Pallet and Pocket Detection Based on Deep Learning Techniques

Daniele Caldana[1,2], Artur Cordeiro[1,3], João Pedro Souza[1], Ricardo B. Sousa[1,2], Paulo M. Rebelo[1,3], António Joaquim Silva[4], and Manuel F. Silva[1,5]

[1]CRIIS – Centre for Robotics in Industry and Intelligent Systems, INESC TEC – Institute for Systems and Computer Engineering, Technology and Science. Portugal. {daniele.caldana,artur.j.cordeiro,joao.p.souza,paulo.m.rebelo}@inesctec.pt

[2]Faculty of Engineering, University of Porto (FEUP). Portugal. up201503004@edu.fe.up.pt

[3]School of Science and Technology, University of Trás-os-Montes e Alto Douro (UTAD). Portugal.

[4]Setúbal Industrial Complex, The Navigator Company. Portugal. antonio.silva@thenavigatorcompany.com

[5]ISEP, Polytechnic of Porto. Portugal. mss@isep.ipp.pt

***Abstract*—The high level of precision and consistency required for pallet detection in industrial environments and logistics tasks is a critical challenge that has been the subject of extensive research. This paper proposes a system for detecting pallets and its pockets using the You Only Look Once (YOLO) v8 Open Neural Network Exchange (ONNX) model, followed by the segmentation of the pallet surface. On the basis of the system a pipeline built on the ROS Action Server whose structure promotes modularity and ease of implementation of heuristics. Additionally, is presented a comparison between the YOLOv5 and YOLOv8 models in the detection task, trained with a customised dataset from a factory environment. The results demonstrate that the pipeline can consistently perform pallet and pocket detection, even when tested in the laboratory and with successive 3D pallet segmentation. When comparing the models, YOLOv8 achieved higher average metric values, with YOLOv8m providing better detection performance in the laboratory setting.**

***Index Terms*—pallet detection, Deep Neural Networks (DNN), RGB-D camera, object recognition**

I. Introduction

The intralogistics within warehouses, distribution centres, and manufacturing facilities demands planning and controlling the internal flow of materials and delivery processes. Traditionally, these operations have relied heavily on human labour, such as inventory management and pallet manual handling. However, the growing need for enhanced efficiency, safer environments, and costs reduction is driving the adoption of autonomous technologies. Autonomous forklifts are a key element in this adoption, by allowing the automation of handling pallets in intralogistics operations.

To achieve a fully autonomous forklift, the system must detect and identify pallets reliably within its operational environment, independently of the lighting conditions, dynamics objects and unforeseen obstacles in the scene. Several works propose the use of 2D laser scanners to detect pallets [1–3]. However, this sensor can only scan in a 2D plane while only retrieving range data from the environment. An alternative is the use of cameras. The current literature explores the usage of conventional computer vision to detect pallets, such as colour-based image segmentation [4–6] and fiducial markers [5, 7–10], exploiting the depth channel information from RGB-D cameras with point cloud-based filtering and segmentation [11, 12] and template matching to the perceived point cloud [13, 14]. With the advancements on the application of Neural Networks (NNs) to recognise objects in the scene, several works propose the usage of NN-based detectors to detect pallets [10, 15–23]. Overall, most existent works focus on recognising the pallet and possibly estimating its pose, not on detecting the pallet pockets. The accurate detection of the pallet pockets allows the estimation of where the forklift's forks should be inserted, allowing compatibility with different types of pallets and, thus, enhancing operational efficiency.

Given the above described ideas, this work proposes a pipeline architecture to recognise the pallet and its pockets in the RGB image and segment the pallet surface from the depth point cloud of an RGB-D camera. The pallet recognition model is the YOLOv8 from the YOLO [24] family. To the best of the authors knowledge, this model is not yet evaluated in the literature for the pallet detection task. Furthermore, all point cloud operations are performed using the Point Cloud Library (PCL). Experiments are performed over a dataset that includes pallets in various configurations, as well as partially obstructed by plastic film, in an industrial environment. The outcomes of the pallet and pocket detection procedures are also evaluated in comparison to the YOLOv5 model. Moreover, the experiments conducted in the laboratory with the YOLOv8 demonstrate that both the pallet and pocket are accurately detected, thus providing a solid foundation for future pocket pose estimation.

The remainder of this paper is organised as follows. Section II reviews the current advancements and developments in the field of pallet detection in industrial environments. Next, Section III outlines the methodology and the steps taken during the development process. Section IV presents the results obtained and provides an analysis of these findings. Finally, Section V concludes the work and discusses future directions.

This work is co-financed by Component 5 – Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021–2026, within project PRODUTECH R3, with reference 60.

2024 7th Iberian Robotics Conference (ROBOT) | 979-8-3503-7636-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ROBOT61475.2024.10797410

II. Related Work

This section describes approaches utilising diverse methodologies and technologies for the task of pallet detection.

A. 2D laser scanner

The presence of 2D laser scanners in Autonomous Mobile Robots (AMRs) is usually related to their autonomous navigation and safety capabilities. Nevertheless, these sensors have been employed in the literature for pallet detection. Bellomo *et. al.* [1] propose a methodology to detect pallets by matching the scan from a 2D laser with a template, using the Iterative Closest Point (ICP) as the matching criteria. The method combines genetic algorithms with direct search to be more efficient in local search. However, the experimental results show mathematical ambiguities when two pallets are close to each other. Instead of only scanning on a single height, Walter *et. al.* [2] actively actuate the forklift's mast to collect 2D laser scans at different heights, requiring user input to provide the volume of interest. Then, a classification algorithm determines if a set of detections is consistent with typical pallets in terms of height.

Another way to interpret range data is by converting it into 2D bitmap images. Given the compatibility of the bitmap format with Convolutional Neural Networks (CNNs), Mohamed *et. al.* [3] use the Faster R-CNN [25] to detect the Region of Interest (ROI) within the pallet range points as a bounding box. A Kalman filter-based tracker matches the potential pallet detections over time, estimating the pallet by its centroid. Overall, the laser-based methods are limited by the sensor characteristics of scanning a single 2D plane.

B. Vision-based systems

Intralogistics operations can have pallets made from different materials (e.g., wood, plastic, and metal) and based on various standard sizes, including the Euro pallet (1200×800 mm) and the North American size (40×48 in). Thus, vision-based systems may also be employed for pallet detection, considering the recent advancements in computer vision and that these systems can get more information other than range data.

In the literature, several works investigate using conventional computer vision techniques for pallet detection in RGB images. Byun and Kim [4], Chen *et. al.* [5], and Cui *et. al.* [6] use the colour to segment the pallet. Byun and Kim [4] consider the RGB space to segment and extract the pallet pockets, the latter using prior information on its geometry. In contrast, Chen *et. al.* [5] segment in the Hue, Saturation, Value (HSV) space due to being more invariant to lighting changes. As for Cui *et. al.* [6], the dominant colours are extracted from the image iteratively and compared to the pallet colour for rejection and clustering. Both [5] and [6] employ opening and closing morphological operators to remove noise, while [6] also apply the Sobel operator to retrieve the edges and retain the closest points to the ground. Although [4], [5], and [6] are able to estimate the pallet pose using features from the segmented image ([4],[6]) or circle-based fiducial markers ([5]), these works rely on the colour of the pallet.

Other works in the literature focus on using RGB-D cameras with conventional computer vision, with the additional depth channel to retrieve 3D data from the scene. Both Molter and Fottner [11] and Zhao *et. al.* [12] segment the three pillars of the pallet to estimate its pose. These two works do not use any RGB information, instead employing a voxel filtering to down-sample the depth image. Work [11] implements geometric pallet detection focused on Euro pallets, where the point cloud is segmented with the region growing algorithm from the PCL. Method [12] employs the PCL statistical filtering and Euclidean clustering algorithms to retrieve the 3 pillars, based on the assumption that the distance between centre and outer pillars is the same.

Furthermore, the ICP matching criteria may also be employed to match the depth point cloud from the scene to a template model. Wang *et. al.* [13] match a known pallet model with nine pillars with the point cloud obtained from a structured light sensor to obtain the pallet pose. Shao *et. al.* [14] remove walls and ground points using RANdom SAmple Consensus (RANSAC). Feature extraction is performed by extracting Intrinsic Shape Signatures (ISS) key points and estimating their descriptors using the Adaptive Color Fast Point Feature Histogram (ACFPFH). Next, the surface matching step matches the feature descriptors between the pallet template point cloud and scene point cloud. Then, RANSAC and ICP are used for coarse and fine point cloud registration, respectively. Still, both methods [13] and [14] require prior knowledge of the template pallet model.

Another trend found in the literature with conventional computer vision methodologies is the usage of fiducial markers to estimate the pallet pose with RGB cameras. Seelinger and Yoder [7] and Chen *et. al.* [5] put circle-based binary fiducial markers in each of the three pillars on the pallet's front side, where the centre pillar is used for translation and the other two to retrieve orientation. Additionally, Seelinger and Yoder [7] add markers to each fork to retrieve the homogeneous transformation of the camera w.r.t. the internal configuration of the forklift's degrees of freedom. Although Aref *et. al.* [8] employ three markers for each pallet's pillar, the fiducial marker is from the ALVAR virtual and augmented reality library. Moreover, Aswad *et. al.* [9] and Kesuma *et. al.* [10] use ArUco markers to extract the pose of the pallet. In contrast to the other fiducial markers-based methods, [9] and [10] only use a single marker centred on the pallet. Overall, fiducial markers-based pallet detection methods require the calibration of the camera's intrinsic parameters to estimate the markers' pose and bring an added cost to the final solution of putting markers in all pallets and ensuring their integrity over time.

C. Neural Networks (NN) on Pallet Detection

More recently, several works have explored the application of NNs to leverage the recent advancements in artificial intelligence. Zaccaria *et. al.* [15] propose the usage of a CNN to detect the ROI of both the pallet front side and its pockets in an RGB image. The authors compare YOLOv4 [24], Faster R-CNN [25], and Single Shot Detector (SSD) [26], where YOLOv4 obtains better detection at greater distances, while Faster R-CNN and SSD achieve higher precision results. Manurung *et. al.* [16] and Caldana *et. al.* [17] also use YOLO models (YOLOv5 and YOLOv3, respectively) to detect the pallet's front side. Additionally, Caldana *et. al.* [17] estimate the pallet's 3D centre point based on the segmented point cloud's centroid from a Time-of-Flight (ToF) camera and the pallet's orientation angle w.r.t. the camera frame, relaxing the pose estimation problem from six to four degrees of freedom. Tsiogas *et. al.* [18] employ the YOLOv5 to detect only the pallet legs, segment them in the depth point cloud, and extract the colinear centroids of the legs to match them to a template pallet pattern. In contrast, Bohacs *et. al.* [19] and Kesuma *et. al.* [10] use the YOLO9000 and YOLOv5, respectively, to recognise solely the pallet pockets in the RGB image without estimating its pose. Also, Li *et. al.* [20] employ the SSD to detect the ROI from the pallet front view in the RGB image, and then segment the front plane from the depth channel. Key points from the intersection of line features extracted from the pallet's front surface are used to estimate the pallet pose.

Other NN-based models than YOLO [24], SSD [26], and Faster R-CNN [25] are used in the literature to detect ROI for pallet detection. Iinuma *et. al.* [21] adopt the EPSNet performing pixel-wise semantic segmentation to detect the front view of a metal pallet (including its cargo), the top of the pallet's cargo, and the background. Then, this semantic information is used to segment the pallet's front surface from the depth point cloud and estimate the surface's position and orientation with RANSAC. Jia *et. al.* [22] employ the AlexNet to extract the ROI of the pallet's front side over the depth data obtained from a ToF sensor. The authors use the Sobel operator on the depth image to extract the front surface, and then, estimate the pallet's centroid. In contrast to other NN-based methods, Wang *et. al.* [23] leverage both RGB and depth channels with the NanoDet Plus network. The model is trained to detect the tray volume space where all the pallet points lay, from which the pallet pose is estimated based on the centroid of the three pillars. In general, most NN-based methodologies focus only on detecting the pallet and estimating its pose, not on recognising and estimating the pockets pose.

D. CNNs architectures

CNNs typically follow two main approaches: two-stage and single-stage models. A well-known example of a two-stage model is Faster R-CNN [27]. In the first stage, a set of bounding boxes is generated, which are likely to contain objects. In the second stage, these boxes are refined and classified to determine the precise location of the objects. This approach generally requires more processing time, as there can be hundreds of proposed regions. However, it achieves higher accuracy. Single-stage models, such as SSD [26] and detectors from YOLO family [24], perform object detection and classification in a single step, significantly reducing processing time. Regarding the models compared in this paper, YOLOv5 uses a modified version of CSPDarknet53 and incorporated several augmentation techniques like Mosaic, Affine, MixUp, and integration with the Albumentations library. A significant advancement in YOLOv5 was the automation of anchor boxes, optimising the selection of anchor boxes based on the dataset, thus improving efficiency. In relation to YOLOv8, it brings substantial advancements in feature extraction and object detection performance. It adopted an anchor-free design, simplifying the model and reducing complexity. Both the YOLOv5 and YOLOv8 models have been released by Ultralytics, and the YOLOv8 is relatively new. However, although the YOLOv5 has been used and compared in some studies [16][10], no comparison has been found between these two models.

III. Proposed Pallet Detection System

The system overview is presented in Figure 1. However, this paper only addresses the blue shaded area, which encompasses 2D localisation followed by 3D segmentation. The entire system is fed by an RGB-D camera. Object detection is achieved using bounding boxes, employing the YOLOv8 model. The 3D segmentation step consists of segmenting the ROI in the point cloud, comprising the face of the pallet. Finally, the pose estimation of the centre of the pallet pockets is regarded as a future development.

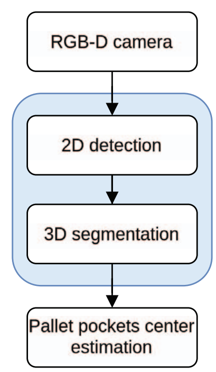

Fig. 1: System overview.

The system is supported by a pipeline which is the continuation of the work presented by Cordeiro *et. al.* [28], although the base structure in pipeline format was introduced by Souza *et. al.* [29]. This pipeline employs a ROS Action Server approach and the system has two modules in the scope of this work: the first handles 2D localisation, and the second focuses on 3D segmentation. A key feature of this system is its flexibility in both the 2D and 3D modules, allowing heuristics to be stacked, in a software pipeline fashion, and enabling the implementation of custom heuristics. This capability facilitates testing various configurations and heuristics to determine the most effective approach. Within the framework of this work, a modification was made to the 2D module. A new heuristic was introduced that employs ONNX Runtime to run inference using a YOLOv8 model converted to the ONNX format. The ONNX format is a standard for representing artificial intelligence models and provides optimisations that can reduce

Fig. 2: Example of dataset images.

inference times [10]. Additionally, all implementations are designed to maximise hardware utilisation. Regarding the point cloud module, no significant changes were made.

The system is developed in two phases. The first consists of preparing the detector to be integrated into the pipeline. The second phase involves launching the pipeline that will carry out the inference and segmentation of the point cloud.

A. Model preparation

This step involves preparing the dataset, training the model, and converting it to the ONNX format. The neural network is trained on a custom dataset that was manually annotated using the Computer Vision Annotation Tool (CVAT). The dataset includes two labels: *pallet* and *pocket*. Apart from locating the face of the pallet, the goal is also to assess the detector's reliability in identifying pallet pockets, which could facilitate the pose estimation of each pocket centre in future implementations of the pipeline. To effectively leverage the detector, images were included with different angles and captured from different cameras and distances. Consistency in annotations was maintained, ensuring that the bounding box fully contains the object and there are no gaps between the object and the annotation. Lastly, the dataset included images where no objects are present to help the model reduce false positives. Images were captured using a Canon R5 camera and a Intel RealSense L515 camera, resulting in a total of 454 images, and then downgraded to a resolution of (640×480) pixels as shown in Figure 2. Furthermore, this is the resolution set-up for training and used in the pipeline.

While the dataset is relatively small at this stage, the stable lighting conditions in the operating area limited the variety of images. To address this limitation, the dataset can be enhanced through the application of data augmentation techniques or through the incorporation of additional datasets. Therefore, it was created a second version of the dataset by applying augmentation techniques, like image rotation, vertical and horizontal flips, random brightness adjustments, and sepia colour transformations, using the Albumentations library [30]. However, it is important to note that increasing the number of images may not always lead to improved detector performance beyond a certain point. Furthermore, indiscriminate augmentation can dilute the relevance of the data, potentially diminishing its impact on training outcomes. It is possible that, in the future, a new dataset may be required for the application to be used in different environments. Consequently, the model would need to be retrained on the new dataset to ensure effective detection across diverse settings.

B. Pipeline execution

Figure 3 shows the pipeline diagram used in this work. Each module includes specific heuristics applied throughout the process. The model, trained and converted to ONNX as outlined in Section III-A, detects pallets and pockets in the camera image. These detection results are passed to the next module, which receives an organised point cloud with a resolution (640×480) equal to the camera image. The first rule extracts regions within the point cloud corresponding to each bounding box. The following step filters the point cloud segments according to the provided class labels, retaining only the pallet class. With this completed, a selection criterion picks the segment with the highest confidence score, resulting in a point cloud containing only pallet points. A voxel grid is then applied to reduce sample density and standardise the point cloud. Next, multiple planes are segmented using RANSAC, with up to five surfaces isolated. Subsequently, a selection metric is used to choose the optimal plane based on a point cloud area measure. Finally, an outlier removal filter is employed to clean any remaining noise from the segment.

IV. Tests and Results

This section presents the training outcomes of YOLOv5 and YOLOv8 models, in addition to the results of the implemented pipeline.

A. Training the detector

Both pre-trained YOLOv5 and YOLOv8 models from the COCO dataset were used. Moreover, the pallet custom dataset was split into 85 % for training and 15 % for validation. Table I shows the distribution of objects for training and validating the model for each class on the dataset with data augmentation. The models were trained for 200 epochs and all training sessions were conducted using an Nvidia GeForce RTX 4090 GPU. Furthermore, the results were evaluated using COCO metrics.

TABLE I: Number of objects for each class with augmentation.

	Training set	Validation set
Pallet side	1978	257
Pocket	3964	513
Total	5492	770

Table II presents the model validation results, where *P* stands for precision, *R* stands for recall, *mAP50* refers to the mean average precision at an Intersection over Union (IoU) threshold of 50 %, and *mAP50-95* represents the mean average precision across IoU thresholds ranging from 50 % to 95 %. Additionally, it should be noted that the class *all* represents the average of both the *pallet* and *pocket* classes.

The cells highlighted in green indicate the size with the highest mean value for the class *all* for each metric within each model. Considering the results of the YOLOv5 model,

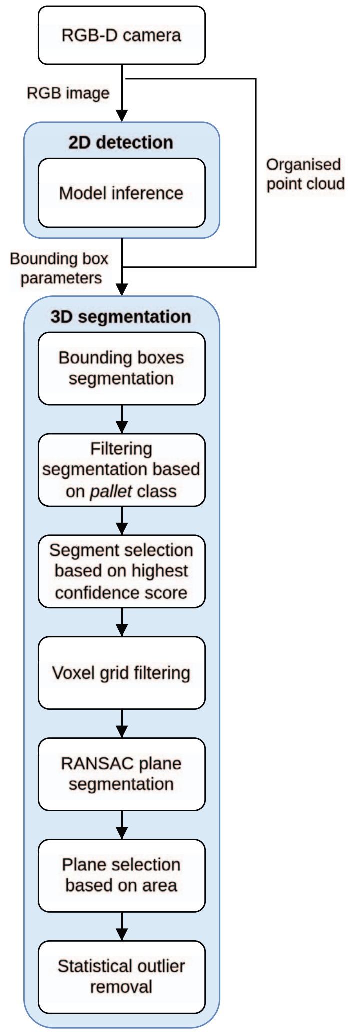

Fig. 3: Pipeline diagram.

TABLE II: Metrics for the detection of pallet and pocket classes. The class *All* represents the mean between pallet and pocket classes. The green highlighting indicates the model size where the mean is higher compared to other sizes within the same model.

Model	Size	Class	P	R	mAP50	mAP50-95
YOLOv5	x	*pallet*	0.931	0.969	0.976	0.857
		pocket	0.954	0.961	0.979	0.764
		all	0.942	0.965	0.978	0.810
	l	*pallet*	0.924	0.961	0.973	0.851
		pocket	0.963	0.961	0.982	0.766
		all	0.943	0.961	0.978	0.808
	m	*pallet*	0.929	0.977	0.981	0.856
		pocket	0.957	0.953	0.98	0.758
		all	0.943	0.965	0.980	0.807
	s	*pallet*	0.932	0.961	0.974	0.833
		pocket	0.957	0.950	0.977	0.755
		all	0.944	0.955	0.975	0.794
	n	*pallet*	0.928	0.947	0.980	0.805
		pocket	0.967	0.919	0.979	0.731
		all	0.947	0.933	0.979	0.768
YOLOv8	x	*pallet*	0.943	0.969	0.979	0.866
		pocket	0.970	0.946	0.989	0.793
		all	0.957	0.957	0.984	0.830
	l	*pallet*	0.932	0.965	0.974	0.869
		pocket	0.947	0.972	0.990	0.789
		all	0.940	0.968	0.982	0.829
	m	*pallet*	0.936	0.969	0.985	0.887
		pocket	0.965	0.955	0.989	0.787
		all	0.950	0.962	0.987	0.837
	s	*pallet*	0.939	0.967	0.982	0.893
		pocket	0.972	0.963	0.990	0.791
		all	0.956	0.965	0.986	0.842
	n	*pallet*	0.942	0.952	0.968	0.865
		pocket	0.978	0.946	0.991	0.789
		all	0.960	0.949	0.979	0.827

YOLOv5n achieved the highest P with a mean of 0.947, while the highest R was recorded for both YOLOv5x and YOLOv5m, at 0.965. Additionally, YOLOv5m excelled in mAP50, attaining a mean of 0.980, whereas YOLOv5x stood out in mAP50-95, achieving a value of 0.810. In contrast, no single size of YOLOv8 consistently outperformed the others across multiple metrics. However, regarding each of the metrics, YOLOv8 values surpassed those of YOLOv5, with improvements ranging from 0.03 to 0.13. As a result, the YOLOv8m was the model selected that successfully detected both the pallet and its pockets in the laboratory setting.

It is also worth noting that experiments were conducted with varying parameters, such as learning rate and epochs, but no improvements were observed during training. Figure 4 shows several images with the detections of pallets and pockets in the industrial environment where the dataset was collected using the trained YOLOv8m model.

B. Pipeline outputs

Tests were carried out in the laboratory, and Figure 5 shows detections using the pipeline for image segmentation with the YOLOv8m model. Although the dataset used for training only contained images from the industrial environment, data augmentation allowed the model to generalise to other completely different environment. The model is also able to detect pockets. However, at distances greater than 4 m, pocket detection becomes less consistent and there are more false negatives, while pallet detection remains successful, suggesting that the model should be trained with more images where the pallets are at greater distances.

Considering point cloud segmentation, Figure 6 illustrates the detections in RGB image and the front of the pallet segmented. The L515 camera controls, such as digital gain, laser power, confidence threshold, receiver gain and post processing sharpening, were adjusted to achieve the best quality of the point cloud. The pocket information is not being used, as in this paper is not proposed pose estimation method for the pockets. However, the results are promising and will serve as the foundation for the next phase of the project.

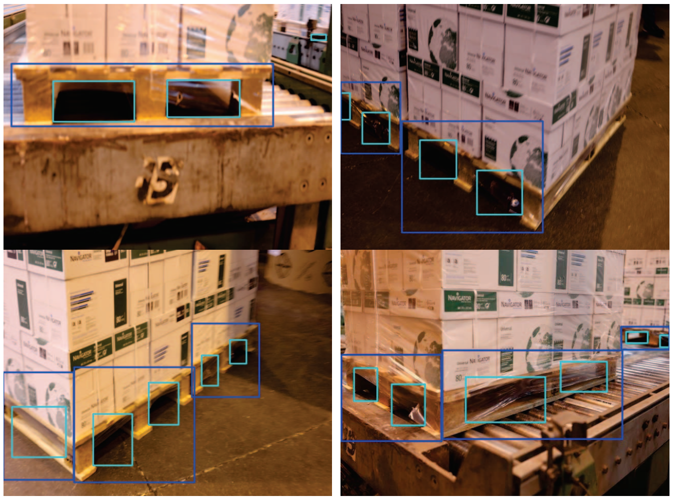

Fig. 4: Detections on industrial environment using YOLOv8m model with Ultralytics framework. Dark blue bounding boxes are pallets, and light blue are the pockets.

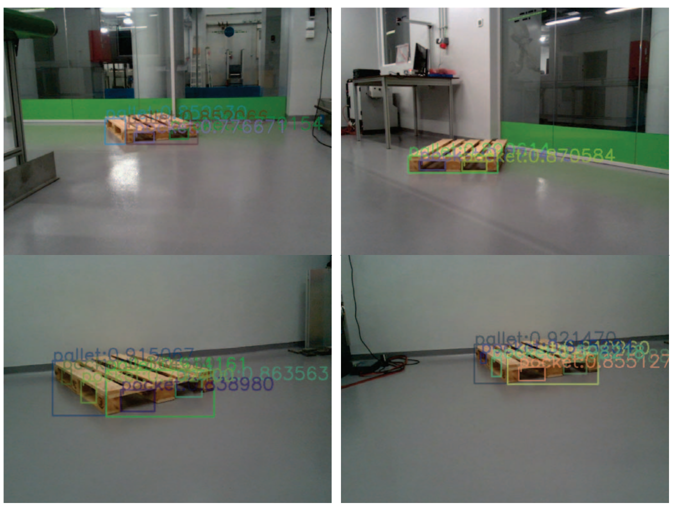

Fig. 5: Detections on laboratory with YOLOv8m model converted to ONNX. The pallets and pockets are labeled with different colours.

V. Conclusions and Future Work

In conclusion, this paper presents the pallet and pocket detection system implemented using the YOLOv8 model converted to the ONNX, followed by 3D segmentation of the pallet surface. The results demonstrate that, although the model was trained on a dataset from an industrial environment, it effectively generalised to detect pallets and pockets in a laboratory setting. Moreover, the modular structure of the pipeline allows for high adaptability of heuristics and facilitates various testing scenarios. The pallet segmentation results are satisfactory and consistent, providing a solid foundation for future implementations. In comparing the YOLOv5 and YOLOv8 models, YOLOv8m showed a slightly superior detection performance in the laboratory setting, with higher average metric results on the validation dataset. Future work will focus on developing a strategy to estimate the pose of the centre of pallet pockets. The robustness and diversity of the detector will also be improved by incorporating images from different environments and exploring the creation of synthetic data. Other detectors, including the latest YOLOv10 model, will be tested using the proposed methodology to evaluate potential advantages for the localisation task.

VI. Acknowledgements

The authors would like to thank The Navigator Company in the scope of the PRODUTECH R3 project for its support and assistance in developing this work, including the availability to retrieve a dataset for pallet detection in the company's Setúbal Industrial Complex.

References

[1] N. Bellomo, E. Marcuzzi, L. Baglivo, M. Pertile, E. Bertolazzi, and M. De Cecco. "Pallet pose estimation with LIDAR and vision for autonomous forklifts". In: *IFAC Proceedings Volumes* 42.4 (2009), pp. 612–617. DOI: 10.3182/20090603-3-RU-2001.0540.

[2] M.R. Walter, S. Karaman, E. Frazzoli, and S. Teller. "Closed-loop pallet manipulation in unstructured environments". In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2010, pp. 5119–5126. DOI: 10.1109/IROS.2010.5652377.

[3] I.S. Mohamed, A. Capitanelli, F. Mastrogiovanni, S. Rovetta, and R. Zaccaria. "Detection, localisation and tracking of pallets using machine learning techniques and 2D range data". In: *Neural Computing and Applications* 32.13 (2020), pp. 8811–8828. DOI: 10.1007/s00521-019-04352-0.

[4] S. Byun and M. Kim. "Real-time positioning and orienting of pallets based on monocular vision". In: *2008 20th IEEE International Conference on Tools with Artificial Intelligence*. Vol. 2. 2008, pp. 505–508. DOI: 10.1109/ICTAI.2008.124.

[5] G. Chen, R. Peng, Z. Wang, and W. Zhao. "Pallet recognition and localization method for vision guided forklift". In: *2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*. 2012, pp. 1–4. DOI: 10.1109/WiCOM.2012.6478603.

[6] G.-Z. Cui, L.-S. Lu, Z.-D. He, L.-N. Yao, C.-X. Yang, B.-Y. Huang, and Z.-H. Hu. "A robust autonomous mobile forklift pallet recognition". In: *2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010)*. Vol. 3. 2010, pp. 286–290. DOI: 10.1109/CAR.2010.5456688.

[7] M. Seelinger and J.-D. Yoder. "Automatic pallet engagment by a vision guided forklift". In: *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*. 2005, pp. 4068–4073. DOI: 10.1109/ROBOT.2005.1570744.

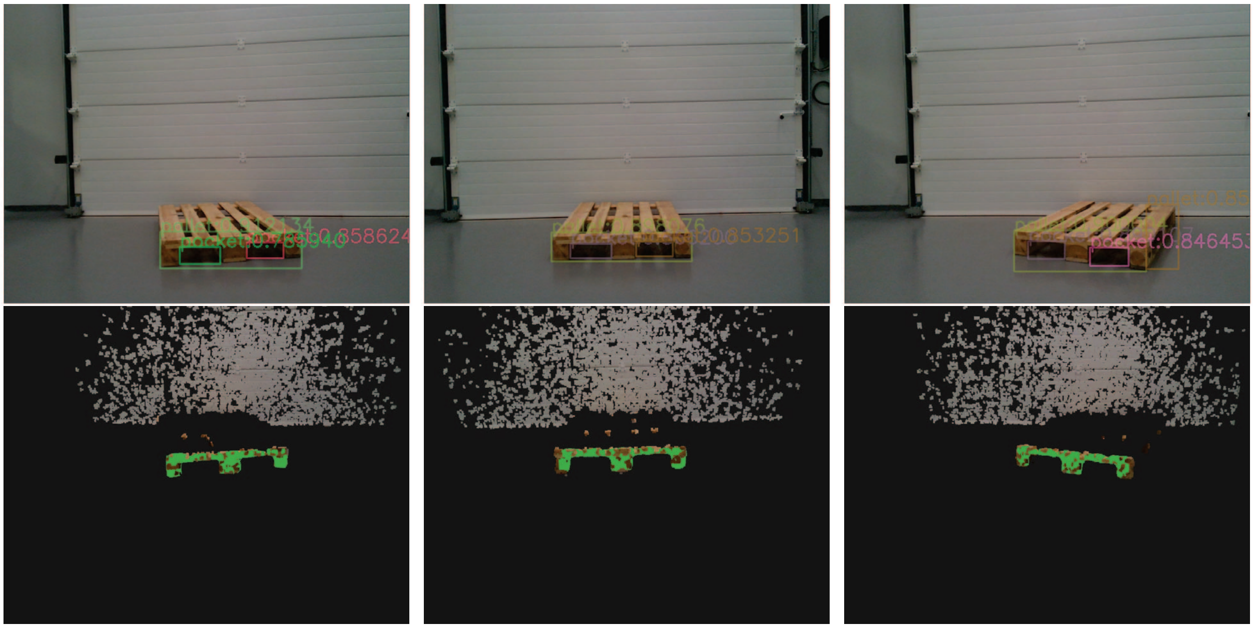

Fig. 6: Point cloud segmentation based on YOLOv8m model detections. RGB images on top represent the detections. Images on bottom are the point cloud segmentation of the front side of the pallet.

[8] M.M. Aref, R. Ghabcheloo, A. Kolu, M. Hyvönen, K. Huhtala, and J. Mattila. “Position-based visual servoing for pallet picking by an articulated-frame-steering hydraulic mobile machine”. In: *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. 2013, pp. 218–224. DOI: 10.1109/RAM.2013.6758587.

[9] M.F. Aswad, P.H. Rusmin, and R.N. Fatimah. “Marker-based detection and pose estimation of custom pallet using camera and laser rangefinder”. In: *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. 2023, pp. 186–191. DOI: 10.1109/ISITIA59021.2023.10221123.

[10] E.S. Kesuma, P.H. Rusmin, and D.A. Maharani. “Pallet detection and distance estimation with YOLO and fiducial marker algorithm in industrial forklift robot”. In: *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 2023, pp. 270–275. DOI: 10.1109/ICAIIC57133.2023.10066999.

[11] B. Molter and J. Fottner. “Real-time pallet localization with 3D camera technology for forklifts in logistic environments”. In: *2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*. 2018, pp. 297–302. DOI: 10.1109/SOLI.2018.8476740.

[12] T. Zhao, J. Liu, and J. Zhou. “High precision pallet detection and location algorithm for AGV”. In: *2023 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. 2023, pp. 170–173. DOI: 10.1109/CIS-RAM55796.2023.10370758.

[13] S. Wang, A. Ye, H. Guo, J. Gu, X. Wang, and K. Yuan. “Autonomous pallet localization and picking for industrial forklifts based on the line structured light”. In: *2016 IEEE International Conference on Mechatronics and Automation (ICMA)*. 2016, pp. 707–713. DOI: 10.1109/ICMA.2016.7558649.

[14] Y. Shao, Z. Fan, B. Zhu, M. Zhou, Z. Chen, and J. Lu. “A novel pallet detection method for automated guided vehicles based on point cloud data”. In: *Sensors* 22.20 (2022), p. 8019. DOI: 10.3390/s22208019.

[15] M. Zaccaria, R. Monica, and J. Aleotti. “A comparison of deep learning models for pallet detection in industrial warehouses”. In: *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*. 2020, pp. 417–422. DOI: 10.1109/ICCP51029.2020.9266168.

[16] P. Manurung, P.H. Rusmin, and R. Yusuf. “Custom pallet detection using YOLOv5 deep learning architecture”. In: *2022 International Symposium on Electronics and Smart Devices (ISESD)*. 2022, pp. 1–6. DOI: 10.1109/ISESD56103.2022.9980635.

[17] D. Caldana, R. Carvalho, P.M. Rebelo, M.F. Silva, P.G. Costa, H.M. Sobreira, and N. Cruz. “Comparison of pallet detection and location using COTS sensors and AI based applications”. In: *ROBOT 2023: Sixth Iberian Robotics Conference. Lecture Notes in Networks and Systems*. Vol. 976. 2024, pp. 495–508. DOI: 10.1007/978-3-031-58676-7_40.

[18] E. Tsiogas, I. Kleitsiotis, I. Kostavelis, et al. “Pallet detection and docking strategy for autonomous pallet truck AGV operation”. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

2021, pp. 3444–3451. DOI: 10.1109/IROS51168.2021.9636270.

[19] G. Bohacs, Z. Rozsa, and B. Bertalan. "Mono camera based pallet detection and pose estimation for automated guided vehicles". In: *LISS 2020*. 2021, pp. 1–11. DOI: 10.1007/978-981-33-4359-7_1.

[20] Y.-Y. Li, X.-H. Chen, G.-Y. Ding, S. Wang, W.-C. Xu, B.-B. Sun, and Q. Song. "Pallet detection and localization with RGB image and depth data using deep learning techniques". In: *2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE)*. 2021, pp. 306–310. DOI: 10.1109/CACRE52464.2021.9501390.

[21] R. Iinuma, Y. Hori, H. Onoyama, T. Fukao, and Y. Kubo. "Pallet detection and estimation for fork insertion with RGB-D camera". In: *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*. 2021, pp. 854–859. DOI: 10.1109/ICMA52036.2021.9512641.

[22] F. Jia, Z. Tao, and F. Wang. "Pallet detection based on Halcon and AlexNet network for autonomous forklifts". In: *2021 International Conference on Internet, Education and Information Technology (IEIT)*. 2021, pp. 86–89. DOI: 10.1109/IEIT53597.2021.00025.

[23] H. Wang, B. Cui, X. Wen, Y. Jiang, C. Gao, and Y. Tian. "Pallet detection and estimation with RGB-D salient feature learning". In: *2023 China Automation Congress (CAC)*. 2023, pp. 8914–8919. DOI: 10.1109/CAC59555.2023.10451178.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: unified, real-time object detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[25] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: towards real-time object detection with region proposal networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg. "SSD: single shot multibox detector". In: *Computer Vision – ECCV 2016*. 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[27] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[28] A. Cordeiro, J.P. Souza, C.M. Costa, V. Filipe, L.F. Rocha, and M.F. Silva. "Bin picking for shipsuilding logistics using perception and grasping systems". In: *Robotics* 12.1 (2023), p. 15. DOI: 10.3390/robotics12010015.

[29] J.P. Carvalho de Souza, C.M. Costa, L.F. Rocha, R. Arrais, A.P. Moreira, E.J.S. Pires, and J. Boaventura-Cunha. "Reconfigurable grasp glanning gipeline with grasp synthesis and selection applied to picking operations in aerospace factories". In: *Robotics and Computer-Integrated Manufacturing* 67 (2021), p. 102032. DOI: 10.1016/j.rcim.2020.102032.

[30] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A.A. Kalinin. "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11.2 (2020), p. 125. DOI: 10.3390/info11020125.