Semiconductor Optoelectronics
ISSN:
1001-5868

FRESH FRUIT BUNCH RIPENESS CLASSIFICATION USING REAL-TIME DETECTION TRANSFORMER

Goh Jin Yu[1], Yusri Md Yunos[1], Usman Ullah Sheikh[1] and Mohamed Sultan Mohamed Ali[1,2]

[1] Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM
Johor Bahru, Johor, Malaysia
[2] Department of Electrical Engineering, College of Engineering, Qatar University, Doha, Qatar.

Abstract— In the field of smart agriculture and automation, the classification of ripeness in oil palm fresh fruit bunches (FFBs) is a critical task with far-reaching implications for productivity. Traditional manual inspections are time-consuming and subjective, prompting the need for automated solutions. This paper introduces the Real-Time Detection Transformer (RT-DETR) as a groundbreaking approach to FFB ripeness classification. RT-DETR is designed to excel in complex agricultural environments, where FFBs are often obscured by lush foliage, occluded, and captured from a distance. Unlike conventional object detection models, RT-DETR leverages the power of the Transformer architecture to capture long-range dependencies effectively. The paper elaborates on the working principle of RT-DETR, highlighting its efficient hybrid encoder, IoU-aware query selection, and decoder with auxiliary prediction heads. These components collectively enable RT-DETR to navigate the challenges of FFB ripeness classification. The study benchmarks RT-DETR against YOLO variants (YOLOv3, YOLOv5, YOLOv6, YOLOv8), demonstrating RT-DETR's superior performance, particularly in mean average precision (mAP) at various Intersection over Union (IoU) thresholds. RT-DETR achieves remarkable mAP values of 0.982 (mAP50) and 0.882 (mAP50-95), significantly outperforming YOLO variants. The exceptional accuracy and real-time capabilities of RT-DETR position it as an ideal choice for precision-demanding tasks in complex settings, especially in FFB ripeness classification.

Index Terms—Object Detection, Fresh Fruit Bunch, Ripeness Classification, Real-Time Detection Transformer and YOLO.

I. INTRODUCTION

The ever-evolving landscape of artificial intelligence has witnessed a surge in the application of advanced deep learning techniques to tackle complex challenges. Within the agricultural domain, a pivotal concern revolves around the classification of ripeness in oil palm FFBs. Accurate and timely identification of FFB ripeness holds profound implications for streamlining harvesting operations, enhancing yield quality, and bolstering agricultural productivity. The common classification process heavily relied on manual inspections conducted by human experts, characterized by their time-intensive nature, subjectivity, and susceptibility to inconsistencies[1]. The emergence of computer vision and deep learning presents an opportunity for transformative change in this field. By harnessing the capabilities of artificial neural networks, it can autonomously analyze FFB images, providing accurate ripeness classification. Real-time object detection models have predominantly focused on optimizing network architectures, anchor-based prediction methods, and efficient feature extraction techniques. However, the challenges stem from the unique characteristics of the palm oil estate context in which these images are captured. FFBs are often situated in lush, natural environments with abundant foliage, branches, and other fruits in close proximity. This results in visually cluttered backgrounds that make distinguishing the fruit bunches from their surroundings challenging. This leads to substantial occlusion, where the boundaries of individual fruit bunches become blurred or obscured, making it difficult to accurately identify and classify each fruit bunch separately. Moreover, the images are taken from a significant distance during field surveys or aerial platforms like drones. This distance results in the fruit bunches appearing relatively small within the images, reducing the amount of visual detail available for analysis. Consequently, assessing ripeness accurately becomes challenging due to limited visual information. There were several comprehensive literature reviews focusing on oil palm FFB ripeness detection methods, with a comparison of approaches for assessing the maturity of these fruit bunches [2–4]. Their findings indicated that computer vision with deep learning-based techniques emerged as the most viable methods for detecting the ripeness of FFBs in field settings. Arkin et al. [5] conducted a survey of object detection methods, specifically focusing on the transition from Convolutional Neural Network (CNN) based approaches to Transformer-based approaches. This paper demonstrated the potential of Transformer-based methods to surpass CNN-based methods in certain object detection tasks, indicating the importance of considering Transformer-based approaches in future research and development in the field of object detection. The hypothesis is further solidified by Samplawski et al. [6] that conducted a benchmarking study of transformer-based object detection models focused on real-time and edge deployment. These models achieve competitive predictive performance without the need for hand-crafted components like non-maximal suppression (NMS) and anchor boxes. In addition to its exceptional performance in object detection tasks, RT-DETR has also garnered recent attention and research focus in diverse applications, including anomaly detection in wireless sensor networks [7], underwater small object detection [8], and marine ship detection [9].

The unique combination of these challenges collectively renders traditional object detection models, which are often designed for more straightforward scenarios with well-defined objects and backgrounds, unsuitable for the classification of FFB. This paper presents innovative solutions and methodologies that can effectively address these complexities in FFB image analysis by utilizing RT-DETR model for the classification. In the following sections, the working principle of RT-DETR will be described, the experiment conducted with the use of synthetic data for training and validating the RT-DETR model will be detailed with comparative analysis.

II. WORKING PRINCIPLE

The RT-DETR, a groundbreaking object detection model developed by Baidu [10], holds immense potential for applications beyond general object detection. This paper demonstrated a novel innovation to utilize the RT-DETR model to perform fruit ripeness classification. The RT-DETR model architecture as illustrated in Fig. 1, with its three main components: an efficient hybrid encoder, an IoU-aware query selection mechanism, and a decoder with auxiliary prediction heads. The model can be tailored to address the challenges of fresh fruit bunch ripeness classification.

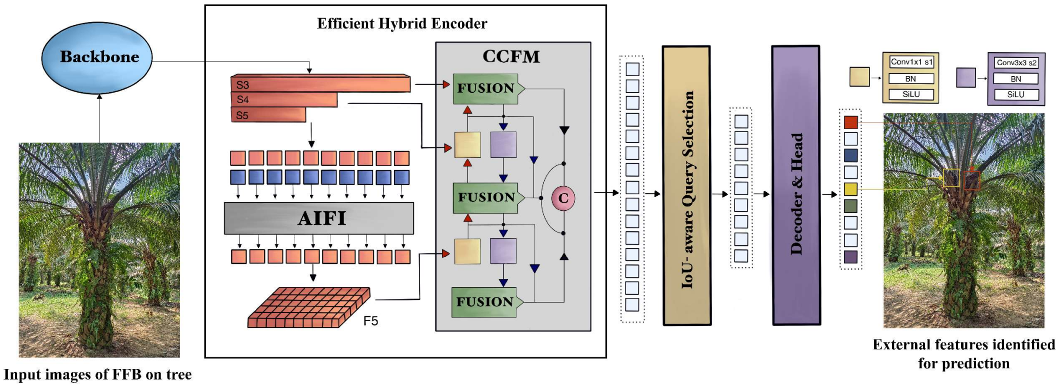

Figure 1. The overall model architecture of RT-DETR [10].

In the context of fresh fruit bunch ripeness classification, the FFB on tree images is feed into the backbone of the model, the backbone used is HGNetv2. The efficient hybrid encoder plays a pivotal role in processing multiscale features from images of fruit bunches. The encoder takes as input the last three stages of the backbone {S3, S4, S5}, allowing it to capture detailed information about the fruit bunches. By decoupling intra-scale interaction and cross-scale fusion through modules like the Attention-based Intrascale Feature Interaction (AIFI) and the Cross-scale Feature-Fusion Module (CCFM), the encoder can help distinguish subtle variations in fruit ripeness, such as color changes and texture differences. This multiscale feature processing is crucial for accurate ripeness classification. Once completed, the process is carried on with the IoU-aware query selection mechanism. By selecting a fixed number of image features as initial object queries, RT-DETR can focus its attention on individual fruit bunches within an image. This level of granularity is essential for precisely identifying and categorizing fruit ripeness. By dynamically adapting the queries based on IoU, RT-DETR can ensure that it concentrates its efforts on the most relevant regions of the fruit bunches, enhancing the accuracy of ripeness assessment. The decoder of RT-DETR takes the initial object queries from the IoU-aware selection mechanism and refines them iteratively. For ripeness assessment, this refinement process can involve identifying specific color patterns, textural changes, and other visual cues indicative of ripeness. The auxiliary prediction heads provide additional supervision during training, allowing RT-DETR to learn and recognize the diverse characteristics associated with varying degrees of fruit ripeness as illustrated in Fig.1.

In the realm of wireless sensor networks, an adaptive transformer model is designed for real-time anomaly detection in dynamic and noisy wireless sensor networks [7]. This model, known as STA-Tran, leverages a dynamic context-capturing deep learning architecture inspired by Transformers, offering a robust solution to the challenge of identifying abnormal data. STA-

Tran's Transformer architecture features an encoder-decoder structure, integrating three essential attention mechanisms: self-attention applied to the input sequence, encoder attention, and decoder attention. The encoder employs a multi-head self-attention mechanism in conjunction with a position-wise fully connected feed-forward network to generate a comprehensive latent representation of the input sequence. Subsequently, the decoder utilizes this latent representation to produce the output sequence, thereby encapsulating crucial contextual information from the input. To further enhance the model's ability to learn effectively from unstructured inputs, Layer Normalization is thoughtfully applied. Additionally, incorporating skip connections enables efficient residual learning within the network. RT-DETR recently gained the attention to being implemented in marine applications such as the detection of marine ships [9] and underwater small objects [8]. This efficiency is derived from its architectural features as illustrated in Fig.1, including multiscale feature processing, IoU-aware query selection, and a refined decoder. These components collectively enable RT-DETR to effectively navigate the challenges posed by underwater settings, such as the ever-shifting dynamics of water, the presence of ghost effects, and the occurrence of multiple detections within a single frame. RT-DETR's adaptability to varying object sizes and appearances in underwater scenes, coupled with its real-time capabilities, makes it adept at capturing and analyzing rapid changes in water features while reducing the impact of reflections and distortions. RT-DETR's working principle, built upon the Transformer framework and deep learning, equips it to excel in underwater environments, facilitating accurate and efficient marine ship and small object detection, ultimately serving critical roles in marine and underwater applications. By harnessing its multiscale feature processing, query selection, and refinement capabilities, RT-DETR can revolutionize the automation of this crucial agricultural task, ensuring consistent and accurate assessments of fruit ripeness across large-scale plantations.

III. RESULT AND DISCUSSION

A. Experiment Setup

This experiment was conducted with the primary objective of introducing and conducting a comprehensive evaluation of RT-DETR in the context of FFB ripeness classification. The experiment aimed to benchmark the accuracy of RT-DETR against state-of-the-art You Only Look Once (YOLO) detectors. The dataset employed was systematically divided into three distinct partitions: 80% for training, 10% for validation, and the remaining 10% for testing, comprising a total of 3000 labeled FFB images. The dataset employed for this paper is a synthetic generated dataset created through domain randomization simulation. In the process of dataset generation, a 3D model of FFB was designed and incorporated into the simulation environment [11]. This 3D model of FFB served as the foreground object to be manipulated during the dataset creation process. The domain randomization approach applied in this simulation encompassed several critical aspects [12,13]. Firstly, it introduced randomized lighting conditions to mimic the transition from morning to evening, ensuring that the dataset included a wide range of lighting scenarios to challenge the object detection model. Additionally, the orientation and position of the FFBs within the images were randomized, contributing to the diversity of object placements. Furthermore, the simulation incorporated randomized background objects and textures, creating a complex visual environment that closely resembled real-world scenarios. This comprehensive approach to domain randomization ensured that the dataset captured a rich spectrum of variations and complexities, effectively challenging the object detection capabilities of the RT-DETR model. The initial dataset structure, originally formatted in the Synthetic Optimised Labeled Objects dataset format, was transformed to align with the prerequisites for training the Convolutional Neural Network (CNN) model, adopting the YOLO dataset format. Hyperparameter encompassing a learning rate set at 0.01, a weight decay coefficient of 0.0005, the utilization of the Adam optimizer, running epoch of 1 and a batch size of 16. These parameters were executed on a INTEL® Core i7-9750H CPU and NVIDIA® 1660Ti GPU, closely adhering to the recommended guidelines provided by Ultralytics®. At the crux of object detection lies the fundamental task of associating bounding boxes with class labels for predictive purposes. To gauge the accuracy of these predictions, our experiment revolved around the calculation of IoU between predicted and ground truth bounding boxes. Precision and recall metrics played a pivotal role in our assessment, contingent upon the imposition of an IoU threshold that delineated the criteria for IoU values. Our evaluation methodology relied on precision-recall curves, characterized by the analysis of the area under the curve, effectively encapsulating the intricate interplay between recall and precision. High precision values indicated a low false positive rate, while high recall values denoted a low false negative rate. Central to our assessment was the computation of the average precision (*AP*) score. In the context of multi-class detection tasks, our scrutiny was extended to the widely recognized mAP score. AP is derived from precision (*p*) and recall (*r*), and mAP is computed by the average is computed of all AP divided by the total number of classes (*Q*), as shown in the following equations (1) and (2):

$$AP = \int_0^1 p(r)dr \qquad (1)$$

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP_q \qquad (2)$$

with $q = 1 \dots Q$ and Q is the number of classes. mAP yields a high value close to 1 when the model demonstrates commendable performance in both recall and precision. On the other hand, the minimum value achievable is zero.

B. Results and comparative analysis

The results of our extensive comparative analysis among YOLO variants (YOLOv3, YOLOv5,

TABLE I. Performance metrics from YOLO variant and RT-DETR detection model.

Detection Model	Metrics	
	mAP50	mAP50-95
YOLOv3	0.263	0.107
YOLOv5	0.219	0.113
YOLOv6	0.447	0.199
YOLOv8	0.602	0.295
RT-DETR	**0.982**	**0.882**

YOLOv6, YOLOv8) and the RT-DETR in the context of object detection reveal a compelling and notable outcome. RT-DETR emerges as the frontrunner, showcasing the highest mean average precision at multiple IoU thresholds, particularly excelling in mAP50 and mAP50-95 scores as tabulated in Table I. In this discussion, the underlying principles and unique contributions that drive RT-DETR's superior performance are delved into. The results of the evaluation provide valuable insights into the performance of different object detection models as shown in Fig. 2, with a particular focus on mAP50 and mAP50-95 metrics. YOLOv3, a well-known model, exhibited a lower accuracy with a mAP50 value of 0.263 and mAP50-95 of 0.107, suggesting limitations in high-precision and robust detection. YOLOv5 and YOLOv6 showed improvements over YOLOv3 but still fell short of RT-DETR's exceptional accuracy, especially in the mAP50-95 range. YOLOv8 demonstrated significant progress in mAP50 with a value of 0.602 but trailed behind RT-DETR in terms of mAP50-95, which had an outstanding value of 0.882, indicating its proficiency in detecting objects with higher IoU thresholds but potential limitations in lower-overlap scenarios. In addition, RT-DETR consistently outperformed all models, excelling in both mAP50 with a value of 0.982 and mAP50-95,

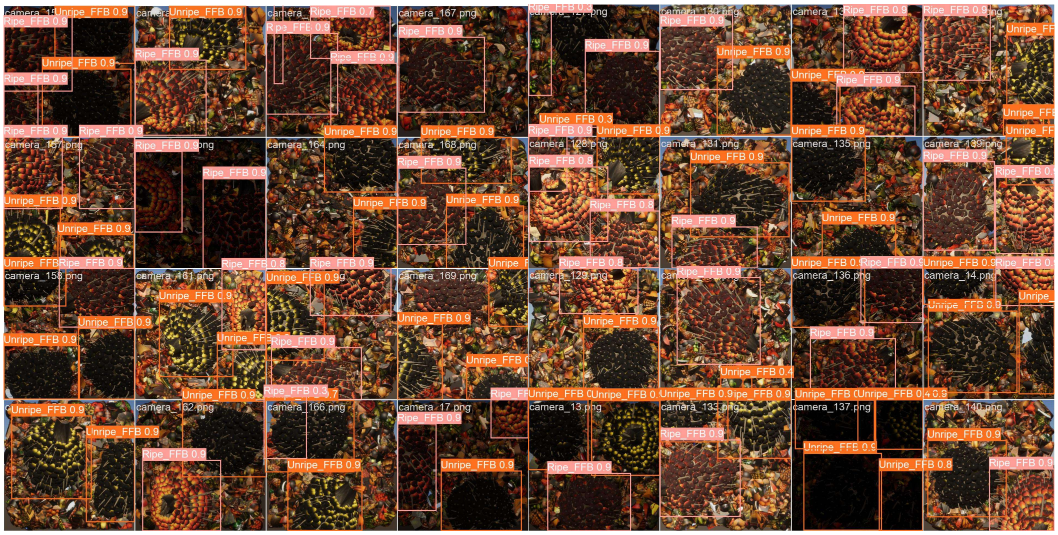

Figure 3. The validation result images of RT-DETR model for FFB ripeness classification. Labeled with ripe and unripe categories of FFB with bounding box.

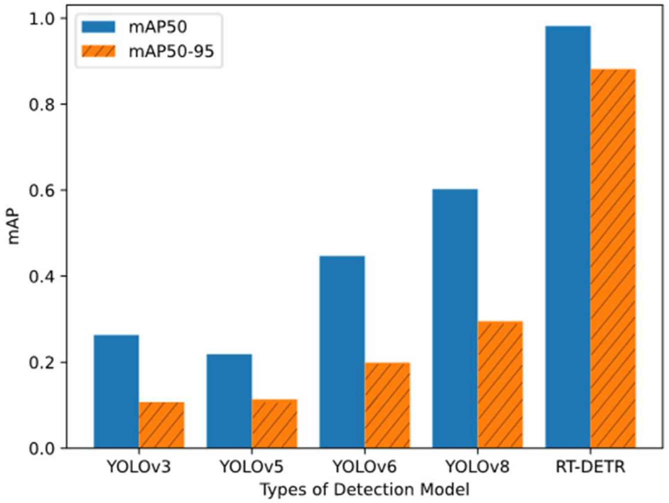

Figure 2. The graph plotting for mAP metrics performance of each detection models.

highlighting its unique ability to deliver precise and reliable object detection across various IoU thresholds. These results emphasize RT-DETR's potential to drive advancements in object detection tasks with critical requirements for precision and robustness. RT-DETR's architecture, tailored for both high accuracy and real-time processing, offers a distinct advantage in complex scenarios like FFB ripeness classification. YOLO models, while known for their real-time capabilities, might grapple with trade-offs between speed and accuracy. In the context of small object detection within cluttered backgrounds, RT-DETR's efficient multiscale feature processing, facilitated by its hybrid encoder, allows it to excel by capturing fine-grained object details across different scales. In contrast, YOLO models may prioritize speed, potentially leading to differences in detection accuracy, especially when confronted with complex backgrounds and small objects. RT-DETR's innovative IoU-aware query selection mechanism proves invaluable when dealing with FFBs in complex settings. By dynamically adapting initial object queries based on IoU criteria, RT-DETR ensures precise focus on relevant regions of an image, overcoming challenges posed by overlapping objects or complex backgrounds as illustrated in Fig. 3, which are common in FFB classification tasks. Traditional YOLO models often rely on fixed anchor boxes, which may not offer the same adaptability in query selection when addressing complex and obstructed object scenarios. In addition, where subtle variations in color and texture hold significance, RT-DETR's decoder, supported by auxiliary prediction heads, excels in refining object queries iteratively. This capacity enhances its ability to interpret intricate visual attributes. YOLO architectures also employ decoding mechanisms, but the extent of refinement and adaptability may differ, potentially affecting detection accuracy, especially when dealing with objects amid complex backgrounds and obstacles. Moreover, RT-DETR's versatility and adaptability extend its applicability to various domains, including agriculture, where it has demonstrated remarkable performance in FFB ripeness classification, a task inherently characterized by small object detection amidst complex backgrounds and obstacles. The model's adaptability to complex environments, real-time capabilities, and high accuracy make it well-suited for scenarios where precision is paramount, even in challenging settings.

IV. CONCLUSION AND FUTURE WORK

This experiment conducted a comprehensive evaluation of the RT-DETR against various YOLO architectures, including YOLOv3, YOLOv5, YOLOv6, and YOLOv8, in diverse object detection scenarios for FFB ripeness classification. The results underscored RT-DETR's remarkable performance, marked by its ability to achieve high accuracy while maintaining real-time processing capabilities. RT-DETR's success can be attributed to its innovative architectural elements, including efficient multiscale feature processing, IoU-aware query selection, and a refined decoding mechanism. These features equip RT-DETR with a unique balance of precision and speed, making it a compelling choice for a wide range of object detection tasks, including those involving small object detection amidst complex backgrounds and obstacles, as evidenced by its excellent performance in fresh fruit bunch ripeness classification.

Moving forward, there are several promising directions for future research and development in the realm of object detection. Real-world deployment of object detection models, including RT-DETR, demands thorough investigation. Conducting field trials and assessing the performance of these models under diverse conditions will be crucial. Additionally, addressing the specific challenges associated with real-world implementation, such as data collection and hardware integration, will be a focus area. Interdisciplinary applications present exciting opportunities. Furthermore, exploring how object detection models like RT-DETR can be adapted to domains like healthcare, agriculture, and environmental monitoring will unlock novel use cases and contribute to broader societal benefits. These avenues of exploration will drive innovation and progress in the field, leading to more reliable, adaptable, and responsible object detection solutions.

ACKNOWLEDGMENT

This work was supported by Ministry of Higher Education under Fundamental Research Grant Scheme FRGS/1/2022/TK07/UTM/02/42 and Universiti Teknologi Malaysia under UTM Fundamental Research Grant Q.J130000.3823.22H55.

REFERENCES

[1] M.K. Shabdin, A.R.M. Shariff, M.N.A. Johari, N.K. Saat, Z. Abbas, A study on the oil palm fresh fruit bunch (FFB) ripeness detection by using Hue, Saturation and Intensity (HSI) approach, in: IOP Conf Ser Earth Environ Sci, Institute of Physics Publishing, 2016. https://doi.org/10.1088/1755-1315/37/1/012039.

[2] X.J. Tan, W.L. Cheor, K.S. Yeo, W.Z. Leow, Expert systems in oil palm precision agriculture: A decade systematic review, Journal of King Saud University - Computer and Information Sciences. 34 (2022) 1569–1594. https://doi.org/10.1016/j.jksuci.2022.02.006.

[3] K.Y. You, F.H. Wee, Y.S. Lee, Z. Abbas, K.Y. Lee, E.M. Cheng, C.S. Khe, M.F. Jamlos, A Review of Oil Palm Fruit Ripeness Monitoring Using Microwave Techniques in Malaysia, IOP Conf Ser Mater Sci Eng. 767 (2020). https://doi.org/10.1088/1757-899X/767/1/012007.

[4] M.Y. Mohamed Ahmed Mansour, A Review of Non-Destructive Ripeness Classification Techniques for Oil Palm Fresh Fruit Bunches, J Oil Palm Res. (2022).

https://doi.org/10.21894/jopr.2022.0063.

[5] E. Arkin, N. Yadikar, X. Xu, A. Aysa, K. Ubul, A survey: object detection methods from CNN to transformer, Multimed Tools Appl. 82 (2023) 21353–21383. https://doi.org/10.1007/s11042-022-13801-3.

[6] C. Samplawski, B.M. Marlin, Towards Transformer-Based Real-Time Object Detection at the Edge: A Benchmarking Study, in: Proceedings - IEEE Military Communications Conference MILCOM, Institute of Electrical and Electronics Engineers Inc., 2021: pp. 898–903. https://doi.org/10.1109/MILCOM52596.2021.9653052.

[7] A. Siva Kumar, S. Raja, N. Pritha, H. Raviraj, R. Babitha Lincy, J. Jency Rubia, An adaptive transformer model for anomaly detection in wireless sensor networks in real-time, Measurement: Sensors. 25 (2023). https://doi.org/10.1016/j.measen.2022.100625.

[8] G. Chen, Z. Mao, K. Wang, J. Shen, HTDet: A Hybrid Transformer-Based Approach for Underwater Small Object Detection, Remote Sens (Basel). 15 (2023). https://doi.org/10.3390/rs15041076.

[9] Z. Xing, J. Ren, X. Fan, Y. Zhang, S-DETR: A Transformer Model for Real-Time Detection of Marine Ships, J Mar Sci Eng. 11 (2023). https://doi.org/10.3390/jmse11040696.

[10] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang, Y. Liu, DETRs Beat YOLOs on Real-time Object Detection, (2023). http://arxiv.org/abs/2304.08069.

[11] S. Borkman, A. Crespi, S. Dhakad, S. Ganguly, J. Hogins, Y.-C. Jhang, M. Kamalzadeh, B. Li, S. Leal, P. Parisi, C. Romero, W. Smith, A. Thaman, S. Warren, N. Yadav, Unity Perception: Generate Synthetic Data for Computer Vision, (2021). http://arxiv.org/abs/2107.04259.

[12] J. Borrego, A. Dehban, R. Figueiredo, P. Moreno, A. Bernardino, J. Santos-Victor, Applying Domain Randomization to Synthetic Data for Object Category Detection, (2018). http://arxiv.org/abs/1807.09834.

[13] O. Maqbool, J. Roßmann, Formal Scenario-driven Logical Spaces for Randomized Synthetic Data Generation, in: Scitepress, 2022: pp. 203–210. https://doi.org/10.5220/0010816400003119.