Advanced Techniques for Corners, Edges, and Stacked Gaps Detection and Pose Estimation of Cardboard Packages in Automated Dual-Arm Depalletising Systems

Santheep Yesudasu and Jean-François Brethé
GREAH, Normandy University, Le Havre, France

Keywords: Automated Depalletising, Cardboard Package Detection, Keypoint Detection, YOLOv8, Point Cloud Data, 3D Pose Estimation, Robotic Manipulation, Industrial Automation, Deep Learning, Computer Vision.

Abstract: This paper introduces advanced methods for detecting corners, edges, and gaps and estimating the pose of cardboard packages in automated depalletizing systems. Initially, traditional computer vision techniques such as edge detection, thresholding, and contour detection were used but fell short due to issues like variable lighting conditions and tightly packed arrangements. As a result, we shifted to deep learning techniques, utilizing the YOLOv8 model for superior results. By incorporating point cloud data from RGB-D cameras, we achieved better 3D positioning and structural analysis. Our approach involved careful dataset collection and annotation, followed by using YOLOv8 for keypoint detection and 3D mapping. The system's performance was thoroughly evaluated through simulations and physical tests, showing significant accuracy, robustness, and operational efficiency improvements. Results demonstrated high precision and recall, confirming the effectiveness of our approach in industrial applications. This research highlights the potential of using different sensors' information to feed the deep learning algorithms to advance automated depalletizing technologies.

1 INTRODUCTION

Automated depalletizing systems play a crucial role in modern logistics and manufacturing by improving package handling efficiency, accuracy, and safety. With the growing demand for automation, there is an increasing need for advanced techniques to enhance the precise detection and manipulation of packages. This paper presents the development of sophisticated methods for detecting key features such as corners, edges, and gaps and estimating the pose of cardboard packages, common in industrial environments.

Traditional computer vision techniques for object detection and pose estimation face challenges in scenarios involving partial occlusions, featureless objects, varying lighting, and tightly packed arrangements, highlighting a critical gap in the automation of depalletizing tasks. This research aims to overcome these limitations and improve the performance of automated depalletizing systems. Leveraging recent advances in deep learning, specifically YOLOv8, and the integration of point cloud data from RGB-D cameras, we achieve more accurate 3D positioning and structural analysis. Our approach significantly enhances the detection and localization of cardboard packages in complex industrial settings.

The contributions of this work are threefold:

1. We develop a novel methodology combining YOLOv8 for keypoint detection with point cloud data, enabling precise 3D localization and structural analysis of cardboard packages.
2. We create and annotate a dataset under diverse conditions, focusing on keypoint detection to optimize model training.
3. We rigorously evaluate the system's performance using various metrics, showing significant improvements over traditional methods.

This paper details the methodology, including dataset creation, YOLOv8-based keypoint detection, point cloud integration, and the evaluation process. We demonstrate substantial performance gains, addressing the identified challenges in automated depalletizing. The findings highlight the potential of integrating deep learning and 3D data for complex tasks in industrial automation, and we conclude by discussing real-world implications and future research directions for further enhancement of these systems.

Yesudasu, S. and Brethé, J.
Advanced Techniques for Corners, Edges, and Stacked Gaps Detection and Pose Estimation of Cardboard Packages in Automated Dual-Arm Depalletising Systems.
DOI: 10.5220/0013016900003822

In *Proceedings of the 21st International Conference on Informatics in Control, Automation and Robotics (ICINCO 2024) - Volume 2*, pages 264-273
ISBN: 978-989-758-717-7; ISSN: 2184-2809

2 RELATED WORK

2.1 Traditional Techniques

Traditional techniques for cardboard package detection and pose estimation have laid the groundwork for modern advancements in automated depalletising systems. These methods, while foundational, often struggle with limitations in complex and dynamic industrial environments. One of the primary traditional methods is RFID-based detection. RFID tags are attached to packages to facilitate identification and tracking throughout the logistics process. For instance, (Bouzakis and Overmeyer, 2010) demonstrated the use of RFID tags to describe the geometry of cardboard packages, enabling automated manipulation by industrial robots. Furthermore, RFID systems can detect package tampering and openings by analyzing changes in the radiation profile caused by the movement of RFID-based antennas, as highlighted by (Wang et al., 2020).

Another technique involves terahertz imaging, which utilizes terahertz waves to screen folded cardboard boxes for inserts or anomalies. This method offers high-speed and unambiguous detection capabilities, as noted by (Brinkmann et al., 2017). Visual monitoring and machine vision systems also play a crucial role. (Castaño-Amoros et al., 2022) explored the use of low-cost sensors and deep learning techniques to detect and recognize different types of cardboard packaging on pallets, optimizing warehouse logistics. Electrostatic techniques, as described by (Hearn and Ballard, 2005), leverage electrostatic charges to identify and sort waste packaging materials, differentiating between plastics and cardboard. Additionally, nonlinear ultrasonic methods, investigated by (Ha and Jhang, 2005), are employed to detect micro-delaminations in packaging by analyzing harmonic frequencies generated by ultrasonic waves.

While these traditional methods provide valuable insights and capabilities, they often face challenges such as accuracy, speed, cost, and environmental interference. These limitations have driven the development and adoption of more advanced techniques, particularly those based on deep learning.

2.2 Deep Learning Techniques

Deep learning techniques have revolutionized the field of cardboard package detection and pose estimation, offering significant improvements in accuracy, robustness, and efficiency. Convolutional Neural Networks (CNN) (Figure 1) form the backbone of these advancements, enabling the development of sophisticated models that can handle complex environments with ease. Models like YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) have set new benchmarks for real-time object detection. These models balance speed and accuracy, making them highly suitable for industrial applications where quick and precise detection is crucial.

Our 2023 study, (Yesudasu et al., 2023) explores the application of YOLOv3 for object detection in automated depalletization systems. YOLOv3 is renowned for its speed and accuracy, making it an ideal choice for real-time detection of cardboard packages on a pallet. The detection process in their study is seamlessly integrated with a pose estimation algorithm, enabling the system to determine the orientation and position of each package. This integration significantly enhances the efficiency and precision of the depalletization task. However, the previous system primarily handled free cardboard boxes without addressing the complexities of varied box locations and orientations. Additionally, it had limitations in detecting gaps between packages, a critical factor for optimizing the depalletization process. By learning hierarchical feature representations directly from data, these models excel in identifying and localizing objects in diverse and challenging scenarios. Deep learning extends beyond CNN to include architectures such as Deep Boltzmann Machines (DBM), Deep Belief Networks (DBN), and Stacked Denoising Autoencoders. These models have been successfully applied to various tasks, including face recognition, activity recognition, and human pose estimation. The versatility of deep learning in handling different computer vision challenges underscores its potential in cardboard package detection and pose estimation.

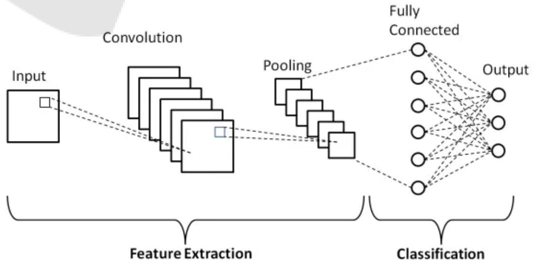


Figure 1: Architecture of a typical Convolutional Neural Network (Monica et al., 2020).

Significant strides have been made in object detection with models like Faster R-CNN, YOLOv3, and SSD. These models use region proposal networks, grid-based prediction, and multi-scale feature extraction to achieve high accuracy and efficiency. For example, Faster R-CNN integrates a region proposal network for efficient object detection, while YOLOv3 achieves real-time performance by dividing the image into grids and predicting bounding boxes and class probabilities for each cell. Deep learning has also found applications in robotics, enhancing perception, decision-making, and control. CNN are widely used for visual perception tasks, enabling robots to interpret and understand their environment in real-time. Recurrent Neural Networks (RNN), especially Long Short-Term Memory (LSTM) networks, handle temporal information, essential for tasks requiring sequence prediction and temporal context. Deep Reinforcement Learning (DRL) combines deep learning with reinforcement learning, enabling robots to learn optimal actions through trial and error. Generative Adversarial Networks (GAN) are used for generating synthetic data to train robots in simulation environments, improving the robustness of robotic perception systems.

2.3 Object Pose Estimation Techniques

Object pose estimation is critical for robotic systems, involving the determination of an object's position and orientation. Various advanced techniques have been developed to enhance the accuracy and efficiency of pose estimation in different applications.

RGB-D camera-based methods leverage depth information from sensors to enhance pose estimation. The Hybrid Reprojection Errors Optimization Model (HREOM) combines 3D-3D and 3D-2D reprojection errors for robust pose estimation in texture-less and structure-less scenes using RGB-D cameras (Yu et al., 2019). Additionally, 3D human pose estimation techniques use RGB-D images to estimate human poses for robotic task learning, enhancing robots' ability to mimic human actions (Zimmermann et al., 2018). Geometric and feature-based methods focus on analyzing the geometric properties of objects. The all-geometric approach utilizes distances between feature pairs and image coordinates for pose estimation with a single perspective view (Chandra and Abidi, 1990). Another technique, 6D pose estimation using Point Pair Features (PPF), employs multiple edge appearance models to handle occlusion-free object detection for robotic bin-picking (Liu et al., 2021). Deep learning-based methods have significantly advanced pose estimation. Deep Object Pose Estimation Networks use synthetic datasets and deep learning algorithms like CNN for 6-DOF pose estimation, achieving high accuracy in complex environments (Zhang et al., 2022). Pruned Hough Forests combine split schemes for effective pose estimation in cluttered environments, enhancing performance for robotic grasping tasks (Dong et al., 2021).

Pose estimation algorithms are crucial for various robotic applications, including navigation, manipulation, and human-robot interaction. Accurate pose estimation enables robots to interact with objects in their environment, perform tasks like assembly and bin-picking, and collaborate effectively with humans.

In summary, the advancements in traditional, deep learning, and object pose estimation techniques have significantly enhanced the capabilities of automated depalletising systems. These techniques address the challenges of accuracy, robustness, and efficiency, making them suitable for complex and dynamic industrial environments. Future research will continue to refine these methods, further improving the performance and reliability of automated depalletising systems.

3 METHODOLOGY

This section outlines the methodology used for detecting corners, edges, gaps, and pose estimation of cardboard packages in automated depalletising systems. Our approach leverages the advanced capabilities of YOLOv8 and integrates point cloud data for enhanced 3D analysis. Additionally, we explore traditional computer vision techniques and discuss their limitations, which led to the adoption of deep learning methods.

3.1 Classical Computer Vision Pipelines

3.1.1 Edge Detection

The initial phase of this research explored various traditional computer vision techniques to detect and grasp cardboard boxes. For edge detection, algorithms such as the Canny Edge Detector and Sobel Operator were employed. The Canny Edge Detector identifies edges by detecting rapid intensity changes, effectively outlining the boxes, while the Sobel Operator computes the gradient of the image intensity to highlight regions with high spatial frequency corresponding to edges.

3.1.2 Thresholding

Thresholding methods like Otsu's Method and Adaptive Thresholding were used to separate cardboard boxes from the background. Otsu's Method automatically finds the optimal threshold value, whereas Adaptive Thresholding adjusts the threshold dynamically for different image regions, useful under varying lighting conditions.

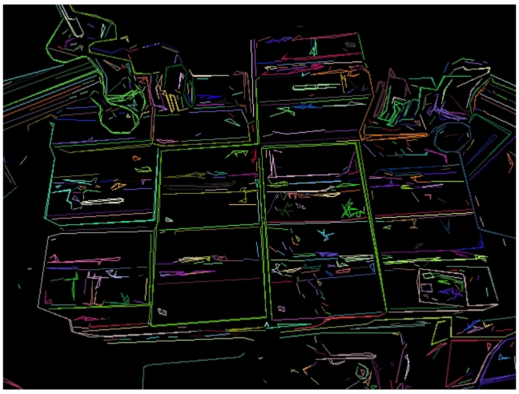


Figure 2: Traditional Computer Vision Techniques.

3.1.3 Contour Detection

Contour detection was implemented using OpenCV's FindContours function, which identifies the boundaries of boxes in a binary image. Shape analysis followed, where bounding boxes were drawn around detected contours (see Figure 2) to determine the location and size of the boxes and aspect ratio analysis was used to distinguish boxes from other objects based on their width-to-height ratio.

3.1.4 Template Matching

Template matching involved techniques like cross-correlation and normalized cross-correlation. Cross-correlation matches a predefined template of a box to the image to detect similar shapes, while normalized cross-correlation provides a more refined match, less affected by lighting and contrast changes.

3.1.5 Morphological Operations

Morphological operations, including erosion and dilation, were applied to remove noise and small irregularities in the binary image, making the boxes more distinct. Additionally, opening and closing operations, which are combinations of erosion and dilation, were used to clean up the image, filling small holes and removing small objects.

3.1.6 Feature Detection

Feature detection methods such as the Harris Corner Detector and FAST (Features from Accelerated Segment Test) were explored. The Harris Corner Detector identifies corners in the image which are common features of rectangular boxes, while FAST provides a quicker corner detection method suitable for real-time applications.

3.1.7 Line Detection

Line detection was performed using the Hough Line Transform and its probabilistic version. The Hough Line Transform detects lines in an image, aiding in identifying the edges and structure of the boxes, with the probabilistic version being more efficient in detecting line segments.

3.1.8 Color Segmentation

Finally, color segmentation was applied using the HSV color space. By converting images to HSV, it became easier to segment cardboard boxes based on color, assuming the boxes had distinct color properties.

3.2 Limitations and Transition to Deep Learning

3.2.1 Performance Issues

Despite extensive experimentation, traditional techniques struggled with accuracy, robustness, and handling occlusions, varying lighting conditions, and featureless surfaces of the boxes. Additionally, the tightly arranged boxes in pallets and the specific camera angles, with the camera located above the head of the robotic system, further complicated the detection process.

3.2.2 Decision to Shift

These limitations highlighted the need for a more advanced approach, prompting a transition to deep learning-based methods. Deep learning techniques offered superior performance in complex environments, providing enhanced accuracy and robustness for cardboard box detection and pose estimation in challenging conditions.

3.3 Dataset Collection and Annotation

To begin with, we collected and annotated 807 images of cardboard packages using the Computer Vision Annotation Tool (CVAT). The dataset was meticulously labeled to capture the precise details required for accurate detection and pose estimation. The keypoints were categorized into three classes based on the number of visible faces on the boxes:

- `boxF-1`: Includes the top four corners as keypoints.
- `boxF-2`: Includes the top four corners plus the bottom two corners of the visible side face.

- `boxF-3`: Includes the top four corners, the bottom two corners of the visible side face, and another bottom corner of an additional visible side face.

Each keypoint was annotated with its position and a visibility factor, indicating whether the keypoint was fully visible, fully occluded, or not labeled. This detailed annotation process ensures high-quality data for training the neural networks.

3.4 Keypoint Detection with YOLOv8

The YOLOv8 model was then trained to detect keypoints and skeletons of cardboard boxes. The model predicts keypoint coordinates and confidence scores, forming the skeletons necessary for structural analysis. Anchor Points and Regression, YOLOv8 employs predefined anchor points for keypoints, facilitating the prediction of the exact positions of keypoints relative to these anchors. For each anchor point, the network predicts parameters such as coordinates (tx, ty), representing the keypoints relative to the bounds of the grid cell, and a confidence score indicating the likelihood of each keypoint's presence. YOLOv8 predicts bounding boxes around detected cardboard boxes, including center coordinates, width and height, objectness score, and class probabilities. YOLOv8 also uses predefined anchor points for detecting the skeletons of cardboard boxes, assisting in predicting the key structural elements by providing skeleton keypoint coordinates and a confidence score for each skeleton keypoint.

3.5 Integration with Point Cloud Data

To enhance 3D positioning and structural analysis, we integrated point cloud data from RGB-D cameras with the detected keypoints and skeletons. This integration allows for precise calculation of box dimensions, gaps, and optimal grasp points. Point cloud data (P) is obtained from RGB-D cameras corresponding to the RGB images, where each point p_i in the point cloud is represented as $p_i = (x_i, y_i, z_i)$. The detected 2D keypoints from the YOLOv8 model are mapped onto the point cloud to determine their 3D coordinates. Detected 2D keypoints $K = \{k_1, k_2, \ldots, k_n\}$, where each $k_i = (u_i, v_i)$ represents the pixel coordinates in the image, are projected to 3D coordinates using the intrinsic camera matrix. The depth (z-coordinate) from the point cloud is matched to get the 3D coordinates $K_{3D} = \{(x_i, y_i, z_i)\}$.

3.5.1 Edge and Face Estimation

Edges are calculated by connecting the projected 3D keypoints, where an edge between two keypoints k_i and k_j is represented as a vector $\overrightarrow{E_{ij}} = \overrightarrow{P_j} - \overrightarrow{P_i}$. The planes representing the box faces are determined using the 3D keypoints, where the plane equation is given by $Ax + By + Cz + D = 0$. The normal vector $\mathbf{n}$ to the plane is calculated using the cross product of two vectors on the plane. For a plane defined by three non-collinear points P_1, P_2, P_3, the normal vector $\mathbf{n} = (A, B, C)$, and the plane constant D is calculated as $D = -(Ax_1 + By_1 + Cz_1)$.

3.5.2 Box Size Calculation

To determine the dimensions (height, width, length) of the boxes, we calculate the distances between the identified 3D keypoints. The height (h) is the vertical distance between the top and bottom keypoints on one face, the width (w) is the horizontal distance between the left and right keypoints on the same face, and the length (l) is the depth distance between the front and back keypoints of the box.

3.5.3 Gap Detection and Size Calculation

Gaps between boxes are identified by analyzing the distances and spatial relationships between the edges and faces of adjacent boxes. To identify gaps between two parallel planes, the distance d between them is calculated using

$$d = \frac{|D_1 - D_2|}{\sqrt{A^2 + B^2 + C^2}}$$

where D_1 and D_2 are the plane constants of two parallel planes with normal vector $\mathbf{n} = (A, B, C)$. The size of the gaps is measured by calculating the Euclidean distance between the nearest edges or corners of adjacent boxes, where for two points P_i and P_j on adjacent boxes, the gap size g is calculated as

$$g = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$$

The predicted grasping approaches were tested in a simulated environment to verify their effectiveness. The simulation provided a controlled setting to refine the algorithms and ensure they could handle various scenarios encountered in real-world operations. Successful simulations were followed by physical testing using the dual-arm manipulator, further validating the grasping strategies.

3.6 Performance Metrics

The system's performance was evaluated using key metrics: detection accuracy, grasping precision, and operational efficiency, showing significant improvements over previous models. Mean Average Precision (mAP) measures detection accuracy by calculating average precision across classes:

$$\text{AP} = \sum_{n=1}^{N} \frac{P(n) \cdot \Delta R(n)}{N}$$

Frames Per Second (FPS) gauges model speed:

$$\text{FPS} = \frac{\text{Number of Frames}}{\text{Total Time Taken}}$$

Intersection over Union (IoU) assesses bounding box overlap accuracy:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Recall evaluates the model's ability to detect all relevant instances, and the F1 score balances precision and recall:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This methodology offers a comprehensive solution for detecting corners, edges, and gaps, as well as estimating the pose of cardboard packages in automated depalletizing systems. By leveraging YOLOv8 and integrating point cloud data, we significantly improve the accuracy, robustness, and efficiency of these systems. Future work will focus on enhancing dataset diversity, optimizing real-time performance, and integrating real-time feedback mechanisms to further refine and improve the system's capabilities.

4 RESULTS AND DISCUSSION

This section presents the results of our methodology for detecting corners, edges, gaps, and pose estimation of cardboard packages in automated depalletising systems. We evaluate the system's performance using various metrics and discuss the implications of these results for real-world applications. The computer system has a high-performance Intel Core i7-10875H processor, 32GB of RAM, and an NVIDIA Quadro RTX 4000 GPU with 8GB of memory.

4.1 YOLOv8 Detection and Validation

The performance of the YOLOv8 model was evaluated across several object classes. As shown 3 Key metrics analyzed include F1-Confidence, Precision-Confidence, Precision-Recall, and Recall-Confidence, providing a comprehensive understanding of the model's accuracy and reliability at different confidence thresholds.

4.1.1 F1-Confidence Analysis

The F1-Confidence metric is essential for evaluating an object detection model's performance, illustrating the trade-offs between precision and recall. Our results show that the F1 score increases rapidly as the confidence threshold rises from 0 to approximately 0.3, indicating high recall but moderate precision. The F1 scores stabilize between 0.3 and 0.8 confidence thresholds, with an average F1 score of 0.91 at a confidence threshold of 0.624 for all classes. As the confidence threshold approaches 1.0, F1 scores decline due to increased precision at the expense of recall. BoxF-1 maintained the highest F1 scores, followed by boxF-2 and boxF-3. The 'all classes' curve demonstrated consistent performance with a high F1 score.

4.1.2 Precision-Confidence Analysis

The Precision-Confidence metric evaluates the model's ability to correctly identify objects without false positives. Precision increased rapidly as the confidence threshold rose to 0.3, stabilized between 0.3 and 0.8, and further increased at high confidence levels, minimizing false positives. BoxF-1 and boxF-2 maintained higher precision levels compared to boxF-3. The 'all classes' curve showed perfect precision (1.00) at a high confidence threshold (0.975), validating YOLOv8's robustness across varying confidence thresholds and making it suitable for tasks requiring high precision.

4.1.3 Precision-Recall Analysis

The Precision-Recall metric assesses the relationship between precision and recall, with the area under the curve (AUC) indicating overall performance. High precision values close to 1.0 were observed at lower recall levels, with a slight decline in precision as recall increased, especially for boxF-2. The 'all classes' curve maintained a high mean average precision (mAP) of 0.939 at an IoU threshold of 0.5. BoxF-1 maintained the highest precision-recall performance, followed by boxF-3 and boxF-2, demonstrating YOLOv8's proficiency in balancing precision and recall.

4.1.4 Recall-Confidence Analysis

The Recall-Confidence metric evaluates the model's ability to capture all relevant instances without missing any. High recall values close to 1.0 were observed at lower confidence levels, stabilizing between 0.3 and 0.8 confidence thresholds, with a decline at high

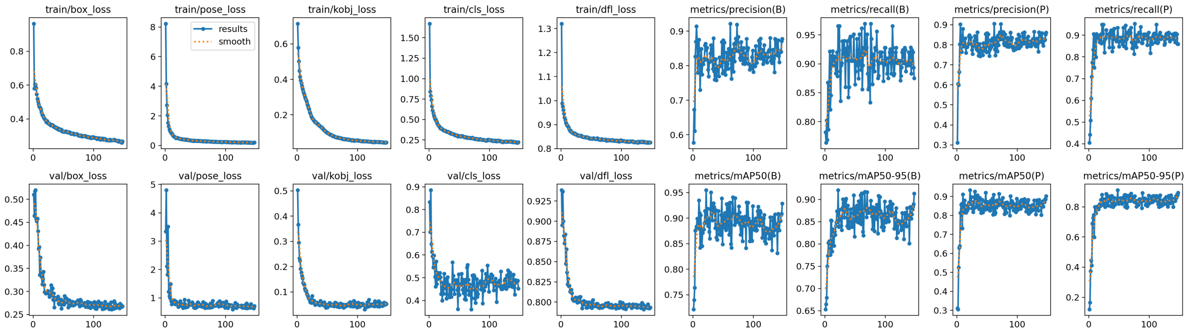


Figure 3: YOLOv8 pose estimation detection across different classes and keypoints. The graphs display various loss metrics, precision, and recall over epochs.

confidence levels due to increased precision. BoxF-1 maintained the highest recall scores, followed by boxF-3 and boxF-2. The 'all classes' curve showed a high recall score (0.98) at a low confidence threshold (0.000), demonstrating YOLOv8's robustness in capturing all relevant instances.

4.1.5 Validation and Test Metrics

Table 1 and Table 2 summarize the validation and test metrics for YOLOv8 object detection. BoxF-1 exhibited the highest precision, recall, and F1 scores, followed by boxF-2 and boxF-3. The combined 'all classes' metrics confirmed YOLOv8's excellent performance across different object classes and confidence thresholds.

Table 1: Validation Metrics for YOLOv8 Object Detection.

Class	Precision	Recall	F1 Score
boxF-1	0.95	0.96	0.95
boxF-2	0.92	0.94	0.93
boxF-3	0.91	0.90	0.91
All Classes	**0.93**	**0.93**	**0.93**

Table 2: Test Metrics for YOLOv8 Object Detection.

Class	Precision	Recall	F1 Score
boxF-1	0.93	0.94	0.94
boxF-2	0.90	0.91	0.91
boxF-3	0.88	0.89	0.88
All Classes	**0.90**	**0.91**	**0.91**

4.2 YOLOv8 Keypoints Detection Results

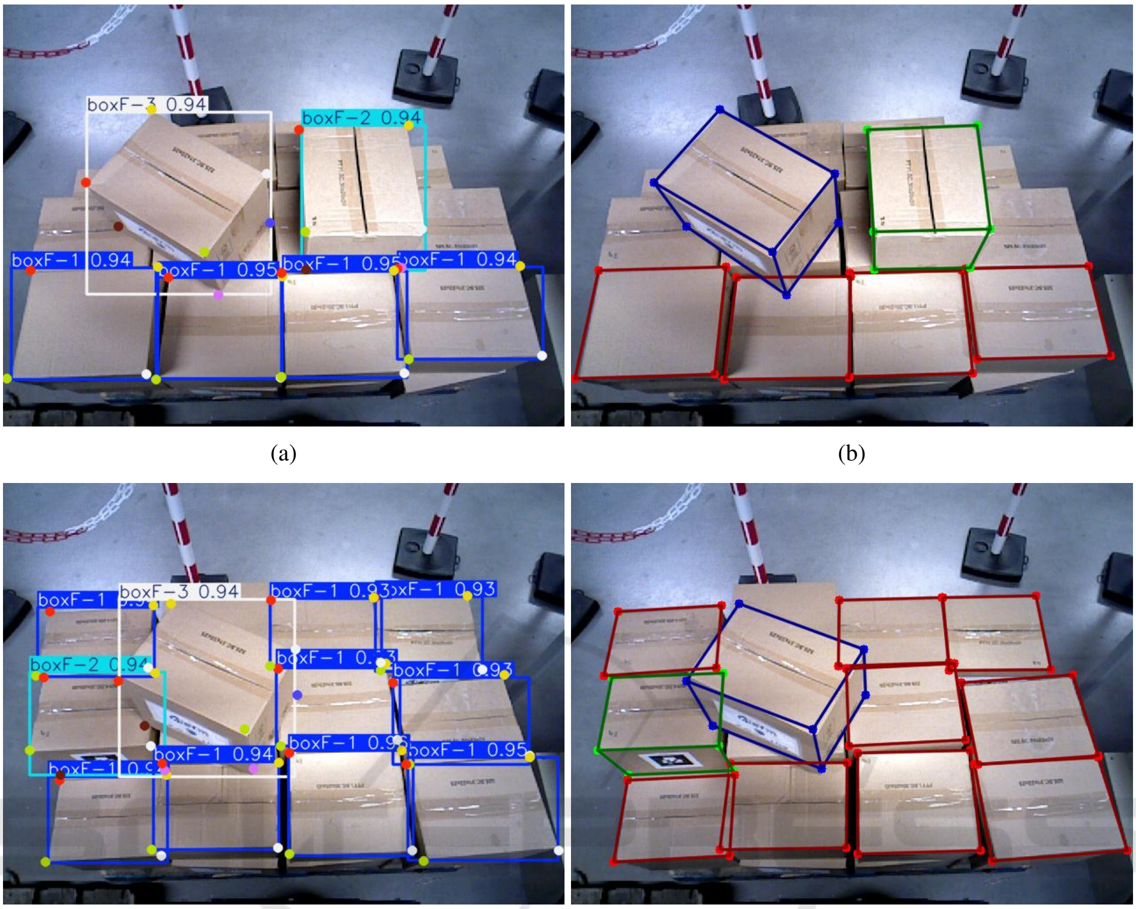
The YOLOv8 model was trained to detect keypoints on cardboard boxes, distinguishing between different faces and edges of the boxes. Figures (5) illustrate the model's output on test images, with annotations indicating the detected keypoints and the respective confidence scores.

The results show high accuracy in detecting keypoints on various faces of the cardboard boxes, as evidenced by the clear and precise annotations. The keypoints, marked with different colors, correspond to the corners and edges of the boxes, facilitating accurate localization. The model effectively handles occlusions and overlapping boxes, demonstrating robustness in detecting partially visible boxes and keypoints in complex arrangements. This capability is crucial for real-world applications where boxes may be tightly packed or partially obscured.

The precise detection of keypoints allows the system to calculate the optimal grasping points and plan the trajectories for the dual-arm manipulator. The ability to identify gaps between boxes, as well as the edges and corners, ensures that the robot can effectively grasp and move the boxes without causing damage or disrupting the arrangement.

The model was trained with the following hyperparameters: 2000 epochs, a batch size of 16, and an input image resolution of 640x640 pixels. A warmup phase of 3 epochs was applied to gradually ramp up the learning rate. The initial learning rate was set to 0.01, with a linear decay to a final learning rate (LRF) of 0.01. A momentum value of 0.937 and a weight decay of 0.0005 were used to stabilize the optimization process. During evaluation, an Intersection over Union (IoU) threshold of 0.7 was employed to balance precision and recall in the model's performance.

An important aspect of our approach was ensuring robustness across diverse lighting conditions. Although RGB-D cameras typically depend on optimal lighting for accurate depth and RGB data, our model mitigates this limitation by supporting low-light environments. This was achieved by training the model on datasets that included both normal and low-light conditions, maintaining consistent detection accuracy even in suboptimal lighting. This adaptability in-



(a) (b)

(c) (d)

Figure 4: (a) & (c) The model's output on test images, indicating the detected keypoints and the respective confidence scores (b) & (d) The results of YOLOv8 skeleton detection for cardboard boxes, even if it is partially occluded.

creases the model's effectiveness in real-world industrial applications, where lighting conditions are often uncontrolled.

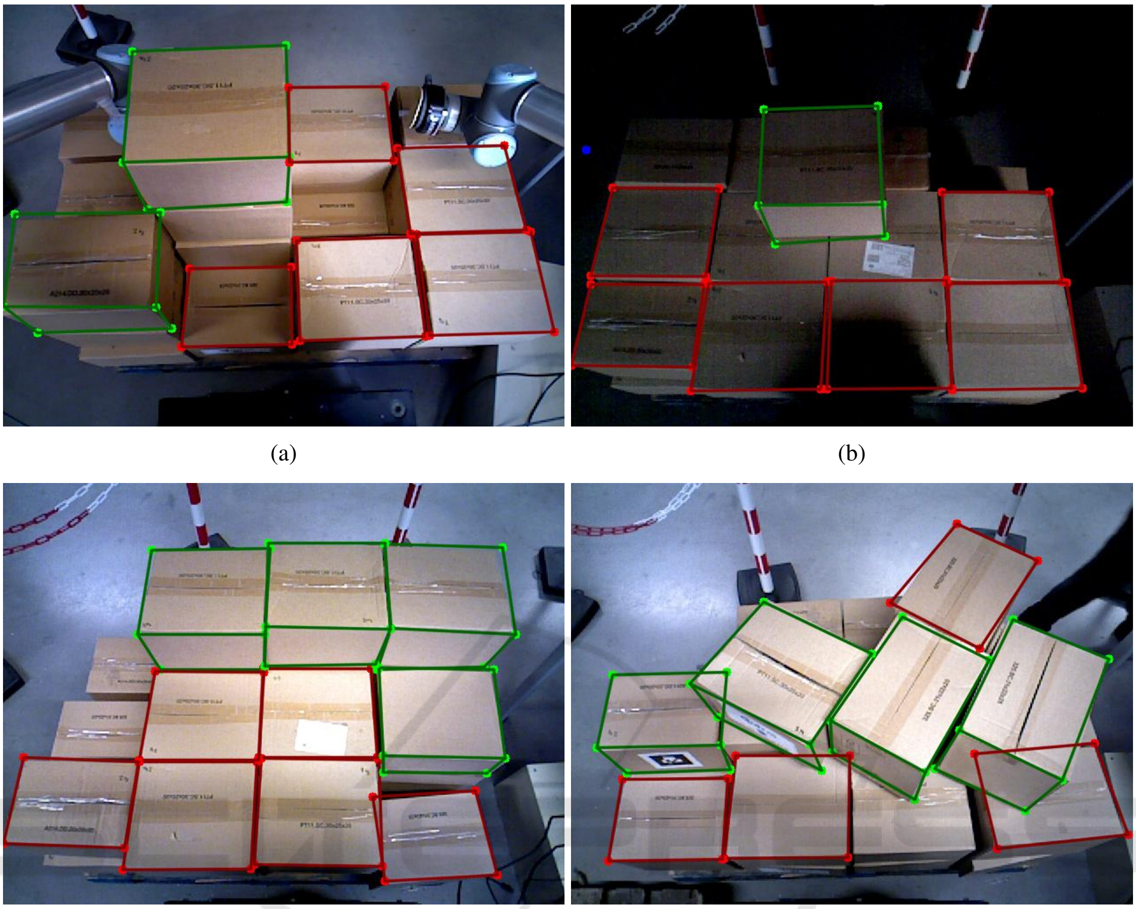
4.3 Skeleton and Prioritized Gap Detection Results

The skeleton detection results, with the identified prioritized grasping points, provide several advantages. The precise identification of keypoints and the prioritized grasping point allows for accurate calculation of the optimal grasping strategy, ensuring secure handling of the boxes. By focusing on the most suitable grasping point, the system can execute grasping actions more quickly and effectively, improving the overall efficiency of the depalletising process. The ability to detect keypoints and determine the best grasping point is robust to variations in box placement and orientation, making the system adaptable to different scenarios and box arrangements.

While the current results are promising, further improvements can be made by enhancing dataset diversity, including a wider variety of box types and environments in the training dataset to improve the model's robustness and generalizability. Integrating real-time force and torque feedback during grasping can further enhance the precision and safety of the manipulation process. Ensuring that the detection and processing can be performed in real-time will be critical for deploying the system in dynamic industrial settings.

5 CONCLUSION

In this study, we have introduced a comprehensive methodology for detecting corners, edges, and gaps and estimating the pose of cardboard packages in automated dual-arm depalletising systems. Leveraging the advanced capabilities of the YOLOv8 model, cou-



(a) (b) (c) (d)

Figure 5: (a) YOLOv8 skeleton detection: Detecting boxes at different levels in the pallet, (b) Handling different lighting conditions, (c) Adapting to different complex environments, (d) Managing rotated boxes and partially occluded boxes.

pled with point cloud data from RGB-D cameras, we have addressed the significant challenges associated with traditional computer vision techniques. Our approach demonstrated marked improvements in detection accuracy, robustness, and operational efficiency, particularly in handling complex scenarios such as occlusions, varying lighting conditions, and tightly packed arrangements. The rigorous process of dataset collection and annotation, combined with the use of sophisticated detection algorithms, has facilitated precise calculations of box dimensions and optimal grasp points. This has significantly enhanced the efficiency and reliability of robotic manipulation, validating our methodology through extensive simulation and physical testing.

While our results are promising, several areas warrant further investigation and enhancement. Expanding the dataset to include a wider variety of box types, colors, and environments will improve the model's robustness and generalizability. Optimizing the model for real-time processing is crucial for its deployment in dynamic industrial settings, ensuring swift and accurate detection and manipulation. Integrating real-time force and torque feedback during grasping can enhance precision and safety, reducing the likelihood of errors and damage during manipulation. Investigating the system's scalability for larger and more varied industrial applications will help understand its limitations and areas for improvement. Exploring the potential for human-robot interaction and collaboration in depalletising tasks can open new avenues for efficiency and safety in industrial environments. In conclusion, this research underscores the potential of integrating deep learning with precise 3D data to advance automated depalletising systems. By continuing to refine and build upon this work, we aim to develop more adaptable, efficient, and reliable automated systems that can meet the evolving demands of modern industries.

REFERENCES

Bouzakis, A. and Overmeyer, L. (2010). Rfid-assisted detection and handling of packages. In *ROMANSY 18 Robot Design, Dynamics and Control: Proceedings of The Eighteenth CISM-IFToMM Symposium*, pages 367–374. Springer.

Brinkmann, S., Vieweg, N., Gärtner, G., Plew, P., and Deninger, A. (2017). Towards quality control in pharmaceutical packaging: Screening folded boxes for package inserts. *Journal of Infrared, Millimeter, and Terahertz Waves*, 38:339–346.

Castaño-Amoros, J., Fuentes, F., and Gil, P. (2022). Visual monitoring intelligent system for cardboard packaging lines. In *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–8. IEEE.

Chandra, T. and Abidi, M. (1990). A new all-geometric pose estimation algorithm using a single perspective view. In *Conference Proceedings*.

Dong, H., Prasad, D. K., and Chen, I. (2021). Object pose estimation via pruned hough forest with combined split schemes for robotic grasp. *IEEE Transactions on Automation Science and Engineering*, 18:1814–1821.

Ha, J. and Jhang, K. (2005). Nonlinear ultrasonic method to detect micro-delamination in electronic packaging. *Key Engineering Materials*, 297-300:813–818.

Hearn, G. and Ballard, J. R. (2005). The use of electrostatic techniques for the identification and sorting of waste packaging materials. *Resources Conservation and Recycling*, 44:91–98.

Liu, D., Arai, S., Xu, Y., Tokuda, F., and Kosuge, K. (2021). 6d pose estimation of occlusion-free objects for robotic bin-picking using ppf-meam with 2d images (occlusion-free ppf-meam). *IEEE Access*, 9:50857–50871.

Monica, R., Aleotti, J., and Rizzini, D. L. (2020). Detection of parcel boxes for pallet unloading using a 3d time-of-flight industrial sensor. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 314–318. IEEE.

Wang, W., Sadeqi, A., Nejad, H. R., and Sonkusale, S. (2020). Cost-effective wireless sensors for detection of package opening and tampering. *IEEE access*, 8:117122–117132.

Yesudasu, S., Sebbata, W., Brethé, J.-F., and Bonnin, P. (2023). Depalletisation humanoid torso: Real-time cardboard package detection based on deep learning and pose estimation algorithm. In *2023 27th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 228–233. IEEE.

Yu, H., Fu, Q., Yang, Z., Tan, L., Sun, W., and Sun, M. (2019). Robust robot pose estimation for challenging scenes with an rgb-d camera. *IEEE Sensors Journal*, 19:2217–2229.

Zhang, H., Liang, Z., Li, C., Zhong, H., Liu, L., Zhao, C., Wang, Y., and Wu, Q. (2022). A practical robotic grasping method by using 6-d pose estimation with protective correction. *IEEE Transactions on Industrial Electronics*, 69:3876–3886.

Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., and Brox, T. (2018). 3d human pose estimation in rgbd images for robotic task learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1986–1992.