International Journal of Technology 14(6) 1206-1215 (2023)
Received May 2023 / Revised September 2023 / Accepted September 2023

International Journal of Technology

http://ijtech.eng.ui.ac.id

Detection and Sizing of Durian using Zero-Shot Deep Learning Models

Mohtady Barakat[1], Gwo Chin Chung[1*], It Ee Lee[1], Wai Leong Pang[2], Kah Yoong Chan[1]

[1]*Faculty of Engineering, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Selangor, Malaysia*
[2]*School of Engineering, Taylor's University, 47500 Subang Jaya, Selangor, Malaysia*

Abstract. Since 2017, up to 41% of Malaysia's land has been cultivated for durian, making it the most widely planted crop. The rapid increase in demand urges the authorities to search for a more systematic way to control durian cultivation and manage the productivity and quality of the fruit. This research paper proposes a deep-learning approach for detecting and sizing durian fruit in any given image. The aim is to develop zero-shot learning models that can accurately identify and measure the size of durian fruits in images, regardless of the image's background. The proposed methodology leverages two cutting-edge models: Grounding DINO and Segment Anything (SAM), which are trained using a limited number of samples to learn the essential features of the fruit. The dataset used for training and testing the model includes various images of durian fruits captured from different sources. The effectiveness of the proposed model is evaluated by comparing it with the Segmentation Generative Pre-trained Transformers (SegGPT) model. The results show that the Grounding DINO model, which has a 92.5% detection accuracy, outperforms the SegGPT in terms of accuracy and efficiency. This research has significant implications for computer vision and agriculture, as it can facilitate automated detection and sizing of durian fruits, leading to improved yield estimation, quality control, and overall productivity.

Keywords: Durian; Grounding DINO; Segment Anything; SegGPT; Zero-shot learning

1. Introduction

Agriculture is the backbone of many economies around the world. Among various plantations, durian, often referred to as the "king of fruits", is of great economic importance in many countries, especially in Southeast Asia (Allied Market Research, 2021). In the countries where it is extensively grown, such as Thailand, Malaysia, and the Philippines, durian farming is a major income-generating activity. For example, in 2020, Thailand, the world's leading producer of durians, harvested about 1.05 million metric tonnes of fruit (FAOSTAT, 2021). On the export front, China's enormous appetite for durians is notable. According to a report from the Thai Office of Agricultural Economics (Saowanit, 2017), in 2016, durian exports to China amounted to over 158,000 metric tonnes, valued at approximately $0.24 billion. Durian's unique traits and high economic value have also led to considerable investments in research and development to improve cultivation techniques, disease resistance, and post-harvest handling. For instance, the Malaysian government has invested about $2.8 million in durian research and development since 2016.

*Corresponding author's email: gcchung@mmu.edu.my, Tel.: +603-83125309
doi: 10.14716/ijtech.v14i6.6640

In recent years, technology has significantly influenced traditional farming practices, a shift often called 'Precision Agriculture' (Haziq *et al.,* 2022; Pedersen and Lind, 2017). One of the critical tasks in precision agriculture is the accurate measurement of fruit size, which assists in yield prediction, harvest planning, and post-harvest processing. Durian's size detection holds substantial importance due to the fruit's variable and large size. While traditional methods of durian cultivation have relied on skilled farmers' expertise to determine the fruit's ripeness and quality, the global agricultural landscape is rapidly evolving. The appeal of integrating advanced deep-learning techniques into durian cultivation stems from several factors: scalability and consistency, data-driven insights, market differentiation, and future-proofing (Kamilaris and Prenafeta-Boldú 2018). With the advent of machine learning and deep learning techniques, automated fruit detection and size estimation have become a possibility. Various models have been explored, including Convolutional Neural Networks (CNNs) (Tan *et al.,* 2022), You Only Look Once (YOLO), etc.

For instance, (Pedersen and Lind, 2017) demonstrated the use of machine learning algorithms for predicting fruit yield, quality, and disease detection. Similarly, image processing and computer vision have been widely studied for their potential in the ripeness classification of oil palm fresh fruit bunches (Mansour, Dambul, and Choo, 2022). Besides that, the authors used deep-simulated learning for fruit counting, thereby highlighting the potential of deep learning in precision agriculture (Rahnemoonfar and Sheppard 2017). The proposed Inception-ResNet architecture showed 91% average test accuracy on real images. Similarly, (Mohanty, Hughes, and Salathé, 2016) used deep CNNs for image-based plant disease detection and achieved nearly 100% accuracy in detection under certain conditions. Last but not least, a game-theoretic model of the species and varietal composition of fruit plantations has also been developed (Belousova and Danilina, 2021).

While traditional machine learning techniques have shown promising results, deep learning, with its ability to learn complex features, has taken a step ahead. The Grounding DINO and Segment Anything models have been proven effective in object detection and image segmentation tasks. (Maninis, Radosavovic, and Kokkinos, 2019) demonstrated the potential of Segment Anything in attentive single-tasking of multiple tasks. The model has shown exceptional results in various image segmentation tasks in terms of reducing the number of parameters while maintaining or even improving the overall performance. On the other hand, Grounding DINO, known for its disentangled non-autoregressive object detection, has also shown promising results (Carion *et al.,* 2020). The zero-shot learning approach used by the Grounding DINO has raised the interest of many researchers recently for its low computational complexity and promising performance (Wang *et al.,* 2019). Another extensive review of zero-shot learning has also been done (Pourpanah *et al.,* 2020).

In this paper, we propose a zero-shot deep learning model using the Grounding DINO and Segment Anything models for durian fruit size detection. Our methodology centers around a two-step process; we first employ Grounding DINO for initial object detection and then use Segment Anything (SAM) to segment and mask each durian fruit for size determination. Notably, we implement a zero-shot training approach, eliminating the need for data collection and manual annotation. We compare the performance of the proposed model with that of the Segmentation Generative Pre-trained Transformers (SegGPT) model. Performance accuracy is analyzed and improved with several different box threshold and text threshold adjustments. Last but not least, the results of the optimized detection models will be demonstrated on practical images. Accurate fruit size detection, which will be examined in this paper, holds considerable implications for precision agriculture. The ability to reliably quantify fruit sizes in real-time can transform numerous farming practices, from yield prediction and harvest planning to post-harvest processing.

This paper is organized as follows: Section 2 presents the research methodology by defining the concept of one-shot learning models such as Grounding DINO, SegGPT, and SAM. It also discusses the design process for the detection and sizing of durian. Section 3 illustrates the results of durian size detection using the above-mentioned models. It includes performance comparison and analysis of the Grounding DINO, SegGPT, and SAM models. Lastly, Section 4 comprises a summary of the durian size detection using our proposed deep learning models as well as future recommendations.

2. Research Methodology

2.1. Zero-Shot Learning

Zero-shot learning is a concept in machine learning that refers to the ability of a model to understand and perform tasks that it has not seen or has not seen much of during training (Xian *et al.,* 2018). The name "zero-shot" comes from the fact that the model receives zero examples, or "shots," of these new tasks during training.

Zero-shot learning is based on the idea of learning a shared semantic space that connects the input data with the corresponding outputs or labels. This shared space is typically learned from the training data. The outputs, or labels, are often represented as vectors in this shared space. For example, in a text classification task, each label could be represented as a vector derived from a language model. When a zero-shot learning model encounters new data, it maps the new data into this shared semantic space. Then, it matches the new data with the output or label that is closest to this shared space. This matching process could be done in various ways, like by computing the cosine similarity between the new data and each output or label vector.

2.1.1. Grounding DINO

Grounding DINO, an Artificial Intelligence (AI) model developed by OpenAI is a significant advancement in the fields of deep learning and computer vision. The model derives its name from a combination of "grounding", a process that connects the understanding of vision and language in the AI system, and "DINO", which stands for Distillation of Knowledge Without Labels, a framework developed by META AI (Caron *et al.,* 2021).

The architecture of Grounding DINO is built upon the principles of Vision Transformers (ViT) and the DINO frameworks. ViT treats an image as a sequence of patches, similar to how transformers in natural language processing treat text as a sequence of tokens. The DINO framework, on the other hand, is a novel approach to self-supervised learning that does not require labeled data for training. It learns visual representations by encouraging agreement between differently augmented views of the same image. This is achieved through a process known as knowledge distillation, in which a teacher network guides a student network to learn the appropriate feature representations. The combined force of the ViT and DINO frameworks makes it a powerful tool for tasks where labeled training data is scarce or nonexistent (Radford *et al.,* 2018).

2.1.2. SegGPT

SegGPT, short for Segmentation with Generative Pretraining, is a deep learning model designed for object detection and segmentation tasks in images. The model builds on the foundations of Open AI's transformer-based models, such as Generative Pre-training Transformer (GPT), and utilizes a similar architecture with adjustments to suit vision-related tasks (Ramesh *et al.,* 2021).

The architecture of SegGPT employs the standard transformer model, known for its exceptional performance in Natural Language Processing (NLP), and extends its capabilities to handle image data. This is achieved through an autoregressive model that allows SegGPT to generate descriptions of images by sequentially predicting bounding boxes and classes for objects present in the image. One of the key features of SegGPT is its ability to operate under a zero-shot learning paradigm, meaning it can identify and segment objects within images without having previously seen examples of those objects during training (Brown *et al.*, 2020).

2.1.3. SAM

SAM, developed by Meta AI, is a state-of-the-art image segmentation model that can return a valid segmentation mask for any given prompt (Kirillov *et al.*, 2023). In this context, a prompt refers to an indication of what to segment in an image and can take multiple forms, including points indicating the foreground or background, a rough box or mask, text, clicks, and more.

The architecture of SAM is composed of three components: an image encoder, a prompt encoder, and a mask encoder. SAM relies on a massive dataset of over a billion annotated images to support its extensive capabilities, known as the "Segment Anything 1 billion Mask" (SA-1B) dataset. This dataset is currently the largest labeled segmentation dataset available and is specifically designed for the developing and evaluating advanced segmentation models.

2.2. Design Process

This research aims to devise an advanced deep-learning solution for detecting and sizing durian fruits from a variety of images. The objective is approached through a two-phase procedure: object detection, segmentation, and sizing, as shown in Figure 1. This involved the application of two different object detection models— SegGPT and Grounding DINO—and a comparison of their performances. Appropriate thresholds will then be tested and chosen to optimize performance. After selecting the most accurate detection model, the SAM model is used for image segmentation and size calculation. SegGPT can also be used for segmentation, but it is not recommended, as elaborated in the result section later. It's worth noting that the entire methodology operated under a zero-shot learning paradigm, bypassing traditional data collection and manual annotation requirements.

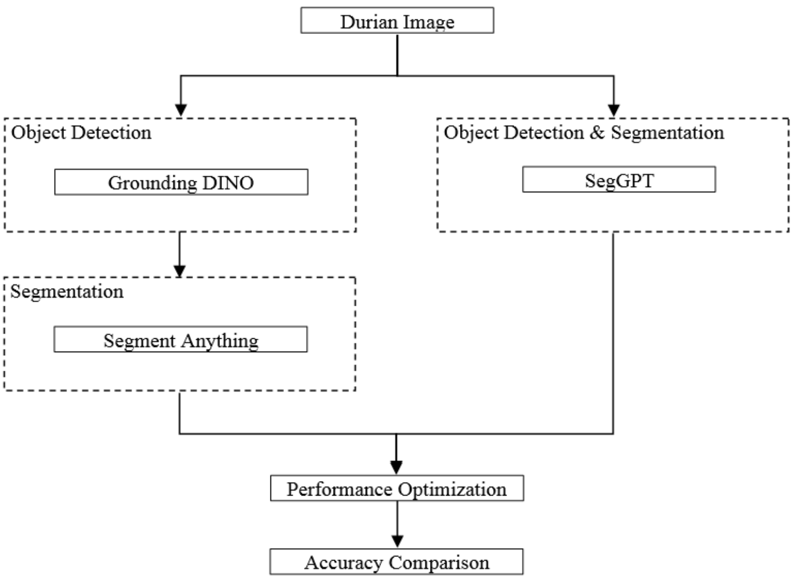


Figure 1 Block diagram of the design process

2.2.1. Phase 1: Object detection

The first phase of the methodology involves applying and comparing two deep learning models designed for object detection tasks: SegGPT and Grounding DINO. Both models have shown effectiveness in object detection and can operate within a zero-shot learning environment, making them ideal candidates for our analysis. They are applied to a set of images, instructing it to identify and mark durian fruits within each image. This is accomplished by having the learning models draw bounding boxes around every detected durian fruit (Carion *et al.,* 2020).

Once both models are applied and have completed their object detection tasks, the next step is the comparative analysis phase. Here, the accuracy of SegGPT and Grounding DINO will be evaluated. Accuracy as shown in Equation 1, is computed as the proportion of durian fruits correctly detected by the model (true positives), t_P relative to the total number of durians present in the images (the sum of true positives and false negatives f_N), $t_P + f_N$.

$$Accuracy = \frac{t_P}{t_P + f_N} \times 100\% \tag{1}$$

2.2.2. Phase 2: Segmentation and size determination

With the object detection phase complete, the next phase is segmentation and size determination. For this phase, the SAM model, an established and potent deep-learning model recognised for its high performance, is recommended for image segmentation tasks. Leveraging the bounding boxes generated by the zero-shot learning model, the SAM model creates accurate segmentation masks for each detected durian fruit. This involved precisely separating the pixels belonging to the fruit from those belonging to the background, thereby providing a detailed and unambiguous representation of each fruit.

Upon obtaining the segmentation masks, the size of each detected durian fruit in pixels is computed (Kirillov *et al.,* 2023). The entire research methodology is conducted within a zero-shot learning framework, eliminating the need for time-consuming and labour-intensive data collection and annotation. By avoiding these traditional steps, it is able to harness the power of advanced deep-learning models efficiently and effectively, demonstrating their potential in real-world agricultural applications.

3. Results and Discussion

The focus of this study is to assess the effectiveness of three cutting-edge deep learning models - Grounding DINO, SegGPT, and SAM, for the task of durian fruit detection and sizing. The experiment process involved zero-shot learning, eliminating the need for extensive data collection or training of these models on durian images specifically. This approach offered an opportunity to test these models' capabilities to generalize their learned knowledge from diverse domains to a new task, i.e. identifying and segmenting durian fruits, as shown in Figure 2.

3.1. Object Detection

The first model tested is Grounding DINO, which is able to ground or localize text in images. In the context of this research, Grounding DINO is used to identify durian fruits based on the textual description, "durian fruit.". Given an input image, the model is prompted with the text "durian," and it returns a list of bounding boxes that it believes contain durian fruits, as shown in Figure 3. The bounding boxes are associated with scores, indicating the model's confidence level for each bounding box.

Figure 2 Original durian image

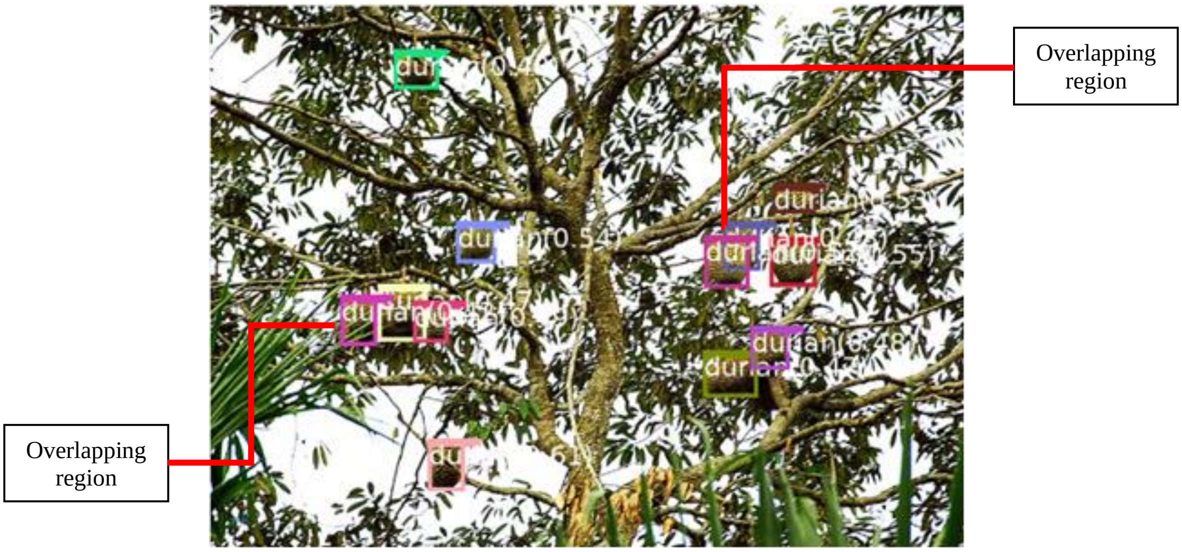
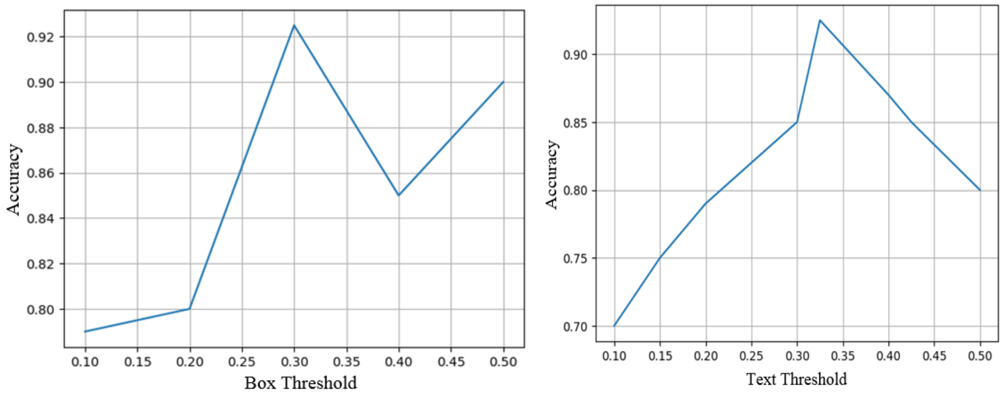


Figure 3 Durian detection using Grounding DINO

To optimize the detection accuracy of both models, we test several different box thresholds and text thresholds, as shown in Figure 4(a) and Figure 4(b). These thresholds are adjusted to identify the sweet spot where the models maximize correct detections (true positives) while minimizing incorrect detections (false positives). After rigorous testing, we conclude that the best results are achieved with a box threshold of 0.3 and a text threshold of 0.325.

Using these thresholds, Grounding DINO is able to achieve an overall accuracy of 92.5% in detecting the durians from the images, as shown in Figure 5. The selection on box threshold and text threshold has been again verified in Figure 5 to give the maximum percentage of accuracy. Nevertheless, one crucial factor to consider in this study is the scenario of overlapping durian fruits. In real-world applications, it is common to find fruits overlapping each other, especially in large-scale plantations or during the storage and transportation process. Our research found that Grounding Dino accurately detected overlapping durian fruits regardless of viewing angles, as shown in Figure 3. The model could effectively discern the individual fruits only when they overlapped each other in a small region. This is a testament to the model's robustness and ability to handle complex image scenarios, a critical feature in practical applications.



] (a) (b)

Figure 4 Performance testing with different (a) box threshold and (b) text threshold

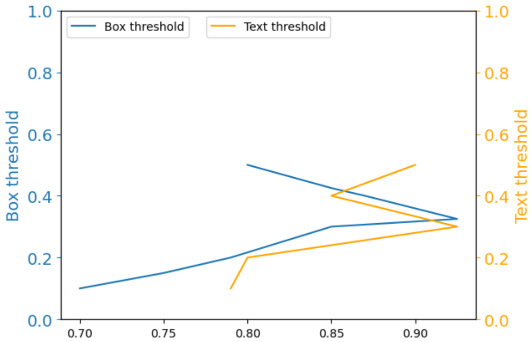


Figure 5 Accuracy performance with different box threshold and text threshold

The second model examined is SegGPT, which generates a binary mask of an image directly from the text. Similar to Grounding DINO, SegGPT is also prompted with the text "durian," and it returns a binary mask where the pixels corresponding to the durian fruits in the image are set to 1, and all other pixels are set to 0, as shown in Figure 6.

Figure 6 Durian detection using SegGPT

3.2. Segmentation and Size Determination

Lastly, the SAM model is employed to segment the identified durian fruits and measure their size in pixels. Using the bounding boxes provided by Grounding DINO, SAM is able to effectively mask each durian fruit, providing a high-quality segmentation mask, as shown in Figure 7. The mask is then used to calculate the size of each durian fruit in terms of the number of pixels. Since the images are digital, each pixel correlates with a specific real-world measurement, allowing for the accurate computation of the fruit size in actual units. The essence of this method lies in the digital nature of the images used. In a digital image, the depicted scene is divided into a grid of tiny squares, each known as a pixel. Therefore, by counting the number of pixels that make up the mask of each fruit, we can estimate its size within the context of the image.

Figure 7 Durian sizing using SAM

However, it is important to mention that this method only provides the size of the fruit in terms of image pixels, which may not directly correlate to real-world measurements. For a direct real-world measurement, depth information is needed. Numerous depth estimation models have been proposed in computer vision in recent years, such as DenseDepth, MiDaS, and MonoDepth (Lu, Xu, and Cao 2021; Alhashim and Wonka, 2018; Godard, Mac-Aodha, and Brostow, 2017). These models predict a depth map where each pixel's value represents its estimated distance from the camera. When combined with our pixel-counting method, these depth maps can enable us to calculate the actual size of each durian fruit in centimeters or any other real-world unit of measurement.

Last but not least, in comparing the above three models, it is evident that Grounding DINO and SAM, when used in combination, provide the most accurate results for durian fruit detection and sizing. Grounding DINO's impressive accuracy in detecting durian fruits and SAM's ability to generate high-quality segmentation masks make them a powerful toolset for the task at hand.

4. Conclusions

This study has highlighted the potential of integrating advanced deep-learning techniques in precision agriculture. By leveraging two cutting-edge models, Grounding DINO, and SAM, we managed to implement a robust methodology for durian fruit size detection. The chosen approach excelled not only in detection accuracy but also in its application efficiency through the use of a zero-shot learning strategy. The results of this paper demonstrate that the integration of advanced deep-learning techniques can revolutionize fruit size detection in precision agriculture. Such advancements pave the way for higher productivity, increased sustainability, and more effective farm management. Although the Grounding DINO shows high accuracy, the overlapping can present unique challenges for this model as it increases the complexity of distinguishing individual fruits.

If the overlapped area becomes too large or the fruits are extensively covered, even Grounding Dino's performance could be compromised. In such cases, acquiring the image from a different angle could help reveal more features of the hidden fruit, thus aiding the model in its detection and segmentation tasks. On the other hand, we see promising potential in integrating depth estimation models with our method in future work. This integration could significantly enhance the accuracy and applicability of durian fruit detection and sizing, moving beyond pixel measurements to providing direct real-world size estimations. While SegGPT showed lower accuracy in detecting durian fruits, future work might also involve exploring other use cases where SegGPT might provide superior results or investigating ways to improve its accuracy for tasks like this. Nevertheless, these models could potentially be used for automated fruit detection with further development and refinement, which includes the actual prediction of the fruit size, the characteristics of the fruit such as ripeness, etc.

Acknowledgments

This research project is fully sponsored by Internal Research Fund (MMUI/220008), Multimedia University.

References

Alhashim, I., Wonka, P., 2018. High-Quality Monocular Depth Estimation Via Transfer Learning. In: *Proceedings of the Asian Conference on Computer Vision*, Springer, Cham, pp. 284–299

Allied Market Research, 2021. Durian Puree Market by Type, Application, and Distribution Channel: Global Opportunity Analysis and Industry Forecast, 2021–2028. Available Online at https://www.alliedmarketresearch.com/durian-puree-market-A10517a, Accessed on March 15, 2023

Belousova, M., Danilina, O., 2021. Game-theoretic Model of the Species and Varietal Composition of Fruit Plantations. *International Journal of Technology*, Volume 12(7), pp. 1498–1507

Brown, T., Mann, B., Ryder, N, Subbiah, M., Kaplan, J,-D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodie, D., 2020. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end Object Detection with Transformers. *In:* European Conference on Computer Vision: 16th European Conference, Glasgow, UK, pp. 213–229

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin., A., 2021. Emerging Properties in Self-supervised Vision Transformers. *In:* Proceedings of The Institute of Electrical and Electronics Engineers (IEEE)/International Conference on Computer Vision (CVF), pp. 9650–9660

Food and Agriculture Data (FAOSTAT), 2021. Food and Agriculture Organization of the United Nations. Available Online at http://www.fao.org/faostat/en/#home, Accessed on March 15, 2023

Godard, C., Mac-Aodha, O., Brostow, G.J., 2017. Unsupervised Monocular Depth Estimation with Left-right Consistency. In: *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE) Conference on Computer Vision and Pattern Recognition*, pp. 270–279

Haziq, M., Pang, W.L., Chan, K.Y., Lee, I.E., Chung, G.C., Wong, S.K., 2022. High-efficiency Low-cost Smart IoT Agriculture Irrigation, Soil's Fertility and Moisture Controlling System. *Universal Journal of Agricultural Research*, Volume 10(6), pp. 785–793

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, Volume 147, pp. 70–90

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.-C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. *arXiv preprint arXiv:2304.02643*

Lu, H., Xu, S., Cao, S., 2021. (Semantic Guided Two-Branch Network) SGTBN: Generating dense depth maps from single-line Light Distance And Ranging (LiDAR). *Institute of Electrical and Electronics Engineers (IEEE) Sensors Journal*, Volume 21(17), pp. 19091–19100

Maninis, K.K., Radosavovic, I., Kokkinos, I., 2019. Attentive Single-tasking of Multiple Tasks. *In:* Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)/International Conference on Computer Vision (CVF) and Pattern Recognition, pp. 1851–1860

Mansour, M.Y.M.A., Dambul, K.D., Choo, K.Y., 2022. Object Detection Algorithms for Ripeness Classification of Oil Palm Fresh Fruit Bunch. *International Journal of Technology*, Volume 13(6), pp. 1326–1335

Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using Deep Learning for Image-based Plant Disease Detection. *Frontiers in Plant Science*, Volume 7, p. 1419

Pedersen, S.M., Lind, K.M. (Eds.)., 2017. Precision Agriculture: Technology and Economic Perspectives. In: *Springer International Publishing*

Pourpanah, F., Moloud, A., Yuxuan, L., Xinlei, Z., Ran, W., Chee-Peng, L., Xi-Zhao, W., Q.M. Jonathan Less, W., 2020. A Review of Generalized Zero-shot Learning Methods. *Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence*, Volume 45(4), pp. 4051-4070

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-training. OpenAI. Available Online at: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, Accessed on April 1, 2023

Rahnemoonfar, M., Sheppard, C., 2017. Deep Count: Fruit Counting Based on Deep Simulated Learning. *Sensors,* Volume 17(4), p. 905

Ramesh, A., Pavlov, M., Goh, Gabriel, Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-Shot Text-To-Image Generation. *In:* International Conference on Machine Learning, pp. 8821–8831

Saowanit, N., 2017. Export Competitiveness of Thai Durian in China Market. *European Journal of Business and Management*, Volume 9(36), pp. 48–55

Tan, H.H., Shahid, R., Mishra, M., Lim, S.L., 2022. Malaysian Vanity License Plate Recognition using convolutional neural network. *International Journal of Technology,* Volume 13(6), pp. 1271–1281

Wang, W., Zheng, V.W., Yu, H., Miao, C., 2019. Semantic Guided Two-Branch Network (SGTBN): A Survey of Zero-shot Learning: Settings, Methods, and Applications*. Association for Computing Machinery (ACM) Transactions on Intelligent Systems and Technology (TIST)*, Volume 10(2), pp. 1–37

Xian, Y., Lampert, C.H., Schiele, B., Akata, Z., 2018. Semantic Guided Two-Branch Network (SGTBN): Zero-shot Learning—a Comprehensive Evaluation of the Good, the Bad and the Ugly, *Institute of Electrical and Electronics Engineers (IEEE) Transactions on Pattern Analysis and Machine Intelligence*, Volume 41(9), pp. 2251–2265