Vision-Enhanced System For Human-Robot Disassembly Factory Cells: Introducing A New Screw Dataset

2022 4th International Conference on Robotics and Computer Vision (ICRCV) | 978-1-6654-8170-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRCV55858.2022.9953178

Georgios Kalitsios
Information Technologies Institute
CERTH
Thessaloniki, Greece
gkalitsios@iti.gr

Lazaros Lazaridis
Information Technologies Institute
CERTH
Thessaloniki, Greece
lazlazari@iti.gr

Athanasios Psaltis
Information Technologies Institute
CERTH
Thessaloniki, Greece
at.psaltis@iti.gr

Apostolos Axenopoulos
Information Technologies Institute
CERTH
Thessaloniki, Greece
axenop@iti.gr

Petros Daras
Information Technologies Institute
CERTH
Thessaloniki, Greece
daras@iti.gr

***Abstract*—Waste from electrical and electronic equipment is exacerbating the global environmental crisis. There is an urgent need to build a robust infrastructure capable of providing effective e-waste disposal options. In this work, a novel hybrid human-robot and system-agnostic application for relevant waste disassembly and recycling has been developed. Working on cells, collaborative robots, enhanced with state-of-the-art computer vision capabilities, can achieve near-real-time performance and high precision in the disassembly process. Additionally, a new screw dataset suitable for three separate computer vision tasks, namely instance segmentation, object detection, and semantic segmentation, is introduced to facilitate future research, which can be utilized almost for any screwing/unscrewing application beyond the current disassembly topic. Experiments demonstrating the robustness of the visual object detection and robotic 3D deprojection modules, which are the core aspects of the proposed architecture, have been conducted.**

***Index Terms*—WEEE recycling, Robotic disassembly, Screw dataset, Robotic vision, Object recognition, Perception systems, Scene analysis**

I. Introduction

Waste from Electrical and Electronic Equipment (WEEE) is now one of the world's fastest-growing waste streams, with experts forecasting that it will continue to grow at a pace of 3 to 5% each year [1]. The technical advancements in robotics, industry 4.0, artificial intelligence (AI), and the needs of factories of the future push robots and humans into close collaboration, with the ultimate goal of increasing industrial productivity and flexibility. Robotic-based WEEE disassembly systems should replace the current hazardous, heavy, and time-consuming processes carried out mostly by human workers. This change reduces the health and safety concerns for human employees posed by potentially hazardous waste items, often handled in plants, and allows workers to focus on higher-skilled, higher-quality, and less intensive tasks. To enable efficient human-robot collaboration [2], such collaborative robots (cobots) perception system should be enhanced with computer vision (CV) capabilities based on deep learning (DL), transforming them into active and effective co-workers. Very few studies, for instance [3], [4] investigated the use of cobots for disassembly systems without any CV method and based their perception only on approximate positions and spiral search of the element of interest. Gil et al. [5] designed a multi-sensorial robotic system to perform disassembly on electronic equipment by identifying covers, wires, batteries, screws, etc. At the detection phase, the later employed a variety of classic CV algorithms such as adaptive thresholding, Douglas–Peucker's algorithm for polygonal approach, Progressive probabilistic hough transform, template matching and edge detection with Canny's detector.

DiFilippo et al. [6] designed a system that combines CV and force-sensing technology. Using a Hough Circle Transform, the overhead camera detected circles as screw candidates once a laptop was put on the workspace. Using another camera mounted on the robot's end-effector, the robot would then move to the positions of these circles and apply CV methods to center the screws. This operation took some time to perform. Bdwidi et al. [7] also developed a workstation for disassembling electric vehicle motors automatically. To detect screws, they used a sensor that could provide depth data, a feature point detector like the Harris detector, and various optimization phases. These classifiers have the problem of being very sensitive to lighting conditions and producing a large number of false positives.

The goal of this work is to introduce a novel framework able to support the challenging task of human-robot disassembly of electronic devices. More specifically, the paper introduces the following innovative features: a) **a novel system-agnostic**

architecture design for the WEEE disassembling system based on cutting-edge components and the most up-to-date software development tools and libraries, where system-agnostic refers to the ability to combine different DL architectures for the visual object detection module as well as its generalisation to different application domains beyond WEEE disassembly b) traditional CV techniques used for the perception system of robots are replaced with **a deep learning-assisted computer vision method**, providing an approach more generalizable, more accurate, and faster than traditional techniques. The method utilized in our system, in particular, operates at 7 FPS on a single GPU, making it a real-time application, and c) to address the challenging task of screw detection, which, according to the above studies, lies at the core of the disassembly process, **a new large-scale screw dataset** is introduced. The proposed screw dataset contains 945 images and over $4,000$ annotated screw instances suitable for three separate computer vision tasks, namely instance segmentation, object detection, and semantic segmentation. The formed dataset can improve detection accuracy of particularly small objects and it is applicable to any screwing/unscrewing tasks beyond WEEE disassembly.

The rest of the paper is organised as follows: the proposed system architecture is presented in Section II. Section III introduces the new screw dataset. The disassembly visual object detector along with experimental results are presented in Section IV, 2D to 3D deprojection scheme and evaluation is being addressed in Section V and the paper is concluded in Section VI.

II. System Architecture

Pivotal to the design of our approach is the observation that the visual object detection and the robotic 3D deprojection modules are two standalone systems that have shareable and co-occurring features. The main objective of our application is to assemble the component sub-systems of the disassembly-related task and to ensure that the sub-systems can efficiently, functionally and physically be integrated into a complete recycling plant solution (working cell). The purpose of the system is to associate the information derived from the visual object detection module with the manipulation processes of the robot. Figure 1 illustrates the overall system's architecture.

World Definition: Our ambition is the system to be used by articulated-stable cobots. In this case, it is necessary to define a working area frame (world coordinate frame). Its purpose is to localize the robot in the world and to use it as a reference for the relationship between the sensor and the actuation frame. To this end, as a preliminary step, the so-called hand-eye calibration problem has to be resolved. This procedure determines and computes a circle of spatial transformations-relationships between a robot and one or more sensors. The outcome is all the transformations needed to estimate the association between the Intel RealSense RGB-D camera sensor and the robot end-effector as well as the transformation from the world coordinate frame to the robot base. Assuming that A is the robot end-effector system, B the camera system and X,Z the unknown transformations matrices, the formula of the hand-eye calibration problem is AX=ZB. In this work, the aforementioned method is system-agnostic, completely automated and embedded in the system's deployment procedure. This is a preparatory step that occurs only at the system's initialization phase. Last but not least, an Aruco [8] marker, attached to the robot's end-effector is used as a calibration object.

Vision Isolation: The AI vision system of the disassembly step is completely isolated and dockerized, communicating with the rest components of the system through a client-server scheme that is based on the ZeroMQ library [9]. RGB frames are efficiently stored in the shared memory - */dev/shm*, giving access to the dockerized object detector to further process them.

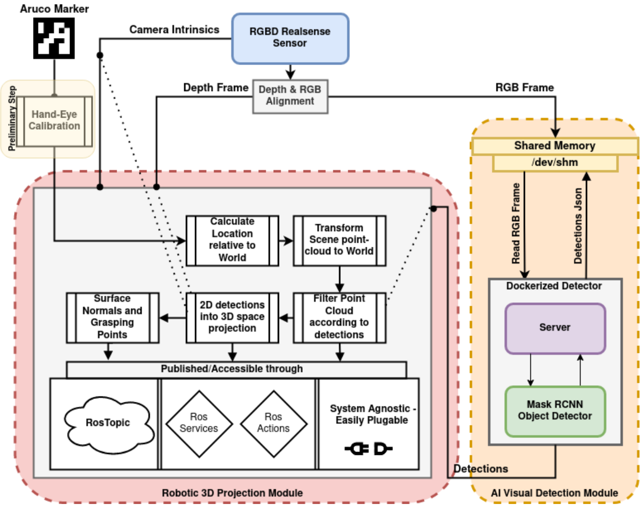
Aruco Marker
Camera Intrinsics
RGBD Realsense Sensor
Depth Frame
Depth & RGB Alignment
RGB Frame
Preliminary Step
Hand-Eye Calibration
Shared Memory
/dev/shm
Read RGB Frame
Detections Json
Calculate Location relative to World
Transform Scene point-cloud to World
Surface Normals and Grasping Points
2D detections into 3D space projection
Filter Point Cloud according to detections
Dockerized Detector
Server
Mask RCNN Object Detector
Published/Accessible through
RosTopic
Ros Services
Ros Actions
System Agnostic - Easily Plugable
Detections
Robotic 3D Projection Module
AI Visual Detection Module

Fig. 1. System Architecture: A Intel RealSense RGB-D sensor feeds the system with RGB-D frames and camera parameters. The dockerized AI Visual object detection is responsible for the 2D detection task, while the 3D projection module is in charge of forming the robot's operation area (world frame), filtering the detected components from the whole scene and projecting them into the 3D space with respect to the World's-Robot's frame.

Scene understanding: A scene may be succinctly characterized as a composition of scenes of objects and their relations. Object detection is a technology, usually based on computer vision, machine learning and image processing, that detects and identifies instances of semantic objects in 2D images. Consequently, object detection can separate and classify objects from a 2D scene. However, point cloud data is needed for the environment perception of robots. Point clouds of a scene can be generated using RGB and depth information captured by RGB-D sensors. Therefore, there is a need to align the RGB and Depth data produced by those sensors in order to produce high-quality point clouds. To this end, Intel's RealSense SDK [10] that provides multiple RGB and Depth alignment techniques has been used.

Point cloud processing: The Point Cloud Library [11] (PCL) has been used to process the point clouds produced by Intel's RealSense RGB-D camera. To avoid broadband issues and failures, the point cloud is filtered so that only specific components of high interest of the scene are transmitted through ROS [12]. More specifically, the point cloud is being cropped based on the bounding boxes and segmentation masks that have been extracted from the vision system.

Reference frame and 3D deprojection: The relationships between the camera and the robot's base are being calculated using the information derived from the Aruco marker and the hand-eye calibration procedure. At this stage, the ROS TF package, which is adequate to keep track of multiple coordinate frames over time, is utilized. Additionally, at the same time, the robot's base is matched and composes the center of the system's world. Next, the 3D coordinates of detections are represented in the camera-centered coordinate system. Using the bounding box information derived from the vision system, the center pixel of the bounding box is calculated. A normal vector is extracted at this point alongside the grasping points. Then, using the camera parameters (intrinsics) the 3D coordinates of the detections are transformed and de-projected to the robot coordinate system.

III. Screw Dataset

DL-based computer vision methods rely heavily on the quantity and quality of labeled data. Our system was tested on a dataset of common WEEE devices (PC Towers, Microwave Ovens, Flat Panel Displays and Emergency lamps) and their components (cables, screws, printed circuit boards (PCB), capacitors, batteries, motors, etc.), which will be called the Four WEEE devices dataset. Moreover, in this work, a new screw dataset is introduced (Figure 2), as the unscrewing task is the first and one of the most challenging steps in the disassembly procedure.

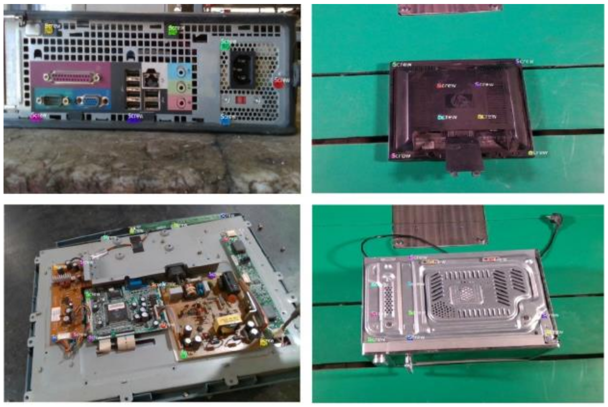

Fig. 2. Screw examples for different WEEE devices.

One of the first attemps to form a large scale screw classification datasets is found in the work of Yildiz et al. [13] where a set of 10, 000 images of screws as positive samples as well as non-screw artifacts have been collected, the images had a resolution of 150x150 pixels and only one screw instance per image. Brogan et al. [14] presented an object detection dataset of 1, 170 close range images with electral devices in good condition, 2, 189 screw instances, and an average of 1.87 screw instances per image.

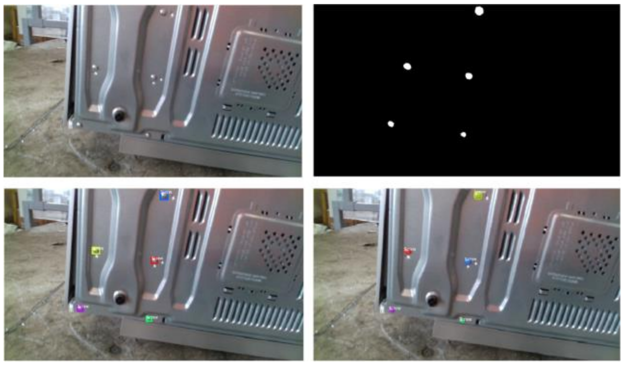

Fig. 3. Screw annotations for Semantic Segmentation, Object Detection and Instance Segmentation respectively.

A. Dataset Collection and Statistics

The introduced screw dataset[1] comprises a wide variety of device types, including damaged and deformable devices, as would be expected in a realistic disassembly scenario. Data were recorded under various lighting conditions, rotations, and device movements were employed. Recordings were made using a multi-camera configuration to obtain more informative views. Images captured from a close distance with a camera mounted on a screwdriver and a hand-held camera from various angles. Additionally, images from a greater distance were captured using cameras in fixed positions at a maximum distance of 90 cm.

TABLE I
Screw Dataset's Statistics.

Images	Resolution	Total Screw instances	Average screws per image
945	1280x720	4,414	4.7

The screw dataset consists of 945 high definition 1280x720 images, 4, 414 screw instances and 4.7 screws per image as shown in Table I. Blur estimation was employed to exclude blurry frames from the annotation process. The dataset includes annotated data in COCO [15] format for three different computer vision tasks: instance segmentation, object recognition, and semantic segmentation as shown in Figure 3. From these images, 52.7% were recorded in a laboratory environment and 47.3% recorded in WEEE recycling plants. In the training, validation, and test sets, there are 765, 90 and 90 fully annotated frames, respectively.

IV. Disassembly Vision Detector

The Deep Learning era has significantly improved the majority of Computer Vision domains. In the case of object detection, replacing the original handcrafted feature extraction method with deep data-driven architectures has shown great potential and produced impressive results. In scenarios where there is a lot of overlap between the disassembly components, recognizing them by their bounding boxes would result in a lot of ambiguity, hence instance segmentation is preferred.

[1] https://vcl.iti.gr/dataset/weee-disassembly-screw-dataset/

Mask R-CNN [16] was selected as the core architecture for this work because of its cutting-edge performance and efficiency. Mask R-CNN is composed of two distinct modules that are responsible for region proposing and classification. A set of candidate regions of predefined shape and size, known as anchors, is uniformly created over the image in the first stage. The RPN then validates each anchor based on its likelihood of containing a ground truth object. The proposed regions are made up of the most reliable anchors in terms of objectivity, and they are then sent to the second stage for additional classification. Furthermore, the Mask R-CNN architecture, which uses the Feature Pyramid Network (FPN) [17] to create discrete feature maps for different object sizes, makes it an even more attractive method in scenarios when small object recognition is required.

A. Implementation details

For feature extraction, the ResNet-101 [18] backbone was combined with the FPN [17] neck. The input images were adjusted so that their largest dimension is 1024 pixels wide while maintaining their aspect ratio. Our training is carried out on a single NVIDIA RTX 2080TI GPU with 11GB of VRAM. The weights of the network were set up using a model that had been pre-trained on the MS COCO dataset. With a momentum of 0.9, the stochastic gradient descent optimizer was applied. The uniform distributed anchors were made with ratios of 0.5, 1, and 2 and scales of [8, 16, 32, 64, 128] to target objects of varied sizes. The model was trained for 300 epochs, with an initial learning rate of 0.001 that declined by a factor of three at epoch 100 and 200. Finally, data augmentation was used to expand the amount of data by rotating it by 90, 180, and 270 degrees as well as flipping it vertically and horizontally during training.

B. Evaluation

The visual object detection module has been tested for detection and recognition of all components (cables, screws, PCBs, capacitors, etc.) on the Four WEEE devices dataset, as well as for screw detection only, using the standard COCO mean Average Precision and Recall metrics. In the former case, due to the complicated nature of the problem, the detection performance is deemed satisfactory (Table II). In the latter case, it is clear that training on the introduced screw dataset to detect screws achieves higher accuracy than training on the Four WEEE devices dataset (Table III), which makes sense given that the proposed dataset is a subset of the Four WEEE devices dataset that has been enhanced with extra annotated screw frames.

TABLE II
EXPERIMENT RESULTS ON THE FOUR WEEE DISASSEMBLY DATASET.

WEEE Device	**AP**$_{0.50:0.95}$	**AP**$_{0.50}$	**AP**$_{0.75}$	**AP**$_{S}$	**AR**$_{0.50:0.95}$	**AR**$_{S}$
PC Tower	0,689	0,836	0,623	0,305	0,591	0,419
Emergency Lamp	0,729	0,828	0,706	0,468	0,623	0,489
Flat Panel Display	0,657	0,641	0,614	0,379	0,644	0,318
Microwave Oven	0,513	0,609	0,435	0,457	0,445	0,391
Average	**0,647**	**0,729**	**0,595**	**0,402**	**0,576**	**0,404**

TABLE III
EXPERIMENT RESULTS ON THE INTRODUCED SCREW DATASET.

Dataset	**AP**$_{0.50:0.95}$	**AP**$_{0.50}$	**AP**$_{0.75}$	**AP**$_{S}$	**AR**$_{0.50:0.95}$	**AR**$_{S}$
Four WEEE devices dataset	0.478	0.532	0.386	0.397	0.412	0.458
Introduced Screw dataset	0.586	0.849	0.442	0.431	0.637	0.537

A disassembly screw detector was implemented to address the difficult task of screw detection, which is at the heart of the disassembly process. This detector incorporates several techniques, including a) Region-proposal tuning, which is an approach aimed at establishing better anchors for tiny objects. b) Multiscales representation, which employs the FPN Module to integrate rich semantic information from high-level feature maps with detailed location information from low-level feature maps. c) Reducing the number of additional component classes for disassembly and training a screw-only model.

V. DEPROJECTION AND EVALUATION

This section describes the deprojection scheme correlating the images (2D) derived by the Intel RealSense to their associated 3D coordinate system and the relationship between those two systems. Also, an evaluation of our application and the deprojection approach is included.

2D Coordinate system (Pixels): Each stream of digital images provided by a Intel RealSense sensor is made up of pixels rows [width] and columns [height] that are associated with the 2D coordinate space. The coordinate [0,0] refers to the most top left pixel of the image while the the positive x and y axis point to the right and down respectively.

3D Coordinate system (Points): In Intel RealSense sensors, each image stream is related with the 3D coordinate space too. The coordinate [0,0,0] indicates the image center, while the positive x,y and z axis point to right, down and forward respectively. The relationship between the 2D and 3D coordinate system or pixels and points is specified by the camera intrinsic parameters. Intel RealSense SDK provides those parameters in the internal structure and that makes mapping operations achievable. More specifically, both mapping a point from 3D coordinate space to a pixel in 2D coordinate frame and the reverse operation are build-in functions of the Intel RealSense SDK. The first operation is called projection while the second one is the deprojection.

A. Setup and Evaluation

The purpose of this task is to evaluate the error derived from the transformations that occur when deprojecting the screw detections from 2D to 3D coordinate system and to estimate the accuracy of our system in the 3D space.

Taking into account the size of the screws on the case of microwave oven, 7mm diameter screws with 4mm cross in the center have been printed. Grid paper has been used as a guide for measuring the distances between every point. Moreover, a hand meter and a ruler were used for confirmation purposes. Using the visual screw detector that presented in section 4, the center (Pixel 2D coordinates) of each screw has been found. Those points have been converted to 3D coordinates using the deproject function by utilizing the intrinsic parameters of

TABLE IV

COMPARISON OF HAND MEASUREMENTS AND DEPROJECTED POINTS DISTANCES.

	Point1	**Point**2	**Point**4	**Point**5	**Point**6
*Hand Measured Distance From **Point**3*	0.200	0.150	0.150	0.156	0.250
*Deprojected Distance From **Point**3*	0.200	0.149	0.148	0.157	0.247
Difference - Deviaton	0.000	0.001	0.002	0.001	0.003

the camera. Table IV exhibits the deviation between grid-hand measurements and the code-calculated distance after the 3D deprojection of six highlighted points in the scene, the average difference-deviation is less than 1.8 millimeters.

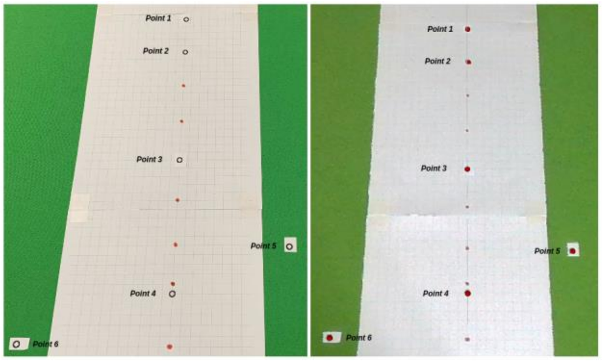

Fig. 4. Left: 2D Printed screws. Right: PointStamped messages after Deprojection in 3D space.

For visualization purposes, PointStamped ROS messages have been created for each point using the aforementioned 3D coordinates. ROS PointStamped messages have been used for visualization in ROS RVIZ. Figure 4 highlights the deprojection accuracy of six screws from 2D coordinate space to 3D.

VI. CONCLUSION

In this work, a novel architecture is introduced for enabling efficient disassembly of WEEE devices and components in a system-agnostic manner. Apart from WEEE disassembly, this system-agnostic architecture can be used for any machine learning application, such as assembly, defect detection, safety inspection, medical imaging, and other similar emerging applications and systems. Traditional CV techniques are replaced with a deep learning-assisted computer vision method, which provides a more generalizable, accurate, and faster approach than traditional techniques. To further boost research in the field, a large-scale screw dataset, significantly broader than most datasets in terms of annotated screws has been formed. The screw dataset is a significant contribution to the scientific community as well as the industry, and it can also be used for small object challenges. Finally, experimental results have been conducted demonstrating the effectiveness of the visual object detection and the robotic 3D deprojection modules. Future work includes the investigation of transferring the proposed solution in various robotic vision applications while increasing the cognitive capacity of cobots by employing advanced perception methodologies.

ACKNOWLEDGMENTS

This work was supported by the European Commission under contract H2020-820742 HR-Recycler.

REFERENCES

[1] Amit Kumar, M. Holuszko, and Denise Espinosa, “E-waste: An overview on generation, collection, legislation and recycling practices,” *Resources, Conservation and Recycling*, vol. 122, pp. 32–42, 07 2017.

[2] Apostolos Axenopoulos, Georgios Th. Papadopoulos, Dimitrios Giakoumis, Ioannis Kostavelis, Alexis Papadimitriou, Sara Sillaurren, Leire Bastida, Ozgur S. Oguz, Dirk Wollherr, Eugenio Garnica, Vasiliki Vouloutsi, Paul F.M.J. Verschure, Dimitrios Tzovaras, and Petros Daras, “A hybrid human-robot collaborative environment for recycling electrical and electronic equipment,” in *2019 IEEE SmartWorld*, 2019, pp. 1754–1759.

[3] Wei Hua Chen, Kathrin Wegener, and Franz Dietrich, “A robot assistant for unscrewing in hybrid human-robot disassembly,” in *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*, 2014, pp. 536–541.

[4] Ruiya Li, Duc Truong Pham, Jun Huang, Yuegang Tan, Mo Qu, Yongjing Wang, Mairi Kerin, Kaiwen Jiang, Shizhong Su, Chunqian Ji, Quan Liu, and Zude Zhou, “Unfastening of hexagonal headed screws by a collaborative robot,” *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1455–1468, 2020.

[5] Pablo Gil, Jorge Pomares, Santiago Puente, Carolina D ´ıaz Baca, Francisco Candelas Herias, and Fernando Medina, “Flexible multi-sensorial system for automatic disassembly using cooperative robots,” *International Journal of Computer Integrated Manufacturing*, vol. 20, pp. 757–772, 08 2007.

[6] Nicholas M. DiFilippo and Musa K. Jouaneh, “A system combining force and vision sensing for automated screw removal on laptops,” *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 887–895, 2018.

[7] Mohamad Bdiwi, Aquib Rashid, and Matthias Putz, “Autonomous disassembly of electric vehicle motors based on robot cognition,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2500–2505.

[8] Sergio Garrido-Jurado, Rafael Mu ñoz-Salinas, Francisco Madrid- Cuevas, and Manuel Marín-Jiménez, “Automatic generation and de- tection of highly reliable fiducial markers under occlusion,” *Pattern Recognition*, vol. 47, pp. 2280–2292, 06 2014.

[9] Zeromq, “libzmq,” https://github.com/zeromq/libzmq.

[10] IntelRealSense, “librealsense,” https://github.com/IntelRealSense/librealsense.

[11] Radu Bogdan Rusu and Steve Cousins, “3D is here: Point Cloud Library (PCL),” in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[12] Stanford Artificial Intelligence Laboratory et al., “Robotic operating system,” .

[13] Erenus Yildiz and Florentin W örgötter, “Dcnn-based screw detection for automated disassembly processes,” in *2019 15th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2019, pp. 187–192.

[14] Daniel P. Brogan, Nicholas M. DiFilippo, and Musa K. Jouaneh, “Deep learning computer vision for robotic disassembly and servicing applications,” *Array*, vol. 12, pp. 100094, 2021.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 740–755, Springer International Publishing.

[16] Kaiming He, Georgia Gkioxari, Piotr Doll ár, and Ross Girshick, “Mask r-cnn,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[17] Tsung-Yi Lin, Piotr Doll ár, Ross Girshick, Kaiming He, Bharath Har- iharan, and Serge Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.