

Image Processing: a Review on Industrial Applications

Michał Affek¹, Manuel Barusco², Diāna Dupļevska³, Kārlis Freivalds³, Gergely Hollósi⁴, Corinna Kofler⁵,
Michał Kopczyński¹, Ruba Almahasneh⁴, Philip Newton⁶, Guntis V. Strazds^{3,7}, Gian Antonio Susto⁸,
Gábor Szűcs⁴, and Marek S. Tatara¹

Abstract—Image processing has become a key technology in industrial automation, supporting quality control, robotic manipulation, and process monitoring. However, deploying deep learning-based vision systems on the factory floor remains challenging. Industrial datasets are often small, imbalanced, and expensive to annotate, while production environments differ substantially from the everyday imagery used to pretrain most vision models. This paper reviews recent advances in industrial image processing and summarizes practical experience from several use cases developed within the AIMS5.0 project. We examine methods for synthetic data generation, multi-camera tracking, and continual learning, with particular attention to their role in data-scarce and evolving production settings. The presented applications cover robotic manipulation, quality inspection, assembly monitoring, and data annotation. For robotic grasping, a hybrid pose estimation pipeline combines AI-based detection and segmentation with Iterative Closest Point fitting. It achieves accuracy comparable to state-of-the-art approaches while requiring only an object point cloud instead of a textured CAD model and offers substantially faster execution. A segmentation-based method for bottle socket detection also provides promising localization accuracy, although further improvements are needed to meet strict mechanical tolerances. For industrial inspection, we show that class balancing, moderate augmentation, and transfer learning improve semiconductor defect classification under severe class imbalance. We further introduce a CycleGAN-based domain adaptation approach that is effective when source and target domains differ in their label distributions. In aluminium corrosion detection, supervised segmentation clearly outperforms unsupervised anomaly detection. Experiments on luminaire assembly show that video action recognition performs well on familiar products but generalizes poorly to visually different variants, illustrating the limitations of vision-only approaches in high-mix, low-volume production. Finally, an iterative pseudo-labeling pipeline substantially improves object detection from a small initial set of annotated images. Based on these results, we identify several open challenges in this area, including data requirements, extreme class imbalance, reliable pseudolabel confidence estimation, automatic image-quality filtering, and scalable human-in-the-loop annotation.

Index Terms—Image processing, deep learning, industry 5.0, data annotation

¹ Faculty of Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland.

² Department of Information Engineering, University of Padova, Italy.

³ Institute of Electronics and Computer Science, Riga, Latvia.

⁴ Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics, Hungary.

⁵ KAI Kompetenzzentrum Automobil- und Industrieelektronik GmbH, Villach, Austria.

⁶ Signify, Eindhoven, The Netherlands.

⁷ University of Latvia, Latvia.

⁸ Department of Information Engineering, University of Padova, Italy.

I. INTRODUCTION

Image processing plays a crucial role in various industrial applications, enhancing efficiency, quality, and automation across different sectors. Categorizing the image processing tasks based on specific criteria helps optimize performance and decision-making. The speed of image processing is critical in industries such as manufacturing and autonomous vehicles, where real-time error detection or object recognition ensures operational efficiency and safety [1]. Real-time processing is particularly important in applications such as surveillance and autonomous driving, where delays can compromise results, while non-real-time processing is sufficient for tasks such as offline medical diagnostics [2] or historical data analysis. Also, in terms of risk [3], there are multiple image processing systems, where an important aim is to minimize the likelihood of wrong decisions. Finally, the magnitude of the impact of decisions varies across industries; for example, in medical imaging, a misclassification can have life-threatening consequences, while minor errors in e-commerce product labeling can only affect the user experience. Together, these factors influence the choice of algorithms, infrastructure, and execution strategies for industrial image processing systems [4].

The integration of image processing into real-time monitoring systems has transformed manufacturing where algorithms capable of edge detection and morphological operations enable immediate feedback on surface quality and defect detection, facilitating timely interventions like tool replacements. This capability not only improves the quality of produced items but also optimizes operational efficiency. One of the proven non-contact methods of Tool Condition Monitoring (TCM) is measuring systems based on image processing technologies that allow assessing the current state of the machined surface and the quantitative indicators of tool wear [5].

In the medical field, deep learning techniques have revolutionized image segmentation processes. These methods allow for precise analysis of medical images, significantly aiding in diagnostics and treatment planning [6].

Image processing helps many researchers around the globe to enhance their models in the agriculture domain. One of the possible tasks is plant classification where the aim is to classify images of plants to species, which is solved by deep learning, e.g. with a pretrained convolutional neural network and transfer learning [7].

Remote sensing is a critical field that involves detecting and

tracking the physical properties of an environment (typically from satellite or aircraft) by the calculation of reflected and emitted radiation from a radius. Many researchers around the world use image data augmentation to improve their models in the remote sensing domain [8].

Hyperspectral imaging poses major challenges for supervised classification methods due to the high dimensionality of the data; and there is an additional challenge to discover and distinguish novel categories for hyperspectral images in scenarios where some classes are unknown [9].

Object pose estimation is a key task in image processing, especially for robotic systems that need to perceive and interact with their environment. It involves determining an object's six degrees of freedom (6-DoF) pose – its 3D position and orientation – from visual input. Precise pose estimation is critical in industrial settings, enabling accurate manipulation and automation. While state-of-the-art approaches often focus on model-free methods requiring little or no prior object knowledge, this doesn't align well with industrial reality. In many industrial applications, high-quality Computer-Aided Design (CAD) models are readily available or easily obtained, and depth sensing is standard alongside color imaging. These practical advantages should be leveraged to develop more efficient, robust, and simpler pose estimation algorithms tailored for industrial use.

Automated image processing systems utilize deep learning algorithms to identify defects with high accuracy in semiconductor manufacturing. This area is characterized by high data complexity and data scarcity and neural networks can be pre-trained on industry-specific data, significantly reducing the amount of sample images required for effective training. This capability allows for real-time quality assurance that rivals human inspectors in accuracy while maintaining industrial speed standards. The state of the art in industrial image processing is characterized by sophisticated AI-driven solutions that enhance quality control, optimize production processes, and support predictive maintenance strategies, all contributing to more efficient manufacturing environments [10, 11].

The landscape of deep learning, particularly in computer vision, is rapidly evolving with the interplay between Convolutional Neural Networks (CNNs) and Transformers. Transformers have emerged as a dominant architecture, especially in Natural Language Processing (NLP) and increasingly in computer vision tasks. Their ability to handle global relationships through self-attention mechanisms has led to significant advancements, outperforming CNNs in various benchmarks for tasks like object detection and segmentation.

Vision transformers have become popular as a possible substitute to CNNs, with their ability to focus on global relationships in images, offer large learning capacity; however, they may suffer from limited generalization. Recently, in vision transformers hybridization of both the convolution operation and self-attention mechanism has emerged, to exploit both the local and global image representations. These hybrid vision transformers, also referred to as CNN-Transformer architectures, have demonstrated remarkable results. While transformers are making significant strides in various domains, the integration of CNNs into hybrid models represents a

promising avenue for enhancing performance in computer vision tasks [12].

Thus, image processing is integral to Industry 5.0, supporting smart manufacturing through automation and real-time data analysis. Combining image processing with robotics and IoT can further enhance industrial automation. In the next sections, applications of image processing in the industrial automation will be presented; these applications continue to expand with advancements in technology and integration with other industrial systems.

A. Synthetic Datasets for Vision

Synthetic datasets have become an increasingly valuable resource in machine learning, particularly in scenarios where collecting and annotating large-scale real-world data is costly or impractical. By leveraging simulation environments or generative models, researchers can produce diverse, labeled datasets with fine-grained control over variables such as lighting, pose, or background complexity.

Recent progress in generative AI has significantly broadened the toolkit available for synthetic dataset creation. Diffusion models, such as Stable Diffusion [13] and Imagen [14], allow for the generating high-resolution, photorealistic images from textual prompts via iterative denoising processes. These models are particularly effective for producing diverse samples. Text-to-image models, like DALL-E [15], leverage transformer-based architectures to generate coherent visuals directly from textual descriptions. These models are widely used to prototype visual scenarios and create labeled data by combining image synthesis with automated captioning or labeling pipelines. Image-to-image translation models, such as Pix2Pix [16] or ControlNet [17], enable the generation of synthetic variants based on structural or semantic constraints in input images. These are particularly useful in domains like medical imaging or autonomous driving, where realism and controlled variation are critical. For instance, diffusion models have recently been applied to generate synthetic organ-on-a-chip cell images, enhancing dataset diversity for biomedical classification tasks [18, 19]. In parallel, Vision-Language Models (VLMs) such as Flamingo [20] and GPT-4V [21] enable coordinated processing of visual and textual data, supporting both generation and interpretation tasks. These models can be used not only to synthesize data but also to provide annotations, questions, or captions-essentially automating the label generation process. Together, these tools enable scalable, diverse, and cost-effective synthetic data generation. However, they also introduce challenges related to domain shift, realism, and potential bias transfer from its training dataset. One way to reduce this domain shift is to combine visual domain randomization in simulation with a limited number of real images, which was shown to improve the robustness of the learned representations in sim-to-real transfer [22]. Despite these issues, synthetic datasets are increasingly essential in pre-training pipelines, domain adaptation, and robustness testing in computer vision.

In the Table 1, we present a selection of representative techniques and models used for synthetic dataset generation,

categorized by their underlying generation strategy. While this overview includes widely adopted tools across multiple domains, it is by no means exhaustive. The field is rapidly evolving, and many domain-specific, hybrid, or proprietary models continue to emerge, especially in medical imaging, robotics, and industrial inspection.

Evaluating the quality of synthetic datasets is essential to ensure their effectiveness and reliability in various applications. This evaluation encompasses three primary dimensions: fidelity, utility, and privacy. Fidelity assesses how closely synthetic data resembles real data; utility measures the usefulness of synthetic data for specific tasks; and privacy evaluates the risk of sensitive information leakage. Table II summarizes key metrics used to assess these dimensions, providing a comprehensive overview for practitioners and researchers.

B. Multi-Camera Solutions of Computer Vision in Industry

Multi-Camera Multi-Target (MCMT) tracking systems [51] [52] have a wide range of applications across various sectors [53], including general mobility, logistics [54] and industry [55–57]. For vehicle and people mobility, MCMT systems reconstruct complete trajectories, enhancing safety, sustainability, and efficiency. In industrial environments, MCMT enhances logistic operations, streamlines manufacturing processes, and improves worker safety. For the last aim, MCMT systems provide continuous surveillance of worker movements and activities across multiple camera feeds. This capability allows for immediate detection of unsafe behaviors or situations, enabling prompt intervention to prevent accidents in the real-time monitoring environments.

MCMT systems can identify potential hazards in the work environment, such as unsafe machinery operation or workers entering restricted areas [58]. This proactive monitoring helps mitigate risks before they escalate into incidents. In environments where workers and machinery operate in close proximity, MCMT systems can track both human and machine movements to prevent collisions. Alerts can be generated if a worker is detected in a dangerous zone, enhancing situational awareness and safety. The data collected by MCMT systems can be analyzed to identify patterns in worker behavior and safety incidents. This analysis enables organizations to implement targeted training programs and safety protocols based on real-world observations. MCMT systems can be integrated with other safety technologies, such as Internet of Things (IoT) devices that monitor environmental conditions (e.g., air quality, noise levels) and wearable sensors that track vital signs. The implementation of MCMT tracking systems in industrial settings not only enhances real-time safety monitoring but also fosters a culture of proactive risk management, ultimately leading to safer working environments for employees [59].

MCMT systems can be used in Intelligent Transportation Systems (ITS) [60] as well, where it utilizes data and technology to improve transportation networks, with applications that enhance traffic management, safety, and sustainability. MCMT tracking systems can be effectively utilized in logistics to optimize operations and enhance efficiency. As the latest results from the MCMT, a comprehensive vehicle tracking

system is developed that performs multiple tasks simultaneously, including tracking vehicles using a single-phase object detector, re-identifying vehicles, and classifying their attributes such as color and type. While re-identification networks and classification models operate independently, ID collisions may occur across cameras. To address this, MCT-MAC [61], a novel multi-camera trajectory matching algorithm was introduced based on hierarchical agglomerative clustering with time and space constraints, allowing trajectory aggregation across multiple cameras. This approach is flexible, supporting various trackers (ByteTrack and DeepSORT) as well.

MCMT systems can track the movement of goods within warehouses, providing real-time data on inventory levels and locations [54]. By monitoring the flow of both goods and personnel, MCMT can identify bottlenecks in the logistics process. This data allows for the optimization of workflows, ensuring that resources are utilized efficiently and that processes are streamlined. MCMT enables the tracking of vehicles transporting goods across multiple camera views, providing a comprehensive overview of shipment locations. This capability is crucial for maintaining supply chain transparency and improving delivery reliability. The analytics provided by MCMT systems allow logistics managers to make informed decisions based on real-time data regarding movement patterns, resource allocation, and operational efficiency. In automated warehouses, MCMT can work alongside robotic systems to monitor their movements and interactions with human workers, ensuring safe and efficient operations. Overall, the implementation of MCMT systems in industry not only enhances operational efficiency but also contributes to improved safety.

C. Continual Machine Learning

Processes enriched with image processing capabilities and decision-making schemes are frequently deployed in an industrial environment. Despite the automated process's ideation, implementation, and integration stages, there is also a step responsible for long-term management and process service. For Machine Learning (ML) models used in these processes, we can identify methods for partial retraining or updating the model as the dataset grows, which fall under the umbrella term Continual Machine Learning (CML).

The main challenge in building an ML solution in an industrial environment is that the solution does not scale up and thus is not easily reconfigurable without a significant amount of resources [62]. Reconfigurability can relate to the changes between machines, tools, processes, or the changes in the distribution of collected data. Concept drift is a data process that influences many implemented models over time. The continued learning methods were introduced to maintain or enhance the efficiency and reliability of solutions, and simultaneously spend fewer resources on retraining the model with the whole dataset. In principle, the methods aim to treat the data classes as equal without the occurrence of catastrophic forgetting. The mentioned phenomenon occurs when the adaptation to the new data without retraining on the whole dataset tends to lose the ability to capture older classes.

Handling catastrophic forgetting requires careful adoption of the CML techniques. Authors in [63] proposed image

TABLE I
OVERVIEW OF GENERATIVE APPROACHES FOR SYNTHETIC DATA CREATION WITH EXAMPLE MODELS.

Technique	Description	Selected Models / Tools
Text-to-Image	Generates realistic images from natural language prompts. Offers fine control over content, style, and scene composition, making it ideal for augmenting datasets or generating domain-specific samples.	DALL-E [15], Imagen [14], Parti [23], Midjourney [24]
Diffusion Models	Use iterative denoising to produce high-fidelity images from noise. These models are currently state-of-the-art in generative realism and are used both for unconditional and conditional generation tasks.	Stable Diffusion [13], DeepFloyd IF [25], GLIDE [26]
Image-to-Image Translation	Translates an input image from one domain to another (e.g., sketches to photos, semantic maps to realistic scenes). Supports controlled generation in highly structured domains.	Pix2Pix [16], CycleGAN [27], ControlNet [17]
Vision-Language Models (VLMs)	Process and relate visual and textual modalities. Useful for tasks like automatic captioning, label generation, question answering, and scene understanding in multimodal contexts.	Flamingo [20], GPT-4V [21], BLIP-2 [28]
Simulation Engines	Use physically accurate 3D environments to render synthetic scenes with automatic ground-truth labels. Widely used in robotics, automotive, and AR/VR settings for controlled dataset generation.	Unity Perception [29], CARLA [30], Blender [31], NVIDIA Omniverse [32]
3D-to-2D Rendering	Projects 3D assets or scans into annotated 2D images with varied poses, lighting, and camera perspectives. Enables generation of multi-view labeled datasets with consistent geometry.	BlenderProc [33], Meshcapade [34], SURREAL [35]
Semantic-to-Image	Synthesizes realistic images from segmentation masks or semantic layouts. Allows control over scene layout while enabling diversity in style and appearance.	SPADE [36], GauGAN [37]
Prompt-Based Annotation	Uses multimodal models to generate captions, tags, or region descriptions given a prompt. Reduces reliance on manual annotations in early-stage dataset development.	CLIP [38], GPT-4V [21], LLaVA [39]
Automatic Mask Generation	Produces high-quality segmentation masks or bounding boxes using pre-trained models. Useful for initializing label sets in detection and segmentation pipelines.	SAM [40], Grounding DINO [41]
Video Synthesis	Generates temporally coherent video sequences either from text, images, or other conditioning. Supports tasks like motion synthesis, animation, and synthetic video pretraining.	Gen-1 [42], Tune-A-Video [43], Video Diffusion [44]

processing methods such as alpha blending and histogram equalization to mitigate this issue. On the other hand, in the cross-modal image-text retrieval domain, authors in [64] proposed the CML method that utilizes the knowledge decomposition and replay, which imitates the knowledge decomposition and replay mechanisms in humans that are put in challenging environments. Besides the methods, researchers also work on more generalized frameworks. For example, [65] presents a differential geometry framework for functionally invariant paths, enabling flexible adaptation of neural networks for secondary tasks. Simultaneously, authors of the mentioned paper claim that the networks resulting from the application of their framework have increased sparsification of the network and can be portrayed as more robust to adversarial attacks. The authors present a representative example of the usage of CML in industrial applications in [66]. The paper describes the development of the framework for visual classification of the defects and demonstrates the approach to new data inspection and labeling with simultaneous aggregation of incoming data for efficient model updates.

II. POSE ESTIMATION OF OBJECTS FOR ROBOTIC SYSTEMS

A. Overview

Reconfigurable AI-based robotic systems require the overall tasks to be decomposed into smaller basic tasks. These basic

tasks are then the building blocks to more complex tasks that can be implemented.

One of the most basic topics in robotic systems is grasping capability, particularly when a robotic arm interacts with and grasps objects. Grasping is performed based on RGB-D images, which provide both color (RGB) and depth (D) information at the pixel level, thereby integrating visual and three-dimensional spatial data. From these images, estimation of the object pose in six degrees of freedom (6-DoF) is required to enable accurate and reliable manipulation.

Pose estimation of objects are performed basically in two different ways: by geometric fit, or using AI-based methods. Geometric fit uses the point cloud of the target object to fit to the reconstructed 3D point cloud of the scene, resulting in 6 degrees of freedom transformation. However, the fit can be problematic, since iterative algorithms developed for the problem can have many local minima resulting in a completely wrong fit. Modern AI-based methods try to enhance the classic methods by using neural network-based solutions. Mostly, these are based on networks trained on images having objects with ground-truth poses. However, ground-truth poses are hard to label, so commonly rendered images are used instead, which on the other hand causes problems in generalization. Common examples of AI-based solutions are PoseCNN [67], FoundationPose [68] or GDR-Net [69].

TABLE II
EVALUATION METRICS FOR SYNTHETIC DATASET QUALITY ACROSS FIDELITY, UTILITY, AND PRIVACY DIMENSIONS.

Dimension	Metric	Description
Fidelity	Fr�chet Inception Distance (FID)	Measures the distance between feature distributions of real and synthetic images using a pre-trained Inception network. Lower FID indicates higher similarity [45].
	Structural Similarity Index Measure (SSIM)	Assesses the similarity between two images based on luminance, contrast, and structure [46].
	Multi-Scale SSIM (MS-SSIM)	Extends SSIM by evaluating image similarity at multiple scales, capturing more perceptual information [47].
Utility	Precision and Recall	Evaluate the accuracy of object detection or segmentation tasks performed on synthetic data compared to real data [48].
	F1 Score	Harmonic mean of precision and recall, providing a balance between the two for classification tasks [48].
Privacy	Exact Match Score	Calculates the proportion of synthetic data records that exactly match real data records, indicating potential privacy risks [49].
	Row Novelty	Assesses the uniqueness of synthetic data records compared to real data, with higher novelty indicating better privacy preservation [49].
	Differential Privacy Metrics	Quantify the privacy guarantees provided by the data generation process, ensuring that individual data points cannot be re-identified [50].

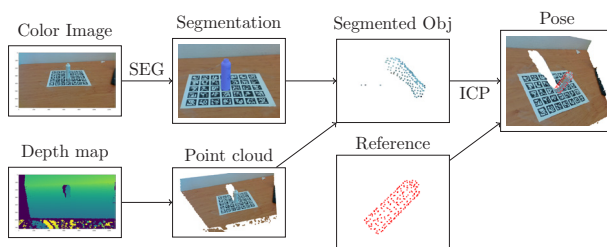


Fig. 1. The pipeline of the hybrid pose estimation algorithm. In the first step the color image is segmented using AI based image processing, then the Depth Map is clipped by the segments. The segmented Point Cloud is fitted using ICP to the reference object.

B. Methods

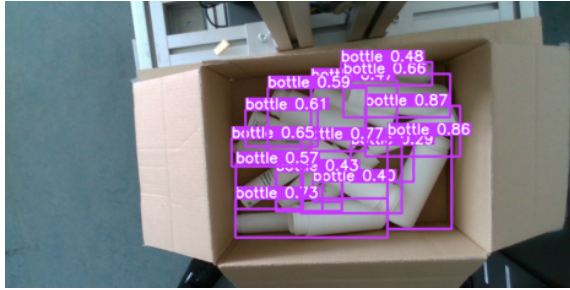
One particular disadvantage of existing pose estimation methods is their reliance on a highly detailed, textured 3D model of the object. This requirement can be a significant limitation, as the accuracy of the pose estimation process is highly dependent on the quality and completeness of the reference model. While this dependency may not pose a challenge in many industrial applications – where precise CAD models are readily available – there are numerous scenarios where obtaining or constructing such high-fidelity CAD models is impractical, time-consuming, or even infeasible. For instance, in applications involving objects that are irregular, subject to frequent changes, or lack pre-existing digital representations, generating detailed 3D models manually can be an unnecessary burden.

To overcome this issue, RGBD images provide an alternative approach by enabling geometry-based methods that align point clouds of objects with reconstructed point clouds of the surrounding environment. One of the most widely used techniques for this task is the Iterative Closest Point (ICP) algorithm ([70]), which iteratively refines the alignment between two point clouds by minimizing the distance between corresponding points. Although ICP has proven effective in

various applications, it suffers from several well-documented limitations that hinder its robustness and reliability, particularly for small or featureless objects. One major drawback is its susceptibility to local minimum, meaning that the algorithm is not guaranteed to converge to the optimal alignment, especially in the presence of noise, occlusions, or insufficient geometric features. Furthermore, the method struggles when significant portions of an object are occluded or outside the field of view – such as back-facing surfaces that are not captured in the depth image –further complicating the alignment process and reducing accuracy. As a result, ICP alone is often insufficient for precise (or even coarse) pose estimation, necessitating the development of more robust hybrid approaches or alternative algorithms that can compensate for these weaknesses.

Building on these findings, we propose a hybrid approach to pose estimation that enhances stability and feasibility. The method begins by identifying potential objects using well-established AI-based object detection and segmentation algorithms. Once the relevant objects are selected, ICP is applied exclusively to the corresponding 3D reconstructed points, significantly improving robustness and reliability. This approach eliminates the need for highly detailed CAD models, as only the point clouds of the objects are required, which are much easier to obtain. The complete pipeline is illustrated in Figure 1. Our method does not depend on a specific object detection and segmentation model – these components can be replaced with more suitable alternatives depending on the application.

For addressing a specific bottle pose estimation use case, the result of the first two steps of the pose estimation pipeline is depicted in Figure 2. In this example, object detection utilizes a RT-DETR model [71] and the segmentation applies a SAM model [40]. After having selected the proper object classes (e.g. in this case, bottles), based on the segmentation, the reconstructed 3D point cloud of the objects are cut from the point cloud of the environment, as shown in Figure 3. After fitting the model, the pose (location and orientation) of the



(a) Object detection results for the bottle objects (RT-DETR)



(b) Object segmentation result for the bottle objects (SAM)

Fig. 2. The first two steps of the pose estimation pipeline: object detection and segmentation



Fig. 3. The result of the point cloud fitting using ICP algorithm for one of the bottles

object is available from the ICP algorithm.

It is important to note that using the method, the poses of multiple objects can be done simultaneously – this can be considered as an advantage e.g., in situations, where robotic grasping requires collision objects to estimate.

Our solution is a hybrid method: it uses AI for segmenting the point cloud based on the RGB image, and after that, it uses iterative point cloud matching, resulting in good generalization, multiple object fit, and high performance.

C. Results

The proposed solution was compared to the state-of-the-art FoundationPose algorithm [68]. The measurement method is based on an Aruco marker-based ground-truth pose estimation to which both the proposed algorithm and the FoundationPose are compared (see Figure 4) [72].

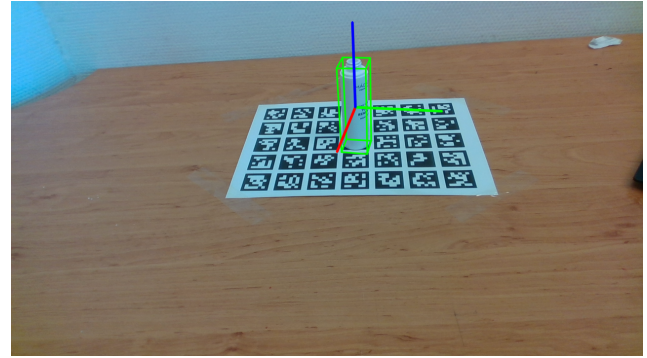


Fig. 4. Evaluating pose estimation utilizing Aruco markers to infer ground-truth poses.

Two separate measurements were conducted, each utilizing a sequence of 300 frames per algorithm. For evaluation, we calculated the translation error and the angle error, where the angle error is the angular difference between the Y axes, since the object was rotational symmetric. The first measurement uses a smooth and slow camera motion (scenario 1) while the second has much more harsh movements with some blurred images (scenario 2). The results can be seen in Figure 5 and Figure 6. Comparing the two algorithms, the FoundationPose has lost tracking in case of harsh movements, but the two methods provide similar accuracy in both angle and location estimation for the pose. However, the proposed solution does not require a textured CAD model but the point cloud for the object. Also, the one-shot speed is much faster, i.e. 100 milliseconds compared to the 2 seconds for the FoundationPose (registration state).

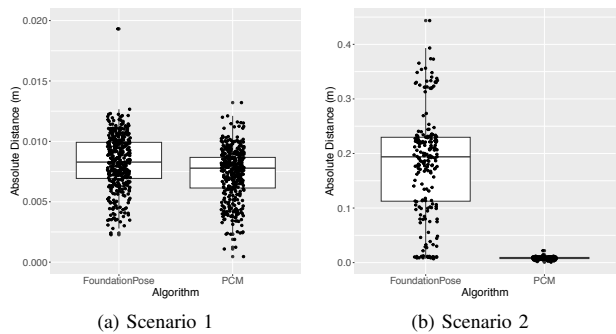


Fig. 5. Comparison of FoundationPose and the proposed solution in estimating the location part of the pose.

III. COMPUTER VISION FOR SOCKET DETECTION

In industrial environments, object detection models play a critical role in automating quality control, assembly verification, and robotic manipulation tasks. These models enable machines to detect and localize specific items, components, or conditions within visual data, significantly reducing the reliance on manual inspection and increasing throughput and consistency. The integration of computer vision systems into production lines helps streamline workflows, reduce human

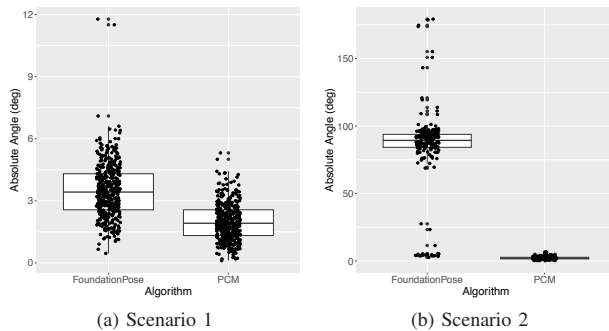


Fig. 6. Comparison of FoundationPose and the proposed solution in estimating the angle part of the pose.

error, and enable real-time decision-making, ultimately improving efficiency and product quality [73].

Modern object detection techniques leverage deep learning architectures trained on large datasets to recognize and localize objects within images or video frames. Among the most prominent object detectors is the YOLO (You Only Look Once) family of models, known for their speed and accuracy. YOLO divides the input image into a grid and simultaneously predicts bounding boxes and class probabilities for each region. This single-shot approach enables real-time performance, making it especially valuable for time-sensitive applications in manufacturing and robotics [74, 75]. The same model family is used for real-time detection in video outside manufacturing as well, for instance in surveillance [76].

In addition to YOLO, other well-known object detection models include SSD (Single Shot MultiBox Detector) and Faster R-CNN. SSD offers a balance between speed and accuracy, making it suitable for mobile or embedded devices [77], while Faster R-CNN provides higher precision at the cost of computational speed, often used in applications where accuracy is critical [78]. These models are typically trained on large-scale annotated datasets, such as COCO [79] or Pascal VOC [80], which contain thousands of labeled images across diverse object categories. However, in industrial scenarios, domain-specific datasets are often required due to the unique appearance of components and the structured nature of the environment - such environments typically involve repetitive patterns, consistent camera angles, and specific lighting conditions, which are not well represented in general-purpose datasets.

In this study, semantic segmentation was applied to a bottle socket detection task within a production context. Given the circular shape of the sockets, manual segmentation was used to create precise pixel-level labels distinguishing between empty and filled sockets. These high-quality annotations make it possible to train both segmentation and object detection models. However, the present results focus on the first segmentation-based experiment. The motivation for choosing semantic segmentation at this stage was to enable accurate estimation of the socket center - a key requirement for subsequent robotic tasks such as inserting a bottle into the socket.

Accurate localization of the socket center is especially critical

in industrial applications involving robotic manipulation. In bottle socket insertion tasks, even small positional errors may result in failed placement due to tight mechanical tolerances. Therefore, spatial accuracy is a key performance requirement in this study, and segmentation was chosen to achieve pixel-level precision in identifying the socket region and its center.

A. Methods

In this work, semantic segmentation was applied to distinguish between two classes: empty sockets and sockets containing bottles. This method involves assigning a class label to each individual pixel in the image, allowing for a highly detailed representation of object boundaries. As a result, it enabled the creation of precisely annotated training data, which were then used to train the YOLOv8m neural network. The model effectively learned to differentiate filled sockets (labeled as bottles) from empty ones (labeled as sockets).

To capture the necessary video data to train and evaluate the bottle socket detection system, an Intel RealSense D435 camera was utilized. The video footage was recorded at a resolution of 1280×720 pixels with a frame rate of 30 FPS. The camera was positioned approximately 8 cm above the conveyor belt and inclined at an angle of 10 to 15 degrees. Variations in this angle were observed across different recordings, indicating potential inconsistencies in camera positioning and highlighting the importance of improved stabilization methods.

A training dataset consisting of 252 images resized to 320×320 pixels was created from these videos. Only the region directly beneath the camera, offering the most perpendicular viewpoint, was included in the dataset to minimize perspective distortions. Manual segmentation was employed to distinguish between two classes: empty sockets (cyan) and sockets containing bottles (blue) (Figure 7). This manual labeling was particularly important due to the visual similarity between the white conveyor belt and the white bottles. The YOLOv8m neural network was trained using this dataset to enhance the detection accuracy of empty sockets.



Fig. 7. Example of image from socket detection dataset

During model training, both validation and test datasets were derived from footage captured with the Intel RealSense

D435 camera. These datasets were used to calculate standard performance metrics such as mean Average Precision (mAP).

Separately, we conducted an additional test using a Logitech Brio 4K camera mounted on an adjustable tripod. This camera provided higher resolution and greater control over capture parameters. The goal of this test was to evaluate how well the trained model performs when applied to data from a different camera. To support this evaluation, an ArUco marker was used for camera calibration and precise localization of the bottle socket.

The detection accuracy evaluation involved comparing the predicted socket center with the manually determined true center, identified as 4 cm horizontally from the center of the ArUco marker. The actual center was visually annotated with a pink cross using OpenCV, serving as a visual reference only.

B. Results

The detection accuracy of the model was quantified using the Root Mean Square Error (RMSE) under various experimental conditions, altering the camera resolution, distance, and angle. Thirty images were analyzed for each experimental scenario. The summarized results are presented in Table III.

TABLE III
ACCURACY RESULTS UNDER VARYING EXPERIMENTAL CONDITIONS

Resolution (px)	Distance (cm)	Angle (°)	RMSE (mm)
2560×1440	60	0	6.6
1280×720	80	0	4.2
1280×720	80	10	3.3

The results indicate that slightly increasing the camera angle improved detection accuracy, whereas higher resolution did not necessarily enhance performance.

To further assess the spatial accuracy of the empty socket detection model, an XY deviation analysis was performed, as illustrated in Figure 8. This analysis compares the predicted socket centers to the ground truth positions under consistent experimental conditions: a resolution of 1280×720 pixels, a camera height of 80 cm, and a tilt angle of 10°. Eight test scenarios were examined, each involving systematic rotation of the sockets to evaluate robustness across orientations. The observed positional deviations ranged from a minimum of 2.1 mm to a maximum of 3.9 mm, demonstrating the model's stable performance within a narrow margin of error.

In conclusion, the achieved RMSE of 3.3 mm in detecting the socket center initially suggests a relatively high model accuracy. However, considering the test socket diameter of 36 mm and bottle diameter of 32 mm, the available clearance is only 2 mm on each side. Since the average deviation exceeds this margin, it is likely that bottles will frequently fail to align properly with the socket, especially in the absence of guiding elements or corrective mechanics. This indicates that the current level of positioning precision is insufficient for reliable insertion under these conditions. To ensure dependable performance, improvements are necessary — either by increasing detection accuracy (e.g., through enhanced calibration or data refinement) or by modifying the mechanical setup, such as enlarging the socket diameter or adding alignment aids.

Pixel-wise Deviation of Detected Socket Center from True Center

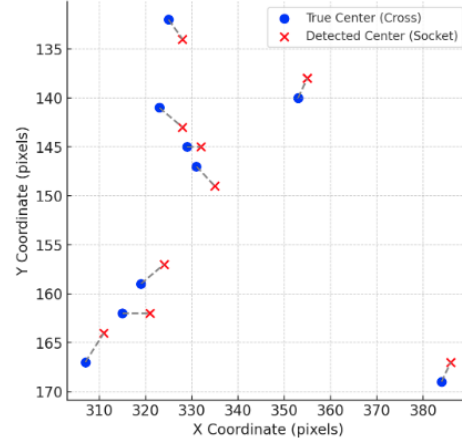


Fig. 8. Socket center comparison

Without such adjustments, the system may not consistently meet industrial requirements for automated insertion tasks.

IV. AI SOLUTIONS FOR SEMICONDUCTOR QUALITY CONTROL

A. Overview

In semiconductor production, ensuring the delivery of high-quality chips and products is of utmost importance due to the critical role semiconductors play in modern technology. To achieve this, extensive inspections and quality checks are conducted throughout the production process, many of them involve the analysis of images. Currently, these defect images are manually analyzed or classified by experts, a process that is time-consuming, labor-intensive, and prone to subjective biases. Automating these inspections through advanced methods such as defect image classification has become a central area of innovation to improve objectivity, efficiency, and scalability. Additionally, given the diverse domains within semiconductor production, which arise from variations in materials, processes, and equipment, it is essential to explore methods for domain adaptation. These techniques enable us to effectively transfer a model trained in one domain to another, addressing the challenge of domain-specific variations without requiring extensive retraining or additional labeled data. Furthermore, in cases where defects are rare or data labeling and collection is expensive, anomaly detection methods are applied to identify unusual patterns. By integrating these approaches, semiconductor manufacturers can enhance product quality, streamline operations, and contribute to sustainability by reducing waste and resource consumption, ultimately creating a more robust and efficient production pipeline. The following subsections will provide a more detailed description of the methods: defect image classification, domain adaptation, and anomaly detection, explaining their contributions to the semiconductor production process.

B. Defect Image Classification

Defect image classification may seem straightforward at first glance, given that the field has been around for quite some time. However, it is far from simple in practice. A review of the literature might suggest that the task can be easily accomplished by using one of the many pre-trained networks on public datasets, applying transfer learning to semiconductor production images, and evaluating the model's performance based on an overall metric, such as the F1-score. While this sounds efficient, the reality is much more complex. Even if we assume that the time-intensive tasks of data collection and labeling have already been completed, significant challenges remain. A major challenge is the highly imbalanced nature of datasets collected over a longer production period, as some defect classes occur less frequently than others. Such imbalances can bias the training process, leading to models that perform poorly on less frequent but critical defect types. Another key difficulty lies in the evaluation criteria: relying solely on overall metrics, such as accuracy or F1-score averaged across all classes, is insufficient for production settings. For critical defect classes, we aim to achieve an F1-score of at least 0.8, as these defects have a disproportionate impact on product quality. Additionally, the evaluation requires a detailed analysis, assessing performance not only at defect class level but also across defect images associated with different base materials. This level of granularity is essential since the dataset includes images originating from three distinct base materials. These additional requirements add layers of complexity to both model development and evaluation. In the following sections, we will describe the approaches we have taken to address these challenges.

Figure 9 illustrates the distribution of the defect image dataset. The y-axis represents the number of images, while the x-axis denotes the different defect classes, sorted in descending order by image count. The different colors in the figure indicate the base materials. Differentiating base materials is important, as defects from base materials 1 and 2 are very similar, while defects from base material 3 differ significantly. Additionally, certain defect classes are exclusive to base material 1 and 2, while others only occur on base material 3.

A closer examination reveals that the dataset is imbalanced in several ways. Most images come from base material 1, followed by base material 3, with very few from base material 2. However, since defects in base material 2 are similar to those in base material 1, the smaller number of images from base material 2 can be disregarded in this context. In contrast, the imbalance between base material 3 and the other two is significant, highlighting the need for more defect images from base material 3 to balance the dataset. Additionally, the dataset is imbalanced across defect classes, as some defects occur far less frequently during production than others.

To address these imbalances and ensure that the model does not become biased toward well-represented defect classes, we apply data balancing and data augmentation strategies to the dataset in the next step. Importantly, these data balancing and augmentation techniques are applied exclusively to the training

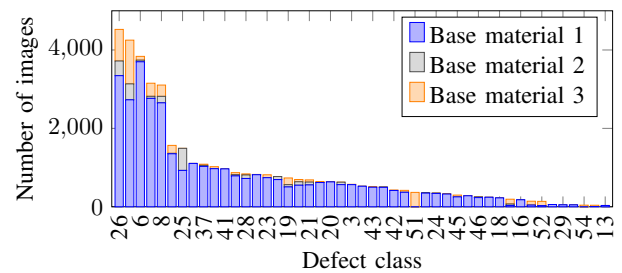


Fig. 9. Dataset distribution per base material

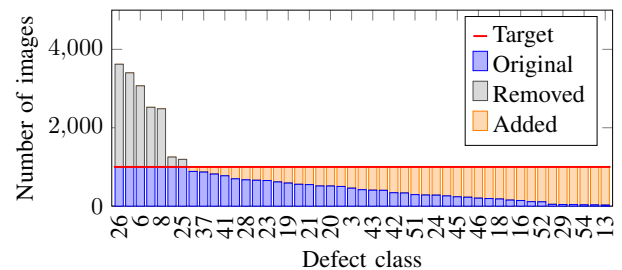


Fig. 10. Data-level balancing

images. The dataset is first split into separate training and validation sets. While the training set undergoes data balancing and augmentation, the validation set remains untouched. This ensures that the validation set reflects the original dataset distribution, allowing for an unbiased evaluation of the model's performance. By maintaining this separation, we can ensure that the results on the validation set accurately represent the model's real-world generalization ability.

These preprocessing steps, data balancing and data augmentation, are designed to improve the robustness of the model and enhance its ability to effectively learn from underrepresented defect classes, as discussed in publications [81–84]. Specifically, when working with datasets from semiconductor production similar to the one in this work, we found that applying data-level balancing in combination with moderate data augmentation yields the best results [85]. Figure 10 visualizes the resulting distribution of the defect image training dataset after resampling. The colors indicate the resampling techniques applied. To address class imbalance, we used a two-fold strategy: for overrepresented classes, excess images were randomly removed, and for underrepresented classes, the dataset was supplemented with replicated images. The blue bars represent the original images, gray bars show removed images from overrepresented classes, and orange bars indicate replicated images added for rare defect classes. The red line at 1,000 images marks the resampling target, determined experimentally as the optimal balance for this dataset.

Figure 11 illustrates examples of the augmentation techniques applied. These techniques are used to diversify the dataset and are dynamically applied to each image before being fed into the model for training. The augmentation methods include shearing, rotating, flipping, shifting, and zooming, applied within carefully defined moderate ranges to

avoid distorting defect characteristics, such as zooming in too much or shifting defects out of the image. During each training iteration, images are augmented with a random combination of transformations and settings within these ranges, ensuring improved variability during training.

For model training, we utilized transfer learning, leveraging the EfficientNetV2S architecture pre-trained on the ImageNet dataset [86]. The balanced and augmented dataset served as the input for fine-tuning the model. By combining resampling, augmentation, and transfer learning, we aimed to maximize the model's ability to generalize effectively while mitigating the challenges posed by the inherent dataset imbalances.

Figure 12 shows the model's performance for each defect class, grouped by base material. The y-axis represents the F1-score, while the x-axis lists defect classes sorted in descending F1-score order. Colors correspond to the base materials, following the same scheme as Figure 9. Transparent circles indicate defect classes with fewer than 20 validation samples, which are excluded from the analysis to ensure reliable evaluation. These classes require additional defect images for better representation. The red line at an F1-score of 0.8 marks the target performance for critical defect classes.

A detailed analysis of the model's performance reveals that the F1-score for certain defect classes falls below expectations, particularly for images associated with base material 3. This issue is primarily attributed to the limited availability of training images for this base material, which significantly affects the model's ability to generalize effectively. To address this, we have made the strategic decision to split the model into two separate ones: one dedicated to base materials 1 and 2, and another one for base material 3. This approach allows us to already use the model for base materials 1 and 2 in production, while iteratively collecting additional images and retraining the model for base material 3 in subsequent steps.

Over time, we will collect more defect images to fill up the underrepresented defect classes and further improve the performance of the models. However, the frequency of updates for the base material 1 and 2 model will be lower compared to the one of base material 3 model. With each base material 1 and 2 model update, we must ensure that the new version maintains or exceeds the performance of the previous version across all critical defect classes, as consistent reliability is crucial for productive usage.

These efforts aim to enhance automation in semiconductor manufacturing, improving efficiency, consistency, and scalability. Advancing these technologies helps reduce resource waste, improve production quality, and meet the growing demand for reliable, high-performance semiconductor products.

C. Domain Adaptation

Domain adaptation is a machine learning technique that transfers a model trained on one dataset (the source domain) to perform well on a different but related dataset (the target domain) where the data distribution is different because of a distribution shift. Instead of needing large amounts of new labeled data, domain adaptation allows models to adapt using

little or no labeled data from the target domain. This makes it especially useful in fields like defect classification, where collecting and labeling new data can be expensive and time-consuming. The two main domain adaptation scenarios are:

- Unsupervised Domain Adaptation (UDA): no target domain labels are available.
- Semi-Supervised Domain Adaptation (SSDA): some target domain samples are labelled and others are not.

The distribution shift in semiconductor manufacturing can occur, and Scanning Electron Microscope (SEM) images can vary significantly across different products, acquisition processes, or acquisition stages.

Domain adaptation methods have shown considerable promise in enabling scalable, robust defect classification across these variations, ensuring that models remain accurate without exhaustive retraining. Despite their potential, domain adaptation approaches have seen limited exploration and adoption compared to traditional defect detection and classification models. Moreover, this gap highlights a critical opportunity: leveraging domain adaptation can significantly mitigate the data scarcity issue, a persistent barrier in deploying machine learning solutions in manufacturing environments.

To tackle the distribution shift challenge in SEM images for defect classification, we developed a CycleGAN-based method named DBACS (DANN-Based Alignment with Cyclic Supervision) [87], which can work in both the previous domain adaptation scenarios.

The closed-source dataset used to develop the model comprises three data domains representing different products, with 2 defect classes: particle defects and point defects. Table IV summarizes the dataset distribution across the three considered domains. While Domain 0 and 1 exhibit a relatively balanced class distribution, Domain 2 presents a significant imbalance, with point defects being more than twice as frequent as particle defects.

TABLE IV
SAMPLE DISTRIBUTION PER CLASS AND DOMAIN IN THE CONSIDERED
CLOSED SOURCE DATASET

Domain	Particle Defects	Point Defects	Total
Domain 0	1706	1516	3222
Domain 1	639	503	1142
Domain 2	577	1196	1773
Total	2922	3215	6137

Examples of the SEM images are presented in Fig. 13

The purpose of the Cycle GAN is to transform an image from a given domain to another without altering the defects present in the image. The model has to change the defect's background while maintaining it unaltered. This is ensured by incorporating the Cyclic Supervision [88] mechanism of the Cycle GANs and additional losses that control the image transformation. As a result, a classifier trained on one domain can perform effectively in another by leveraging the image transformation performed by the Cycle GAN. We focus our experiments on the UDA and SSDA scenarios. Our method, DBACS, was compared with AdaMatch [89], a state-of-the-art domain adaptation method which can work in the UDA and SSDA scenarios.

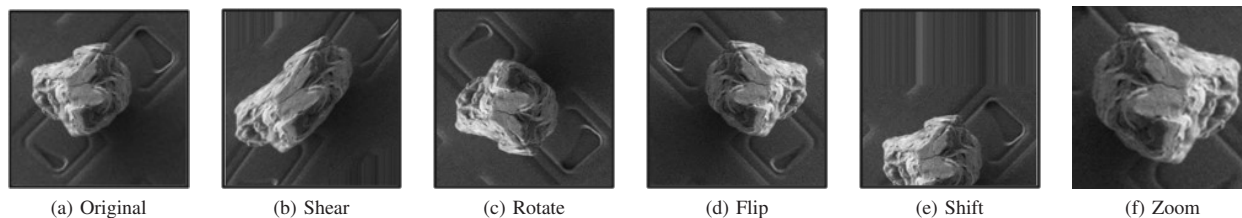


Fig. 11. Applied augmentation techniques

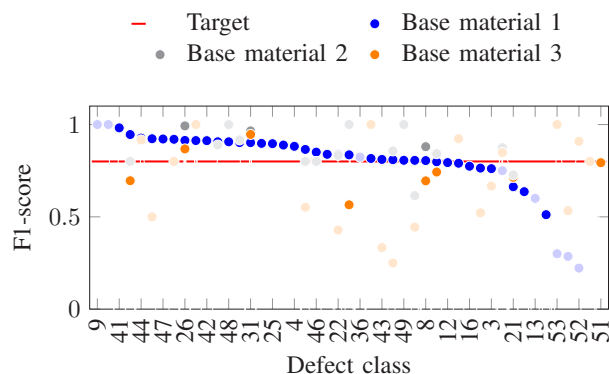


Fig. 12. Model evaluation based on the F1-score for each defect class in the validation dataset.

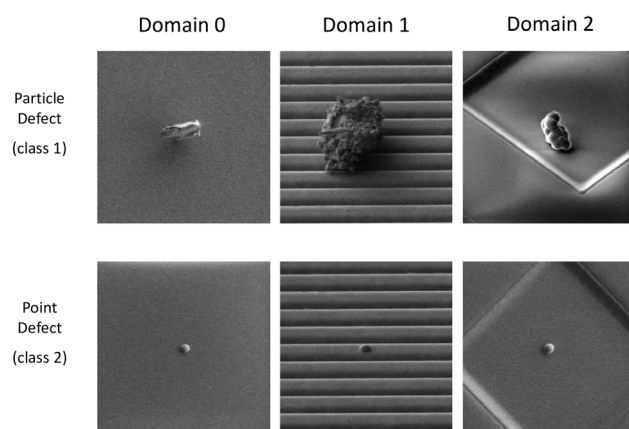


Fig. 13. We address the domain adaptation problem in image classification in the context of semiconductor manufacturing, considering two defect classes: particle defect and point defect. Each column represents a distinct domain.

The experiment results for the UDA scenario are shown in Tables V and VI. The results for the SSDA scenario are shown in Table VII. For every scenario, we tested three different networks: a MobileNet [90], a Resnet50 and a Resnet101 [91]. The three models are pre-trained on ImageNet and fine-tuned on the source domain. After that, the domain adaptation technique is applied, and the model is tested on the target domain test set. In every scenario, the Oracle is a model trained on the labelled target domain and is an upper bound for the performance. The reported values are the classification

accuracies on the test sets. The results show that AdaMatch performs well in all the considered scenarios and outperforms DBACS. However, in the UDA scenario with source domain 1, our method DBACS outperforms AdaMatch. This is caused by the entirely different label distributions between the considered source and target domains. AdaMatch, during its processes, considers this distribution equal and performs poorly in this case.

D. Anomaly detection

In semiconductor manufacturing, defect detection and classification remain challenging tasks due to the high complexity of wafer images and the inherent rarity of defects, which leads to significant data imbalance in training datasets. Anomalies such as aluminium corrosion or surface defects are often visually subtle, difficult to identify manually, and typically require domain expertise for accurate classification. Several state-of-the-art unsupervised anomaly detection methods, including PatchCore [92] and CFA [93], have been explored for this purpose. However, these approaches tend to fall short in high-complexity scenarios where defects are insidious and hard to distinguish from normal patterns.

A challenging problem in this context is the recognition of aluminium corrosion. Developing deep learning models capable of reliably detecting this type of defect is highly difficult. The datasets available for training are typically small, noisy, and feature complex images, exacerbated by a limited number of samples and severe class imbalance: common characteristics in semiconductor defect detection, where anomalies are rare, difficult to collect, and require expert labeling. Furthermore, the considered images are not SEM images but optical images with a very low resolution. In this setting, unsupervised anomaly detection approaches failed to achieve satisfactory results, demonstrating the limitations of unsupervised methods for complex defect types. As a result, a shift toward supervised learning techniques becomes necessary. A similar observation was made for anomaly detection on image representations of radio spectrum data, where autoencoder features gave the best results when combined with a supervised classifier [94].

In the broader field of computer vision, U-Net-based architectures [95] have proven highly effective for segmentation tasks across domains, including medical imaging and beyond. When adapted to anomaly detection in wafer images, the U-Net model can accurately detect and segment defects, pro-

TABLE V
UDA MODELS ACCURACY - SOURCE DOMAIN 0.

Source → Target	0 → 1			0 → 2		
Model	MobileNet	ResNet50	ResNet101	MobileNet	ResNet50	ResNet101
DBACS	79,3	72,4	80,2	64,53	61	65,64
AdaMatch	81,89	84,15	85,3	79	81,23	84,32
Oracle	97,3	85,8	98,2	95,9	81,1	97,1

TABLE VI
UDA MODELS ACCURACY - SOURCE DOMAIN 1.

Source → Target	1 → 0			1 → 2		
Model	MobileNet	ResNet50	ResNet101	MobileNet	ResNet50	ResNet101
DBACS	67,84	69,53	78,75	55,18	63,24	65,27
AdaMatch	56,65	60,78	65,74	52,86	54,33	60,24
Oracle	97,6	92,1	98,1	95,9	81,1	97,1

TABLE VII
SSDA (5% OF LABELED DATA IN THE TARGET DOMAIN) MODELS ACCURACY.

Source → Target	0 → 1			0 → 2		
Model	MobileNet	ResNet50	ResNet101	MobileNet	ResNet50	ResNet101
DBACS	81,67	82,74	82,86	70,03	77,87	79,83
AdaMatch	88,1	89,11	90,12	85,25	86,75	87,13
Oracle	97,3	85,8	98,2	95,9	81,1	97,1

TABLE VIII
DISTRIBUTION OF CLASSES IN THE DATASET

Class	Train	Validation	Test
Corrosion	273	35	34
No Corrosion	1391	175	172
Total	1664	210	206

TABLE IX
DISTRIBUTION OF LABELS IN THE TRAINING SET AFTER DATA AUGMENTATION

Corrosion	No Corrosion	Total
1365	1391	2756

viding enhanced model interpretability through explicit defect localization.

The approach involves training the model in a supervised way, using ground truth mask annotations that focus learning on defect patterns. Furthermore, a carefully designed data augmentation pipeline is applied during training to counteract class imbalance and improve robustness. This strategy significantly enhances performance on the aluminium corrosion recognition task, demonstrating that a supervised segmentation approach can outperform unsupervised methods in complex, data-scarce manufacturing environments.

Tables VIII and IX show the distribution of classes in the considered dataset and the distribution of labels in the training set after data augmentation. The data augmentation was done with vertical flip, horizontal flip, and transposition on both diagonals to images with corrosion defects. This ensures a balanced distribution in the dataset.

In Table X, a comparison of classification results (the Bal-

anced Accuracy Score is reported) and the number of trainable parameters of the different networks trained in a supervised way on the aluminium corrosion detection task is reported. The reported Shallow-CNN comprises 6 convolutional layers and a fully connected network on top, this model was trained from scratch on the considered dataset. The EfficientNetV2S considered was pretrained on the ImageNet dataset and then fine-tuned on our dataset. The ResNet101 models considered were pretrained on the ImageNet dataset or on the PCAM dataset and then fine-tuned on our dataset. Only the U-Net model can do segmentation; the other models were employed for simple classification. As can be seen from the results, the best model is the U-Net model, which outperforms the other classification-based methods and can also segment the input images, allowing model interpretability.

V. LUMINAIRE ASSEMBLY

One of the challenges in applying AI based computer vision in assembly is the cost/benefit tradeoff. Signify can manufacture and sell small volumes customized to specific requirements of a lighting project with short lead times. The variation in the products we manufacture, and the low volumes is at odds with traditional supervised machine learning approaches where AI algorithms learn from human feedback. In supervised machine learning the algorithm is trained to recognize parts on large datasets of human annotated data. Several innovations, such as fine tuning can reduce the amount of training data needed to achieve a high accuracy. But creating a dataset of several hundred to thousands of images costs time and money to create, especially if the trained algorithm does not generalize well to recognize assembly of functionally similar but different looking luminaires. One example we encountered recently was a trail in one of our factories where

TABLE X
COMPARISON OF CLASSIFICATION RESULTS AND NUMBER OF TRAINABLE PARAMETERS

	BACC	Precision	Recall	F1-Score	No. Parameters (in mln)
Shallow-CNN	50.0%	0.0%	0.0%	0.0%	20.8
EfficientNetV2S	93.0%	77.5%	91.2%	83.8%	20.3
ResNet101 (ImageNet)	89.8%	66.7%	88.2%	76.0%	42.7
ResNet101 (PCam)	95.9%	78.6%	97.1%	86.8%	42.7
U-Net	97.1%	86.8%	97.1%	91.7%	31.0

we aimed to recognize the timing of several steps in the assembly of an outdoor luminaire.



Fig. 14. Assembly process.

Here we mounted a camera above a workstation where the driver is installed in an outdoor luminaire. Fig. 14 shows an image from the camera. The goal is to determine the timings of each step in the assembly process. There are seven steps involved in mounting the driver and connecting the wires. A recording of approximately 45 minutes was taken and annotated using an open-source video annotation tool [96]. Fig. 15 shows an example. Once the entire video was annotated, we obtained 380 clips over seven classes.

To determine whether we could learn an action recognition model to identify the seven classes we fine-tuned a VideoMAE model. The model was already pre-trained on the Kinetics 400 dataset. The dataset contains 400 human action classes, with at least 400 video clips for each action. The clips are taken from different Youtube videos and cover a broad range of classes, such as playing instruments, shaking hands, dancing or doing yoga.

A challenge with using open-source models is that often they are trained on datasets outside of an industrial setting, this makes that fine-tuning may perform less well in practice when applied on datasets from industry vs academia. In academic papers the datasets used are more based on common everyday tasks, which aligns much more with what's present in the original dataset. Another challenge is the availability of well-structured data. Metadata on products is partially manually entered into a system. The reliability of this data depends on the human operator and humans are prone to making mistakes especially when trying to meet tight deadlines.

We finetuned the pre-trained model on our own dataset. The model was tested against 10% of the clips left out of the training set. The model performed quite well on this test set (82%). Later we collected more videos on the same line, where different sizes of the luminaire were being assembled with the same operations: mounting and connecting the driver. On those luminaires that had a different size, the model performed poorly. Visually the luminaires look different but to a human, the operations, are clearly similar. This is where AI vision-only models fail, as they rely on visual cues only and have no real semantic understanding of what is going on in the video. An approach that may prove more successful is a staged approach, where the parts and the movements of the operator are identified by several AI models, for example objection detection and hand tracking, and this is then fed to a third model to determine the assembly steps. But this would prove costly as it requires a more extensive dataset, for object detection and for training the assembly step model. We experimented with some newly released advanced multi-modal models to see whether these could play a role to reduce the cost associated with creating datasets. Zero-shot object detectors and multi-modal AI may play a role as they advance further and are trained on industrial datasets but for now these didn't work sufficiently well on our data to play a role.

More experimentation with an external company is ongoing but for cost benefit reasons it's important that any trained model generalizes well to other luminaires. A theoretical 20% percent improvement on efficiency of the line (which is very high) would be needed to come close to the costs associated with a commercial solution.

VI. SEMI-SUPERVISED APPROACH FOR INDUSTRIAL DATA ANNOTATION

A. Introduction

Manual data annotation remains a critical bottleneck in supervised machine learning, particularly in applications that require large-scale, domain-specific, or expert-curated labels. While high-quality annotations are essential for training performant models, the cost and time associated with manual labeling can exceed the foreseen return on investment. Automated annotation methods aim to alleviate this burden by leveraging machine learning models to generate labels with minimal human intervention. These techniques often combine elements of supervised, weakly supervised, and semi-supervised learning to bootstrap model performance and iteratively enhance label quality.

Semi-supervised learning (SSL) has gained considerable attention as a promising strategy for making use of vast



Fig. 15. Image annotation process.

amounts of unlabeled data. In typical SSL pipelines, a model is first trained on a small labeled dataset and then used to generate pseudo-labels for the remaining unlabeled data. These pseudo-labels, usually filtered based on confidence, are iteratively included in the training set to reinforce the model's learning. Methods such as consistency regularization [97], entropy minimization [98], and teacher-student paradigms [99] have shown remarkable success in reducing the reliance on labeled data while maintaining high accuracy. A comprehensive overview of semi-supervised techniques can be found in recent surveys [100, 101].

B. Processing Pipeline

The proposed annotation pipeline consists of four key stages. The first stage, *Initial Training*, involves training a model on a small set of clean, manually annotated data to establish a solid foundation for inference. In the second stage, *Semi-Supervised Learning*, the model is used to generate pseudo-labels for new, unlabeled data. These pseudo-labeled samples are then incorporated into the training process, allowing the model to learn from both the initial ground truth and its own confident predictions. This approach gradually improves the model's generalization capability as it is exposed to a broader data distribution.

Once performance has sufficiently improved, the pipeline proceeds to the *Re-Inference* phase, where the updated model re-labels the dataset. If the new labels are of higher quality, they are used for additional retraining, further reinforcing the model performance. This iterative process, core to the semi-supervised paradigm-enables the model to refine its own outputs over time, making increasingly accurate predictions without additional human annotation. Finally, in the *Results Acquisition* stage, the final labeled dataset is aggregated and

analyzed to evaluate label quality and overall model performance.

To quantify the effectiveness of the pipeline, we compare the mean Intersection over Union (IoU) scores [102] from the initial training stage and the end of the pipeline. As shown in Figure 17, the model's performance improves substantially, with the final mean IoU nearly doubling. This highlights the advantage of iterative pseudo-labeling and retraining in semi-supervised learning setups. The model used for this task was RT-DETR [103]. The dataset was an internal collection of images acquired from a battery production facility, with the objective of detecting batteries. It consisted of 100 images, all of which were manually annotated by our team.

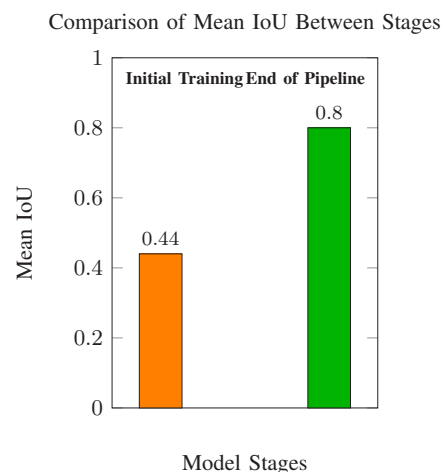


Fig. 17. Comparison of mean IoU scores between the initial training and final model stage.

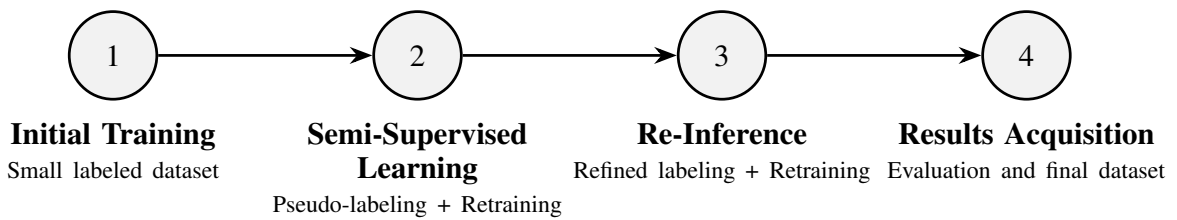


Fig. 16. High-level overview of the proposed semi-supervised annotation pipeline.

C. Summary

Semi-supervised annotation pipelines offer a scalable and cost-effective alternative to fully manual labeling by leveraging a small annotated seed dataset to iteratively train models and expand the labeled set using pseudo-labels. This method is particularly well-suited for domains where manual labeling is expensive or time-consuming. The effectiveness of such pipelines depends heavily on the quality of initial labels, pseudo-label confidence filtering, and the robustness of the underlying model. Modern approaches incorporating consistency regularization, self-training, and teacher-student frameworks continue to push the boundaries of what can be achieved with minimal supervision, opening new possibilities for fully automated or human-in-the-loop annotation systems.

VII. IDENTIFICATION OF THE RESEARCH GAPS FOR INDUSTRY4.0/5.0

A large part of adopting AI in industry revolves around data – collecting sufficient, domain-relevant samples; labeling and preprocessing them; ensuring long-term storage and maintenance; and managing large datasets over extended periods.

In defect image classification, an unresolved challenge is the effective handling of highly imbalanced datasets. Utilizing pretrained models, such as those developed on the ImageNet dataset, and applying transfer learning to specific domains like semiconductor images allows for the effective use of a limited number of images. However, handling classes with extremely few examples continues to pose a challenge. While techniques such as resampling and data augmentation can enhance the dataset, Generative Adversarial Networks (GANs) can also be employed to artificially generate new images. Nonetheless, these approaches require a baseline amount of existing data, which leads to the critical question of how much data is necessary per class to train a robust GAN or classification model. This question often necessitates empirical testing, as there is currently no established methodology or algorithm that clearly determines the optimal number of images required.

Despite recent advances in semi-supervised and active learning, automated annotation still faces significant obstacles. First, pseudo-labeling often propagates errors, especially in the early stages of training, leading to the reinforcement of incorrect annotations. Second, domain shift between labeled and unlabeled datasets can significantly degrade model performance if not properly addressed. Moreover, balancing model confidence with annotation quality remains an open issue—thresholds for pseudo-label acceptance are typically heuristic and task-dependent. Another challenge is the efficient selection

of informative samples for retraining, especially when data arrives in batches or from different sources. Also, integrating human-in-the-loop feedback in a scalable and seamless manner is still underexplored, especially in industrial or real-time contexts. These problems highlight the need for more robust, adaptive, and generalizable annotation strategies that can operate reliably across varying data distributions and tasks.

In addition to the quantity of data, the quality of the images is another vital consideration. It is essential to identify which images may be too noisy for inclusion in the training dataset. Furthermore, accurately labeling images can be challenging; borderline cases often result in different experts categorizing the same image into different classes, similar to how a dog might be confused for a cat. Therefore, it would be beneficial to develop methods that can automatically identify and remove low-quality images from such datasets, thereby improving the overall training process and model performance.

Looking ahead, several improvements appear promising. These include the development of more reliable confidence estimation mechanisms, dynamic thresholding strategies for pseudo-label acceptance, and domain adaptation techniques that can reduce the impact of dataset shift. We also aim to explore intelligent sampling methods for more effective retraining and tighter integration of human feedback loops, supported by intuitive interfaces and lightweight annotation tools. Ultimately, our goal is to make the annotation pipeline more autonomous, yet flexible enough to adapt to specific use cases and data modalities.

The future of deep learning in computer vision is likely to move toward more integrated approaches that combine CNNs and Transformers. Hybrid architectures will continue to be explored, alongside growing interest in model interpretability, real-time performance, and expanded applications in fields such as agriculture, autonomous vehicles, and environmental monitoring.

ACKNOWLEDGMENT

The research leading to these results was conducted in the AIMS5.0 project, which received funding from the Chips Joint Undertaking under grant agreement No. 101112089 and from the partners' national funding programmes.

REFERENCES

- [1] K. Han, H. Pei, R. Huai, T. Huang, and S. Qin, "Design of high-speed image processing system for weakdim target based on fpga," in *2022 7th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2022, pp. 783–789.

Image Processing: a Review on Industrial Applications

- [2] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.
- [3] S. Hussain, I. Mubeen, N. Ullah, S. S. U. D. Shah, B. A. Khan, M. Zahoor, R. Ullah, F. A. Khan, and M. A. Sultan, "Modern diagnostic imaging technique applications and risk factors in the medical field: a review," *BioMed research international*, vol. 2022, no. 1, p. 5 164 970, 2022.
- [4] Y. Li, "Research and application of deep learning in image recognition," in *2022 IEEE 2nd international conference on power, electronics and computer applications (ICPECA)*. IEEE, 2022, pp. 994–999.
- [5] D. Y. Pimenov, L. R. da Silva, A. Ercetin, O. Der, T. Mikolajczyk, and K. Giasin, "State-of-the-art review of applications of image processing techniques for tool condition monitoring on conventional machining processes," *The International Journal of Advanced Manufacturing Technology*, vol. 130, no. 1, pp. 57–85, 2024.
- [6] M. E. Rayed, S. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, p. 101 504, 2024.
- [7] B. P. Gyires-Tóth, M. Osváth, D. Papp, and G. Szűcs, "Deep learning for plant classification and contentbased image retrieval," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 88–100, 2019.
- [8] N. E. Khalifa, M. Loey, and S. Mirjalili, "A comprehensive survey of recent trends in deep learning for digital images augmentation," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2351–2377, 2022.
- [9] R. Mao, M. Al-Shouha, and G. Szűcs, "Spectral generalized category discovery by training on combined labels," in *2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 2024, pp. 000 397–000 402.
- [10] B. Tang, L. Chen, W. Sun, and Z.-k. Lin, "Review of surface defect detection of steel products based on machine vision," *IET Image Processing*, vol. 17, no. 2, pp. 303–322, 2023.
- [11] R. Archana and P. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artificial Intelligence Review*, vol. 57, no. 1, p. 11, 2024.
- [12] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and their cnn-transformer based variants," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 2917–2970, 2023.
- [13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CVPR*, 2022.
- [14] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, 2022.
- [15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *ICML*, 2021.
- [16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.
- [17] L. Zhang, L. S. Manevitz *et al.*, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.
- [18] M. Ivanovs, L. Leja, K. Zviedris, R. Rimsa, K. Narbutė, V. Movčana, F. Rūmnieks, A. Strods, K. Gillois, G. Mozolevskis *et al.*, "Synthetic image generation with a fine-tuned latent diffusion model for organ on chip cell image classification," in *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2023, pp. 148–153.
- [19] V. Movčana, A. Strods, K. Narbutė, F. Rūmnieks, R. Rimša, G. Mozolevskis, M. Ivanovs, R. Kadiķis, K. G. Zviedris, L. Leja *et al.*, "Organ-on-a-chip (ooc) image dataset for machine learning and tissue model evaluation," *Data*, vol. 9, no. 2, p. 28, 2024.
- [20] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.
- [21] OpenAI, "Gpt-4 with vision (gpt-4v)," <https://openai.com/research/gpt-4v-system-card>, 2023.
- [22] A. Béres and B. Gyires-Tóth, "Enhancing visual domain randomization with real images for sim-to-real transfer," *Infocommunications Journal*, vol. XV, no. 1, pp. 15–25, 2023.
- [23] J. Yu *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, 2022.
- [24] Midjourney, "Midjourney ai," 2022, <https://www.midjourney.com>.
- [25] DeepFloyd Team, "Deepfloyd if: A modular text-to-image diffusion model," 2023, <https://deepfloyd.github.io/IF>.
- [26] A. Nichol and P. Dhariwal, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.
- [27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *ICCV*, 2017.
- [28] J. Li *et al.*, "Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.
- [29] Unity Technologies, "Unity perception package," 2021, <https://github.com/Unity-Technologies/com.unity.perception>.
- [30] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," *CoRL*, 2017.
- [31] Blender Foundation, "Blender - open source 3d creation suite," 2022, <https://www.blender.org>.
- [32] NVIDIA, "Nvidia omniverse," 2023, <https://developer.nvidia.com/nvidia-omniverse>.
- [33] M. Denninger *et al.*, "Blenderproc: Procedural blender pipeline for photorealistic scene generation," 2019, <https://github.com/DLR-RM/BlenderProc>.
- [34] Meshcapade GmbH, "Meshcapade - realistic human models," 2023, <https://meshcapade.com>.
- [35] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," *CVPR*, 2017.
- [36] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *CVPR*, 2019.
- [37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *CVPR*, 2018.

- [38] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *ICML*, 2021.
- [39] H. Liu *et al.*, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.
- [40] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: <https://arxiv.org/abs/2304.02643>
- [41] S. Liu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *CVPR*, 2023.
- [42] Runway, "Gen-1: Video generation from a single image," 2023, <https://research.runwayml.com/gen1>.
- [43] J. Z. Wu *et al.*, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," *arXiv preprint arXiv:2212.11565*, 2022.
- [44] Anonymous, "Video diffusion models," 2023, <https://video-diffusion.github.io>.
- [45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: <https://arxiv.org/abs/1706.08500>
- [46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [47] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The 37th Asilomar Conference on Signals, Systems & Computers*, vol. 2, 2003, pp. 1398–1402.
- [48] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020. [Online]. Available: <https://arxiv.org/abs/2010.16061>
- [49] A. M. Alaa, B. van Breugel, E. Saveliev, and M. van der Schaar, "How faithful is your synthetic data? samplelevel metrics for evaluating and auditing generative models," *arXiv preprint arXiv:2102.08921*, 2021. [Online]. Available: <https://arxiv.org/abs/2102.08921>
- [50] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.
- [51] T. I. Amosa, P. Sebastian, L. I. Izhar, O. Ibrahim, L. S. Ayinla, A. A. Bahashwan, A. Bala, and Y. A. Samaila, "Multi-camera multi-object tracking: A review of current trends and future advances," *Neurocomputing*, vol. 552, p. 126 558, 2023.
- [52] X. Yang, J. Ye, J. Lu, C. Gong, M. Jiang, X. Lin, W. Zhang, X. Tan, Y. Li, X. Ye *et al.*, "Box-grained reranking matching for multi-camera multi-target tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3096–3106.
- [53] D. Papp, D. Lovas, and G. Szűcs, "Object detection, classification, tracking and individual recognition for sea images and videos." in *CLEF (Working Notes)*, 2016, pp. 525–533.
- [54] J. Rutinowski, H. Youssef, A. Gouda, C. Reining, and M. Roidl, "The potential of deep learning based computer vision in warehousing logistics," *Logistics Journal: Proceedings*, vol. 2022, no. 18, 2022.
- [55] D. Chiaravalli, G. Palli, R. Monica, J. Aleotti, and D. L. Rizzini, "Integration of a multi-camera vision system and admittance control for robotic industrial depalletizing," in *2020 25th IEEE international conference on emerging technologies and factory automation (ETFA)*, vol. 1. IEEE, 2020, pp. 667–674.
- [56] S. Wang, H. Sheng, Y. Zhang, D. Yang, J. Shen, and R. Chen, "Blockchain-empowered distributed multicamera multitarget tracking in edge computing," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 1, pp. 369–379, 2023.
- [57] X. Gong, X. Tan, and Y. Xiang, "Contrastive mean teacher for intra-camera supervised person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [58] Y. Jeon, A. S. Kulinan, T. Kim, M. Park, S. Park *et al.*, "Vision-based motion prediction for construction workers safety in real-time multi-camera system," *Advanced Engineering Informatics*, vol. 62, p. 102 898, 2024.
- [59] J. Kim, W. Shin, H. Park, and J. Baek, "Addressing the occlusion problem in multi-camera people tracking with human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5463–5469.
- [60] L. Fei and B. Han, "Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review," *Sensors*, vol. 23, no. 8, p. 3852, 2023.
- [61] G. Szűcs, R. Borsodi, and D. Papp, "Multi-camera trajectory matching based on hierarchical clustering and constraints," *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 44 879–44 902, 2024.
- [62] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024.
- [63] J. Orgil, S. Karungaru, and K. Terada, "Continual learning: linear layer classifier concatenation using image processing transform functions," in *Fifteenth International Conference on Quality Control by Artificial Vision*, vol. 11794. SPIE, 2021, pp. 338–343.
- [64] R. Yang, S. Wang, H. Zhang, S. Xu, Y. Guo, X. Ye, B. Hou, and L. Jiao, "Knowledge decomposition and replay: A novel cross-modal image-text retrieval continual learning method," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6510–6519.
- [65] G. Raghavan, B. Tharwat, S. N. Hari, D. Satani, R. Liu, and M. Thomson, "Engineering flexible machine learning systems by traversing functionally invariant paths," *Nature Machine Intelligence*, vol. 6, no. 10, pp. 1179–1196, 2024.
- [66] W. Sun, R. Al Kontar, J. Jin, and T.-S. Chang, "A continual learning framework for adaptive defect classification and inspection," *Journal of Quality Technology*, vol. 55, no. 5, pp. 598–614, 2023.
- [67] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," 2018. [Online]. Available: <https://arxiv.org/abs/1711.00199>

Image Processing: a Review on Industrial Applications

- [68] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," 2024. [Online]. Available: <https://arxiv.org/abs/2312.08344>
- [69] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdrnet: Geometry-guided direct regression network for monocular 6d object pose estimation," 2021. [Online]. Available: <https://arxiv.org/abs/2102.12145>
- [70] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [71] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on realtime object detection," 2024. [Online]. Available: <https://arxiv.org/abs/2304.08069>
- [72] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and M. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320314000235>
- [73] Z.-Q. Zhao, P. Zheng, S. tao Xu, and X. Wu, "Object detection with deep learning: A review," 2019. [Online]. Available: <https://arxiv.org/abs/1807.05511>
- [74] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. [Online]. Available: <https://arxiv.org/abs/1506.02640>
- [75] G. Jocher, A. Chaurasia, J. Qiu *et al.*, "Yolov8: Cuttingedge object detection models," 2023, <https://github.com/ultralytics/ultralytics>.
- [76] R. Mustafovski, T. Shuminoski, and A. Risteski, "YOLOv8-based firearm detection for real-time video surveillance application," *Infocommunications Journal*, vol. XVIII, no. 2, pp. 72–78, 2026.
- [77] 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*. Springer International Publishing, 2016, p. 21–37. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_2
- [78] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: <https://arxiv.org/abs/1506.01497>
- [79] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: <https://arxiv.org/abs/1405.0312>
- [80] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010. [Online]. Available: <http://dblp.uni-trier.de/db/journals/ijcv/ijcv88.html#EveringhamGWWZ10>
- [81] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri, "The imbalanced training sample problem: Under or over sampling?" in *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*. Springer, 2004, pp. 806–814.
- [82] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem," *Int. J. Advance Soft Compu. Appl.*, vol. 5, no. 3, pp. 176–204, 2013.
- [83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [84] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [85] C. Kofler, C. A. Dohr, J. Dohr, and A. Zernig, "Datacentric model development to improve the cnn classification of defect density sem images," in *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2022, pp. 1–6.
- [86] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.
- [87] A. Poniatowski, N. Gentner, M. Barusco, D. D. Pezze, S. Salti, and G. A. Susto, "Domain adaptation for image classification of defects in semiconductor manufacturing," *IEEE Transactions on Automation Science and Engineering*, vol. 23, pp. 3818–3828, 2026.
- [88] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.
- [89] D. Berthelot, R. Roelofs, K. Sohn, N. Carlini, and A. Kurakin, "Adamatch: A unified approach to semisupervised learning and domain adaptation," 2022. [Online]. Available: <https://arxiv.org/abs/2106.04732>
- [90] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [92] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," 2022. [Online]. Available: <https://arxiv.org/abs/2106.08265>
- [93] S. Lee, S. Lee, and B. C. Song, "Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," 2022. [Online]. Available: <https://arxiv.org/abs/2206.04325>
- [94] S. L. Takács and K. Kajdy, "Convolutional and variational autoencoders with supervised classifiers for RF spectrum anomaly detection," *Infocommunications Journal*, vol. XVIII, no. 2, pp. 55–60, 2026.
- [95] O. Ronneberger, P. Fischer, and T. Brox, "Unet: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: <https://arxiv.org/abs/1505.04597>
- [96] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019. [Online]. Available: <https://doi.org/10.1145/3343031.3350535>
- [97] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, and H. Zhang, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

- [98] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2005.
- [99] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
- [100] J. E. Van Engelen and H. H. Hoos, "A survey on semisupervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2020.
- [101] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semisupervised learning algorithms," in *Advances in neural information processing systems*, vol. 31, 2018.
- [102] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," in *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2. Springer, 2010, pp. 303–338.
- [103] Y. Zhao *et al.*, "Rt-detr: Real-time detection transformer," *arXiv preprint arXiv:2304.08069*, 2023.