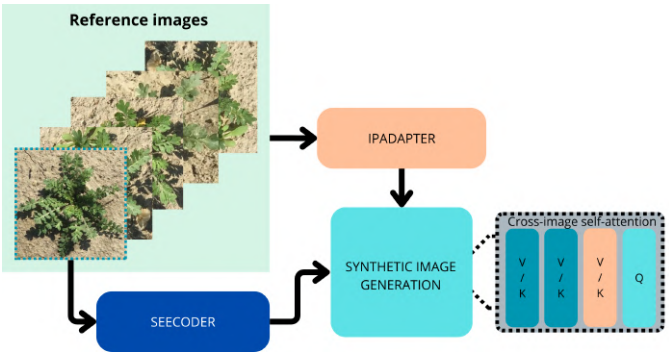

Fig. 2. Combined use of IP-Adapter and SeeCoder with DINOv2-clustered guide images for enhanced synthetic data generation.

D. Experimental Setup

The experiments were conducted using the YOLOv8 detection network across all configurations to ensure architectural consistency. The models were trained on the CottonWeedYolo dataset [16], which consists of 4150 images distributed across

14 weed species: Crabgrass, Eclipta, Goosegrass, Morningglory, Nutsedge, PalmerAmaranth, Prickly Sida, Ragweed, Sicklepod, Spottedspurge, SqurredAnoda, Swinecress, Waterhemp, and Purslane. This dataset was partitioned into 80% training, 10% validation, and 10% testing subsets. The models were trained for 10 epochs with an image size of 640, using an automatic batch size determined by the Ultralytics package on an RTX 4060TI GPU. The inherent class imbalance in the original dataset poses significant challenges for detection models. Seven experimental configurations were designed to address this and systematically evaluate the role of synthetic data in mitigating the imbalance and enhancing detection performance.

- **Experiment 1:** Baseline training using the original imbalanced dataset without modifications.
- **Experiment 2:** Add 300 synthetic images per class using DreamBooth, as recommended in [20], to ensure that all classes are enhanced equally.
- **Experiment 3:** Add synthetic images proportionally to correct the initial imbalance, generating a varying number of synthetic images for each class based on their original distribution, ensuring that all classes end up with the same total number of images.
- **Experiment 4:** Reduce the number of images in the classes more represented to balance the proportions in the original dataset without adding synthetic data.
- **Experiment 5:** Add 300 synthetic images per class generated by the proposed method to balance the dataset.
- **Experiment 6:** Reduce the original dataset to 10%, without any addition of synthetic images, to evaluate the performance on a smaller, real-world dataset.
- **Experiment 7:** Use the 10%-reduced dataset, augmented with 300 synthetic images per class, to assess the impact of synthetic data on model performance when using a limited real dataset.

In all experiments, the synthetic images generated were filtered and semi-automatically labeled using AnyLabeling [28], which incorporates state-of-the-art object detection and segmentation algorithms. This tool includes a SAM2-based [29] bounding box tool, which, with manual reference points, generates both bounding boxes and segmentation masks—features that could be beneficial for future tasks.

IV. Results and Analysis

This section provides a comprehensive analysis of the proposed approach, including both qualitative and quantitative evaluations. It begins with a quantitative and qualitative assessment of the quality of the generated synthetic images, followed by a qualitative comparison of different dataset formation strategies. Finally, the computational cost of the proposed method is qualitatively analyzed.

A. Synthetic data evaluation

The quality of synthetic images generated with Stable Diffusion was evaluated using the CMMD metric, which provides a robust and reliable assessment [25]. Based on CLIP model embeddings and MMD distance, CMMD enables an accurate analysis of differences between the real and synthetic image distributions and is used in this work for this comparison. Table I presents the results obtained for different synthetic image generation methods, where * represents the prompt "crop *specie*" and *specie* each weed species name used.

TABLE I
CMMD Metric for different methods.

Methods	CMMD (↓)
Only SeeCoder [21]	2.314
Only IP-Adapter* [23]	1.647
Dreambooth* [14]	1.153
Ours	1.317

*Models trained with a "crop *specie*" prompt.

The CMMD metric assesses the proximity between real and synthetic image distributions, where lower values indicate greater similarity. DreamBooth achieved the best performance with a CMMD of 1.153. Our method, which combines the capabilities of IP-Adapter and SeeCoder to generate synthetic images, is closely followed by 1.317. Unlike DreamBooth, our approach does not require retrained data, prompt engineering, or computationally intensive training.

Individually, SeeCoder and IP-Adapter obtained CMMDs of 2.314 and 1.647, respectively. By integrating both, our method significantly improves synthetic image generation, enabling high-fidelity outputs with minimal setup and no need for specialized hardware. With only a 0.164 difference from DreamBooth, our solution offers a practical, cost-effective alternative for rapid synthetic image generation.

Regarding computational cost, our method can be executed even on devices with just a CPU or minimal GPU requirements, such as 2GB of VRAM. In contrast, DreamBooth, which achieves the second-best quality in synthetic image generation, requires 8GB of VRAM for training. Additionally, our execution time is nearly identical to the IP-Adapter.

Figure 3 presents a qualitative analysis of four species from the CottonWeedDet12 dataset [16]. The first column shows the original images, while the following rows display synthetic images generated by different methods. We applied a Dinov2d-based clustering tool to group visually similar images, forming five clusters when necessary.

The results highlight SeeCoder's inconsistency, as it often introduces artifacts despite capturing general shapes and textures. The IP-Adapter improves visual quality but produces stylized, cartoon-like outputs. DreamBooth performs best, accurately capturing shape, texture, and realism. This success is attributed to leveraging similar images during training, allowing the model to learn the distinctive characteristics of each group. Our method achieves high realism, faithfully reproducing key image features while outperforming SeeCoder and IP-Adapter. Though slightly stylized, it remains a strong alternative to DreamBooth in terms of quality and fidelity.

Moreover, it is important to emphasize that synthetic images are not guaranteed to preserve all taxonomically significant

traits of the species, as sometimes the models can just simplify them.

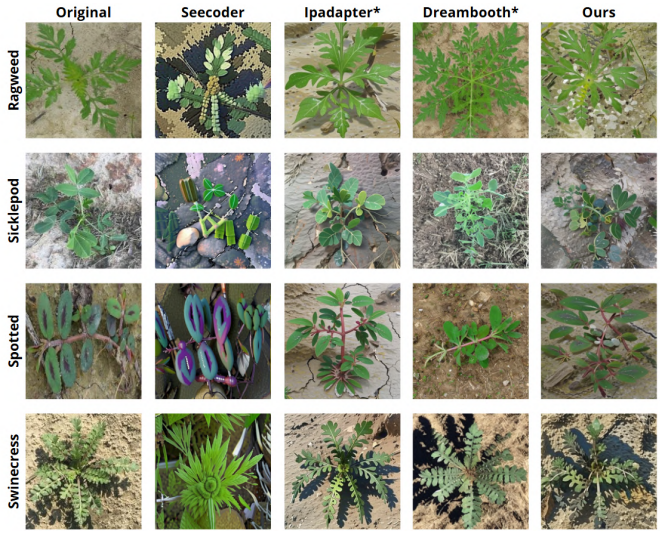

Fig. 3. Qualitative Comparison of Image Generation Methods for CottonWeedDet12 Species (*indicates prompt-based methods).

B. Evaluation of Dataset Formation for Detection

In the quantitative analysis, we evaluated the results obtained in all seven experiments for each class. In this assessment, the results were expressed through quantitative metrics that indicate the accuracy of the enhancement process. The metric used was the mean Average Precision (mAP), which is widely used to measure the performance of object detection models. The mAP was calculated for each class individually, considering an Intersection over Union (IoU) threshold of 0.5 and as the average of the Average Precisions (APs) of all classes.

Figure 4 presents the mAP values obtained in each experiment, comparing the impact of different training configurations on the performance of the YOLOv8 model. Experiment 1 shows low performance compared to its counterparts using synthetic data. Experiment 2 achieves the best results across multiple classes, although it exhibits overall instability in the model with low precision for some species. However, these shortcomings are minimal. In Experiment 3, despite adding more synthetic images, the model's overall performance decreases. This can be explained by the fact that some species have more synthetic images than others to balance the entire dataset.

On the other hand, Experiment 4 does not include synthetic images and shows significantly lower overall precision. However, when comparing Experiment 3 and Experiment 4, it is observed that, despite its lower performance, the model with synthetic images performs better overall. Finally, Experiment 5 stands out as the best on average, demonstrating the highest overall stability without sacrificing the performance of certain species in favor of others, as observed in Experiment 2.

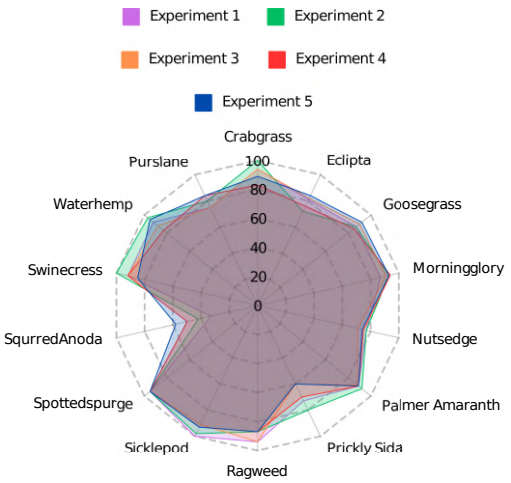

Fig. 4. mAP values obtained in the first five experiments, analyzing the impact of different training configurations on the performance of the YOLOv8 model for detecting 14 weed classes.

In comparing the results of Experiments 6 and 7, as shown in Figure 5, it is observed that number 7, which has more data (even if artificially generated), contributes to improving the model's overall precision. In Experiment 6, the dataset used is very limited, resulting in poor performance, indicating that the model failed to understand the data properly. On the other hand, with synthetic images, the model demonstrates some learning. However, asserting that the model generalizes well is still impossible, as the dataset remains quite restricted.

Thus, it can be concluded that synthetic data is useful but does not replace the need for real data, which should be as diverse as possible. The combination of synthetic and real data is essential for developing robust networks, making it crucial to prioritize quality over quantity in real data collection.

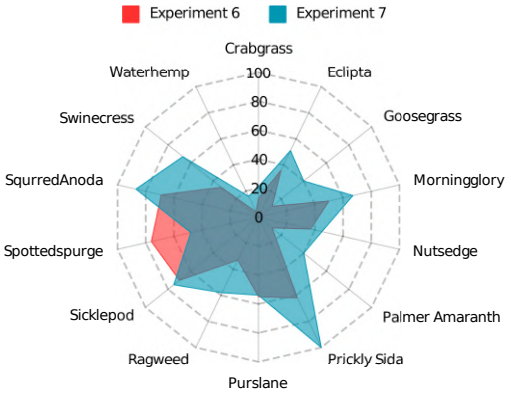

Fig. 5. Comparison of mAP between Experiments 6 and 7, highlighting the impact of adding synthetic data on the detection accuracy of the 14 weed classes with the scarcity of real data.

V. CONCLUSION

This study demonstrates the effectiveness of synthetic data generation in addressing the challenges of training datasets

for weed detection systems. Dreambooth achieves slightly better performance in generating synthetic images (CMMD of 1.153 compared to our 1.317). Our proposed pipeline eliminates the need for retraining, specialized hardware, or complex prompt engineering in the image generation process. Qualitative analysis confirms the ability of our method to generate highly realistic images. Through extensive experimentation with different dataset formations, we found that integrating synthetic images (generated by our method) and real data (Experiment 5, with 300 synthetic images per class) achieved the best overall stability and performance across all weed species. Although our image generation method has a higher computational cost (8.95s versus Dreambooth's 3.37s), this trade-off is justified by eliminating expensive training processes and reducing the need for manual data collection. These findings represent a significant advance in synthetic data generation and optimal dataset formation for agricultural applications. In particular, they pave the way for the next stage in developing robotic weed detection systems, which could significantly enhance the efficiency and sustainability of modern farming practices. By integrating synthetic data with robotic systems, we can further improve weed management solutions' scalability, precision, and adaptability, making it a crucial step toward more sustainable agriculture.

REFERENCES

[1] Bernhardt, Heinz, Bozkurt, Mehmet, Brunsch, Reiner, Colangelo, Eduardo, Herrmann, Andreas, Horstmann, Jan, Kraft, Martin, Marquering, Johannes, Steckel, Thilo, Tapken, Heiko, et al. Challenges for Agriculture through Industry 4.0. *Agronomy*, v. 11, p. 1935, set. 2021.

[2] Goyal, Rajni, Nath, Amar, e Niranjan, Utkarsh. Weed detection using deep learning in complex and highly occluded potato field environment. *Crop Protection*, v. 187, p. 106948, 2025.

[3] G. Yang, J. Wang, Z. Nie, H. Yang, and S. Yu, "A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention," *Agronomy*, vol. 13, no. 7, p. 1824, 2023.

[4] F. Xiao, H. Wang, Y. Xu, and R. Zhang, "Fruit detection and recognition based on deep learning for automatic harvesting: An overview and review," *Agronomy*, vol. 13, no. 6, p. 1625, 2023.

[5] Silva, Lucas; Drews, Paulo; de Bem, Rodrigo. Soybean weeds segmentation using VT-Net: A convolutional-transformer model. *Proceedings of the 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, p. 127–132, 2023. IEEE.

[6] H. Moreno, A. Gómez, S. Altares-López, A. Ribeiro, and D. Andújar, "Analysis of Stable Diffusion-derived fake weeds performance for training Convolutional Neural Networks," *Computers and Electronics in Agriculture*, vol. 214, p. 108324, 2023.

[7] F. Xiao, H. Wang, Y. Li, Y. Cao, X. Lv, and G. Xu, "Object detection and recognition techniques based on digital image processing and traditional machine learning for fruit and vegetable harvesting robots: an overview and review," *Agronomy*, vol. 13, no. 3, p. 639, 2023.

[8] Traversi, Nelson De Faria; Evald, Paulo Jefferson Dias De Oliveira; Dos Santos, Juliana Veiga; Junior, Paulo Lilles Jorge Drews; Botelho, Silvia Silva Da Costa. Development of Comprehensive Fertilizer Datasets: Enhancing Precision Agriculture through Data-Driven Insights. *Proceedings of the 22nd IEEE International Conference on Industrial Informatics (INDIN)*, p. 1–6, 2024. IEEE.

[9] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, 2023.

[10] S. Xie, C. Hu, M. Bagavathiannan, and D. Song, "Toward robotic weed control: detection of nutsedge weed in bermudagrass turf using inaccurate and insufficient training data," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7365–7372, 2021.

[11] K. Islam, M. Z. Zaheer, A. Mahmood, and K. Nandakumar, "DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27621–27630, 2024.

[12] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," *arXiv preprint arXiv:2302.07944*, 2023.

[13] Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

[15] B. Deng, "Stable Diffusion for Data Augmentation in COCO and Weed Datasets," *arXiv preprint arXiv:2312.03996*, 2023.

[16] F. Dang, D. Chen, Y. Lu, and Z. Li, YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems, *Computers and Electronics in Agriculture*, vol. 205, p. 107655, 2023.

[17] J. Guo, J. Zhao, C. Ge, C. Du, Z. Ni, S. Song, H. Shi, and G. Huang, "Everything to the Synthetic: Diffusion-driven Test-time Adaptation via Synthetic-Domain Alignment," *arXiv preprint arXiv:2406.04295*, 2024.

[18] Y. Zhang, J. Xing, E. Lo, and J. Jia, "Real-world image variation by aligning diffusion inversion chain," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] Y.-Y. Tsai, F.-C. Chen, A. Y. C. Chen, J. Yang, C.-C. Su, M. Sun, and C.-H. Kuo, "GDA: Generalized Diffusion for Robust Test-time Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23242–23251, 2024.

[20] J. Xie, W. Li, X. Li, Z. Liu, Y. S. Ong, and C. C. Loy, "Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation," *International Journal of Computer Vision*, pp. 1–20, 2024.

[21] X. Xu, J. Guo, Z. Wang, G. Huang, I. Essa, and H. Shi, "Prompt-free diffusion: Taking 'text' out of text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8682–8692, 2024.

[22] Radford, Alec; Kim, Jong Wook; Hallacy, Chris; Ramesh, Aditya; Goh, Gabriel; Agarwal, Sandhini; Sastry, Girish; Askell, Amanda; Mishkin, Pamela; Clark, Jack; et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, v. 139, p. 8748–8763, 2021.

[23] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[24] Heusel, Martin; Ramsauer, Hubert; Unterthiner, Thomas; Nessler, Bernhard; Hochreiter, Sepp. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, v. 30, 2017.

[25] Jayasumana, Sadeep; Ramalingam, Srikumar; Veit, Andreas; Glasner, Daniel; Chakrabarti, Ayan; Kumar, Sanjiv. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 9307–9315, 2024.

[26] de Deijn, Ricardo; Batra, Aishwarya; Koch, Brandon; Mansoor, Naseef; Makkena, Hema. Reviewing FID and SID Metrics on Generative Adversarial Networks. *arXiv preprint arXiv:2402.03654*, 2024.

[27] Oquab, Maxime, Darcet, Timothée, Moutakanni, Théo, Vo, Huy V., Szafraniec, Marc, Khalidov, Vasil, Fernandez, Pierre, Haziza, Daniel, Massa, Francisco, El-Nouby, Alaaeldin, et al. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research, 2024. *Weed Technology*, v. 37, p. 1-13, fev. 2023. doi:10.1017/wet.2023.5.

[28] V. A. Nguyen, "AnyLabeling - Effortless data labeling with AI support," [Online]. Available: https://github.com/vietanhdev/anylabeling. [Accessed: Jan. 19, 2025]. Licensed under GPL-3.

[29] Ravi, Nikhila; Gabeur, Valentin; Hu, Yuan-Ting; Hu, Ronghang; Ryali, Chaitanya; Ma, Tengyu; Khedr, Haitham; Rädle, Roman; Rolland, Chloe; Gustafson, Laura; et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.