Ref. Ares(2026)6579208 - 30/06/2026

MAGICIAN
AUTONOMOUS DEFECTS DETECTION AND REPAIR IN MANUFACTURING

PROJECT NO

101120731

PROJECT ACRONYM

MAGICIAN

PROJECT TITLE:

IMMERSIVE LEARNING FOR IMPERFECTION DETECTION AND REPAIR THROUGH HUMAN-ROBOT INTERACTION

CALL/TOPIC:

HORIZON-CL4-2022-DIGITAL-EMERGING-02-07

START DATE OF PROJECT:

01.10.2023

DURATION:

48 MONTHS

DELIVERABLE D3.3

Third delivery of perception systems

DUE DATE OF DELIVERABLE:

30.06.2026

ACTUAL SUBMISSION DATE:

30.06.2026

Work Package	**WP3 - Data acquisition and skills learning**
Associated Task	**T3.2 - Perception System, data acquisition and processing**
Deliverable Lead Partner	**IIT**
Main author(s)	**Ammar Qammaz, Luigi Palopoli, Iason Oikonomidis, Antonis Argyros, Annemarijn Blom, Matteo Dalle Vedove, Elena Basei, Enrico Turco, Gionata Salvietti, Domenico Prattichizzo, Nikos Tsagarakis**
Internal Reviewer(s)	**Daniele Fontanelli**
Version	**1.0**

DISSEMINATION LEVEL		
PU	Public	X
SEN	Sensitive - limited under GA conditions	

CHANGE CONTROL

DOCUMENT HISTORY

VERSION	DATE	CHANGE HISTORY	AUTHOR(S)	ORGANISATION
0.1	15.06.2026	First Draft	Gionata Salvietti, Ammar Qammaz, Luigi Palopoli, Iason Oikonomidis, Antonis Argyros, Annemarijn Blom, Matteo Dalle Vedove, Elena Basei, Enrico Turco, Domenico Prattichizzo, Nikos Tsagarakis	IIT, FORTH, UNITN, PIPPLE

0.2	26.06.2026	Internal Review	Daniele Fontanelli	UNITN
0.3	30.06.2026	Final Version	Ammar Qammaz, Iason Oikonomidis, Antonis Argyros, Gionata Salvietti.	IIT, FORTH
0.4	30.06.2026	Quality check	Annalisa Baldo	UNITN
1.0	30.06.2026	Submitted version	Gionata Salvetti, Daniele Fontanelli	IIT, UNITN

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

This deliverable is part of a project that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101120731.

EXECUTIVE SUMMARY

This deliverable presents, building upon the developments documented in Deliverables D3.1 and D3.2, the substantial progress achieved in the design, implementation, integration, and validation of the perception technologies that enable autonomous defect detection, human-aware operation, and knowledge transfer from expert operators to robotic systems.

The work described in this deliverable covers three complementary research directions. First, the visual perception system has been significantly refined through the development of a new industrial-grade sensing platform, improvements in optical design and illumination, expanded data acquisition and annotation procedures, and extensive evaluation of deep-learning-based defect classification methods. The resulting system demonstrates improved robustness, repeatability, and real-time performance under representative industrial conditions while providing reliable detection and classification of manufacturing imperfections.

Second, the tactile perception system has evolved into a complete multimodal sensing solution capable of complementing visual inspection through contact-based exploration. The developed tactile module integrates force/torque sensing, vibration measurements, and compliant mechanical interfaces to characterize surface defects that may be difficult to detect visually. Dedicated data acquisition procedures, together with machine learning models specifically designed for tactile information, have demonstrated accurate defect classification while also providing valuable insights into the influence of scanning strategies, contact conditions, and operator variability.

The deliverable also reports significant advances in human motion perception and prediction. The proposed framework enables robust human pose estimation, motion analysis, and contextual scene understanding using RGB sensing, supporting safe and efficient human-robot collaboration in industrial environments. Furthermore, learning methodologies based on human demonstrations have been further developed to facilitate the transfer of expert inspection knowledge into robotic systems, contributing to the long-term objective of autonomous defect inspection and handling.

A major achievement of the reporting period has been the successful integration of the visual, tactile, and human perception modules into a unified perception framework operating on the MAGICIAN robotic platform. The integrated system has been deployed and validated during project integration activities in representative industrial environments, demonstrating interoperability between hardware and software components, real-time operation, and compatibility with the planning and control modules developed in other work packages. These integration activities mark the transition from individual laboratory prototypes to an operational perception ecosystem suitable for system-level demonstrations. Notice that this document complements the integrated prototype, that is also part of the deliverable and currently operating in ALT.

Overall, the results presented in this deliverable demonstrate the maturity reached by the MAGICIAN perception technologies and their readiness to support the final project

validation activities. The developed perception framework provides reliable multimodal sensing, robust defect classification, human-aware perception, and the foundations for learning inspection skills from expert operators. These achievements constitute a key technological contribution toward the MAGICIAN vision of intelligent, trustworthy, and human-centred robotic systems capable of supporting autonomous defect detection and repair in industrial manufacturing environments.

DEVIATIONS

No deviation is foreseen from the planned path.

TABLE OF CONTENTS

1 INTRODUCTION 16
1.1 PURPOSE AND SCOPE 17
1.2 CONTRIBUTION TO PROJECT OBJECTIVES 18
1.3 RELATION TO OTHER WORK PACKAGES 19
1.4 STRUCTURE OF THE DOCUMENT 20
1.5 PERCEPTION SYSTEM OVERVIEW 20
1.5.1 Requirements and Specifications 20
1.5.2 System Architecture 21
2 VISUAL PERCEPTION FOR IMPERFECTIONS DETECTION 22
2.1 INTRODUCTION 24
2.2 OBJECTIVES AND REQUIREMENTS 26
2.3 CAMERA SYSTEM 27
2.4 DATA ACQUISITION AND ANNOTATION 34
2.4.1 DATA ACQUISITION 34
2.4.2 DATA ANNOTATION 37
2.5 EMPLOYED METHODOLOGIES 45
2.6 RESULTS AND FINDINGS 50
2.7 CHALLENGES AND LIMITATIONS 57
3 TACTILE PERCEPTION SYSTEM FOR IMPERFECTIONS DETECTION 59
3.1 INTRODUCTION 59
3.2 STATE OF THE ART 59
3.3 OBJECTIVES AND REQUIREMENTS 60
3.4 TACTILE SENSORS 61
3.5 DATA ACQUISITION AND ANNOTATION 62
3.6 METHODOLOGIES EMPLOYED 66
3.7 RESULTS AND FINDINGS 68
3.8 CHALLENGES AND LIMITATIONS 73
4 HUMAN MOTION PERCEPTION 74
4.1 INTRODUCTION 74
4.2 OBJECTIVES AND REQUIREMENTS 74
4.3 METHODOLOGIES EMPLOYED 76
4.3.1 Techniques for Pose detection 76

4.3.1.1 DATA ACQUISITION AND ANNOTATION 77
4.3.1.2 HUMAN POSE DETECTION METHOD 79
4.3.1.3 SENSING HUMANS AND THEIR CONTEXT 83
4.3.1.4 RESULTS AND FINDINGS 89
5 CONCLUSIONS 91
6 REFERENCES 92

LIST OF TABLES

Table 1: KPIs related to the Visual Perception for Imperfections Detection. 26
Table 2: KPIs related to the Tactile Perception System for Imperfections Detection. 61

LIST OF FIGURES

Figure 1: The Hardware & Software prototypes of the perception system integrated with the Doosan robot during the MAGICIAN 6th integration meeting in ALT premises. 17
Figure 2: The mock-up prototype V1 (a) of the video sensor created to test various optics configurations and bootstrap the development of perception software completed its life and gave its place to (b) the CAD designed metal V2 sensor that is suitable for industrial deployment. The (c) prototype electronics where similarly updated to a (d) CAD designed PCB that is significantly more robust. 22
Figure 3: Left: V1 Camera mounted on the MAGICIAN robot in IIT premises. Right: V2 Camera mounted on the same robot during the December integration meeting in IIT. The new sensor weights 4.9Kg instead of 1.9Kg for the old plastic ones (The robot can handle 25Kg). 23
Figure 4: Left: Annotated defect examples from the shipment of TOFAS to FORTH being scanned to provide a training dataset. Right: the perception system deployed in ALT and tested against never seen defects. 25
Figure 5: The MAGICIAN robot scanning surfaces and performing real-time defect classification during the 6th integration meeting in ALT premises. The image features an overlay of the various measurements of interest. Assuming a hovering distance of 60mm over the car surface from the bottom of the sensor each frame offers a 153 mm x 129 mm effective scanning area. The Gigabit ethernet standard allows reception of images at 2464x2056x4 resolution with a 23 Hz rate. The F5.6 setting of the 16mm zoom lens (Figure 9:) and expected defect range allows us to move even closer (e.g. 40mm) while staying in focus. 26
Figure 6: Degree of Linear Polarization visualizations of a Class A Negative Dent comparing the improved V2 vs mock-up V1 lighting system in adverse conditions. Up: The V2 lighting system produces a much cleaner polarization signal that is not affected by ambient light. Down: Extreme ambient sunlight (e.g. during the Trento/Pergine experiments) dominated the recorded signal provided by the V1 lighting system. 28

Figure 7: (a) The initial camera lighting system was based on LED COBs originating from workbench torches with a nominal power of 10W lab measurements however revealed they operated at approximately 3W. (b) After careful review they were superseded by 4x 3.7W LED COBs, driven with 100W of power for ~650 microseconds. (c) Custom cases were developed to house them, (d) polarizers were adjusted to their size, and they were assembled (e) and installed in the camera system (f) to provide better lighting. *29*

Figure 8: (A): All aspects of the hardware design are now CAD defined and made from high quality and durable materials suitable for industrial applications. (B) The physical components we designed and sourced o manufactured. *30*

Figure 9: Top left: the MAGICIAN Main Board PCB. Bottom left: the camera lens configuration that uses an F stop of 5.6. Right: The V2 camera assembled mounted and working on the robot. *31*

Figure 10: To provide enough boards for all partners (including backups) 8x PCB main-boards where sourced, hand-soldered, stress tested and distributed to the partners of the project. *32*

Figure 11: (a) The V2 camera assembled and ready to be mounted on the robot. (b) The camera electronics compartment open, and the wire harnesses removed exposing the internal wiring. (c) The SONY Polarsense camera lens sitting in the middle of the sensor. (d) The 6x LED COBs with polarizers that provide illumination. (e) Close up on one of the 3x ToF distance sensors that passively provide a distance to the scanned surface. 32

Figure 12: A newly assembled vision sensor needs a calibration session to focus the optics and correct/account for possible lens aberrations. This typically needs to be performed once. However, to automate the intrinsic calibration of the vision sensor and make optics focusing easier for adopters of the plat a calibration pattern was included in the base of the robot. We also recorded a set of robot joint configurations that provides an easy calibration procedure. *33*

Figure 13: The sensing robot active on the ALT premises working cell with a laptop connected and viewing the last recorded dataset. *33*

Figure 14: ArUco marker detection was included in the camera system stack to more easily provide passive 6D mesh registration. *34*

Figure 15: The MAGICIAN 1.0 grabber utility listing all available commands when started using –help. *35*

Figure 16: The partners of the project as well as third parties (OC partners etc.) that sign the data sharing agreement receive a username and password with which they can access recorded defect detection datasets through the annotation GUI. *38*

Figure 17: Even areas that look "relatively" dark (Left image) encode information as seen when using the Degree of Linear Polarization (DoLP) visualization. *39*

Figure 18: The left image visualization shows 0,45 and 90 degrees polarizations mapped to R G and B channels as well as the output of the vision classifier. The right image visualization shows the Degree of Linear Polarization view of the data with the red circle marking ground truth. Since there is ground truth and the neural network has correctly responded there is a green tick on the dent on the left image. *40*

Figure 19: Degree of Linear Polarization (DoLP) visualization from the Magician Camera Sensor. Red arrows point to the various defects. a) Welding Spatters, b) Class A positive dent, c) Class C positive dent, d) Sealing residual, e) Class A negative dent, f) Class C negative dent. *41*

Figure 20: The Dataset Creator dialog (`Tools → Create Dataset`) allowing the operator to configure tile size, step size, class inclusion, and focus threshold before converting locally stored annotations into an archive ready

for classifier training. *42*

Figure 21: The tile viewer allows Principal Component Analysis (PCA) allowing easy visualization of defect types and identifying variances and sample outliers. *42*

Figure 22: KNN search results in the Tile Viewer: a query tile (top left) with the K most visually similar tiles retrieved from all loaded datasets, supporting exploratory identification of underrepresented defect sub-types *43*

Figure 23: On-image distance measurement in the annotator: the operator clicks on two points on the image and the system draws a line between two points on the polarisation image, and reports the physical distance in millimetres using the stored camera calibration and ToF stand-off distance *44*

Figure 24: The "Stream To Shared Memory" dialog in the annotator, configured to replay a stored dataset into the shared-memory ring buffer. Downstream ROS nodes receive the replayed frames indistinguishably from a live sensor. *45*

Figure 25: Performance graphs of the various classifiers measured while running on the Magician Pilot PC in ALT premises during the 6th integration meeting. All classifiers in this experiment handle all defect types/severities (one-pass configuration) for two different step sizes (classification densities) applied flatly over the whole image. Framerates above 23 Hz can process the incoming input feed without dropping any frames. For networks beneath the 23 Hz threshold, a larger step size can bring them to an acceptable framerate. The maximum theoretical step size at which all pixels in the image are accounted for exactly once is 48 (non-overlapping). Step sizes 14 and 18 oversample the image by approximately 11× and 7× respectively relative to that baseline, providing multiple independent classifier votes per surface location at the cost of proportionally higher compute load. *46*

Figure 26: When training on Camera V1 and Evaluating on Camera V2, Resnet 18 and ResNeXT 50 are the top performers. It is worth noting that the vehicle present in the ALT presences features only a subset of defects, for example no Positive Dents of Class C. High visibility defects such as Class A defects and welding spatters seem to be reliably detected, however smaller defects seem to not perform well on this cross-camera experiments. *47*

Figure 27: We perform a 5% train/test validation split on V2 Camera data from ALT integration meetings since at the time of writing this deliverable the recently recorded V2 datasets with the newly assembled cameras from FORTH, UNITN and IIT have not yet been annotated. We observe different classifiers taking the lead in accuracy, with Densenet 121 and Regnet 800 mf taking the lead. We also observe much more robust results when comparing to V1 -> V2 cross validation experiments (Figure 26). *47*

Figure 28: The vision classification neural networks employed and tested for the MAGICIAN pilot use-case. Each of them is currently trained to handle 48 × 48 × 4 inputs in several thousands of training samples and output soft-max classification scores identifying the various defects. *48*

Figure 29: All available data from the annotator tool gets infused into the training .PNG tiles and allows fine-grained analysis of results based on criteria such as the active light, distance from defects etc. *50*

Figure 30: Defect detection running in real-time on the MAGICIAN platform tested on ALT premises during the 6th integration meeting. *51*

Figure 31: Detection of a novel defect (created using a screwdriver and a hammer by an expert technician from TOFAS) during ALT trials during the 6th integration meeting. The defect is not part of the training corpus, and

we can see the network classifiers agreeing on the Defect Type (Negative Dent) but slight disagreement between different tiles, with some detecting it as Class B and some as Class C. 52

Figure 32: Detection of novel unannotated defects from metal samples shipped to IIT from TOFAS. Using a step size of 16, typical detections output for "large" Class A, defects look like this example with multiple detections finely covering the area of the defect offering a dense localized classification that can be used to perform the repair. 52

Figure 33: The plot shows the precision of each model in the validation set when doing a 0.05 train/validation split on available Camera V2 data. 53

Figure 34: The plot shows the recall of each model in the validation set when doing a 0.05 train/validation split on available Camera V2 data. 54

Figure 35: The plot shows the accuracy of each model in the validation set when doing a 0.05 train/validation split on available Camera V2 data. 54

Figure 36: Efficiency graphs showing balanced accuracy when also considering the model size (Millions of parameters). We observe different efficiencies (ideally, we would want a network that offers very good accuracy without needing a lot of computation) that hint at the architectures that can competently handle our classification tasks without wasted computations. 55

Figure 37: 4D throughput characterisation: batch size (x-axis), step size (z-axis), effective FPS (y-axis), and achieved accuracy (colour channel) for all evaluated architectures. The plot captures how well each model parallelises on the available GPU and quantifies the accuracy contribution of step size 56

Figure 38: Confusion matrices for the best configurations experimentally derived from 18 candidate models, evaluated on 787,238 tile samples from the ALT vehicle door on the side of the robot that features a Negative Dent (Class B) and Welding Spatters (Class A). 57

Figure 39: Overview of the Tactile Perception Module. (a) Exploded view and (b) assembled prototype of the tactile module. The system integrates a six-axis force/torque sensor (A), a tri-axial accelerometer (B), and a compliant TPU-based comb-like contact interface (C) designed to transmit surface-induced vibrations during sliding exploration. 62

Figure 40: Human-Guided Trajectories Acquisition framework for manual exploration. (a-b) The operator scans the metallic surface using a handheld probe integrating the Tactile Perception Module (highlighted in blue). (b) The spatial kinematics are recorded using a Vicon tracking system and passive retroreflective markers (highlighted in green). (c) The acquired data includes the real-time position and orientation of the device as it traverses the panel, providing the kinematic templates for subsequent robotic replication. 63

Figure 41: Robotic data acquisition setup for the MAGICIAN tactile perception module. (a) Tactile Perception Module integrated into a custom-designed flange for robotic mounting. (b) TPM mounted as the end-effector of a Franka Research 3 robot during controlled surface exploration. (c) Example of the standardized linear scanning trajectory executed around the nominal defect location, used as a repeatable benchmark for robotic tactile inspection. 64

Figure 42: Robotic data acquisition at 6th Integration Week in ALT 64

Figure 43: Setup for tactile data acquisition in ALT 65

Figure 44: Performance evaluation and generalization analysis of the tactile classification models. (a) Comparison between ensemble classifiers and baseline models under the Stratified Group K-Fold protocol,

showing that Bagging (BC), Gradient Boosting (GBC), and Random Forest (RFC) clearly outperform the Majority Classifier (MC) and Random Classifier (RC). (b) Comparison between ensemble classifiers under the LOGO-ScanStyle validation regime, assessing generalization to previously unseen demonstrated scanning styles. (c) Comparison between ensemble classifiers under the LOGO-Panel validation regime, assessing generalization to previously unseen physical panels. *69*

Figure 45: Best-performing confusion matrices for each robotic exploration modality. The matrices are obtained under the Stratified Group K-Fold protocol and normalized by row to show class-wise recall for (a) User Trajectories with Orientation (UOT), (b) User Trajectories with fixed orientation (UT), and (c) Linear Trajectories (LT). All modalities show high performance in separating defect and no-defect samples, while the standardized Linear Trajectory provides the strongest discrimination between positive dents (PD) and negative dents (ND). *70*

Figure 46: Statistical analysis of the factors affecting classification performance. (a) Effect sizes under LOGO-ScanStyle validation, where exploration modality and scanning speed dominate performance. (b) Effect sizes under LOGO-Panel validation, where scanning speed becomes the main factor and the interaction between exploration modality and contact force remains relevant. *71*

Figure 47: Identification of the most robust operating condition across the two validation regimes. (a) Ranking of the operating configurations under LOGO-ScanStyle validation, using F1-score as performance metric. (b) Ranking of the operating configurations under LOGO-Panel validation, using accuracy as performance metric. The green markers indicate configurations that are robust in both validation regimes, while red markers indicate configurations that are statistically non-inferior in both regimes. *72*

Figure 48: The framework of Human-Aware Motion Planning. *75*

Figure 49: The forces recorded using the ATI Force/Torque sensor during the fourth demonstration of the first expert operator at TOFAS. The resting phase (0–15 s, near-zero forces) transitions sharply to the active rework phase as the technician engages the panel. *78*

Figure 50: Deployed Intel Real-sense camera on the worker cell in ALT premises. It is worth noting that we use only the RGB stream of the camera, however the additional depth output that it features can be contrasted to the 3D skeleton derived and used as an additional safety check and guarantee for the robustness of the 3D pose estimation fidelity. *79*

Figure 51: The 3D Pose estimation stack running on an RTX 1000 Ada Laptop GPU and tracking multiple persons while also performing ArUco marker localization and only utilizing a fraction of the available GPU resources *79*

Figure 52: Stress testing D-PoSE in ALT premises when tackling the whole factory floor (left) and tracking a person interacting with the robot arm highlighting the accuracy of the method. *80*

Figure 53: Comparison of HPS errors on the EMDB and 3DPW datasets. SD denotes standard realistic datasets, while BL denotes training exclusively with synthetic datasets (BEDLAM and AGORA). *81*

Figure 54: Comparing complexity of TokenHMR and D-PoSE with respect to GFLOPs, number of parameters and inference time (CPU). *81*

Figure 55: Interactions close to the robot tracked by the human pose estimation module of Magician during the 6th Integration Week in ALT. *82*

Figure 56: To record the car rework, we used four different cameras with different specifications and lens

configurations. Two of them had zoom lens (NIKON based) to help focus on details and two wide-angle lenses (GoPro) to simultaneously capture a wider area. To facilitate accurately tracking the various ArUco markers we created a camera position calculator tool to make an informed decision about camera placement..... 83

Figure 57: Successfully tracking workers performing defect repairs on TOFAS premises. 3D pose estimation is conducted using D-Pose, ArUco marker 3D transformations are overlayed on the visualization and the QR time-sync codes emitted by the laptop screen synchronize all streams with each other and also with respect to the ATI F/T sensor whose data is recorded using the same laptop. 85

Figure 58: Successfully tracking people performing defect sensing on TOFAS premises..... 86

Figure 59: Successfully running D-Pose while also performing defect detection from the same computer system host during the 1st integration meeting in IIT. 86

Figure 60: Left: The Y-MAP-Net (https://rasevents.org/presentation?id=177998) network architecture. Right: The network extracts 2D pose, normal vectors and depth maps for the whole scene from monocular RGB something useful in the context of industrial applications..... 87

Figure 61: Successfully using Y-MAP-Net [Qammaz2026] to extract 2D pose, depth maps, normal vectors and segmentation maps in the observed scene. 88

Figure 62: Employing Y MAP Net on monocular videos from a security camera located in the TOFAS factory and extracting 2D joints, person and vehicle segmentation masks, Depth map, Normal maps and Part Affinity Fields using one-pass evaluation in real-time (19Hz @ RTX 4070 GPU)..... 89

Figure 63: Y-MAP-Net oral presentation during the IEEE International Conference on Robotics and Automation (ICRA) 2026 90

LIST OF ABBREVATIONS

ACRONYM	DESCRIPTION
D	Deliverable
EC	European Commission
WP	Work package
WT	Work task
CR	Cleaning robot
SR	Sensing robot
2D	Two dimensional
3D	Three dimensional
BMI	Body Mass Index
CAD	Computer Aided Design
CMOS	Complementary metal oxide semiconductor
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
DMP	Dynamic Motion Primitives
GDPR	General Data Protection Regulation
GigE	Gigabit Ethernet
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HMR	Human Mesh Recovery
KPI	Key Performance Index
LED	Light Emitting Diode
MOCAP	Motion Capture
MP	Megapixel
NN	Neural Network
PIR	Passive infrared
PSD	Power Spectral Density

RAM	Random Access Memory
RGB	Red, Green, Blue
RGBD	Red, Green, Blue + Depth
ROS	Robot Operating System
SDK	Software Development Kit
ViTs	Vision Transformers
WP	Work package
WT	Work Task
YOLO	You Only Look Once

1 INTRODUCTION

This deliverable constitutes the third periodic report on Work Package 3 (WP3) of the MAGICIAN project (Horizon Europe grant agreement No. 101120731), covering the development of the sensing modules for data acquisition and skills learning. It reports on activities completed from the start of the second reporting period (M19, April 2025) through to the preparation of this document (May 2026, M32), providing a comprehensive update on the state of all WP3 tasks relative to the baseline established in D3.2 (submitted July 2025). The second reporting period has been characterised by three major strategic milestones. First, all MAGICIAN perception software has been consolidated into a set of accessible, thoroughly documented repositories under the `magician-project` GitHub organisation (https://github.com/magician-project), establishing a stable, versioned, and reproducible software baseline for the project.

Second, the complete sensing stack was deployed at an industrial site for the first time during the series of integration events hosted by ALT (Figure 1). All major sensing and processing nodes the sensing camera, the custom MAGICIAN PCBs, the multi-modal data acquisition engine (*magician_grabber*), tile-based vision classifier (*magician_vision_classifier*), the body-pose estimation node (*magician_body_pose_estimation*), general perception node (*magician_general_visual_perception*) were co-located, integrated, and exercised on real automotive bodywork provided by ALT and TOFAS. Similarly, the learning by demonstration stack (*magician_learning_by_demonstration_multicamera_object_tracking*) as well as the PCB design and firmware for the V2 Camera (ma*gician_main_board* repository). This marks the transition of the MAGICIAN WP3 effort from laboratory prototyping to industrial-site validation and constitutes the first demonstration of the full perception pipeline under conditions representative of the target operating environment.

Third, each individual software component of the perception module has matured substantially in terms of both software quality and measured performance. The two-stage ensemble defect classifier exceeds 90% precision, recall, and accuracy on the currently available Camera V2 train & validation datasets and runs above the 23 Hz target allowed by the GigE available bandwidth on the pilot hardware even when operating concurrently with the body-pose estimation and general-visual-perception nodes. The MAGICIAN Main Board PCB has been manufactured in a batch of eight and has proven reliable throughout the ALT integration weeks.

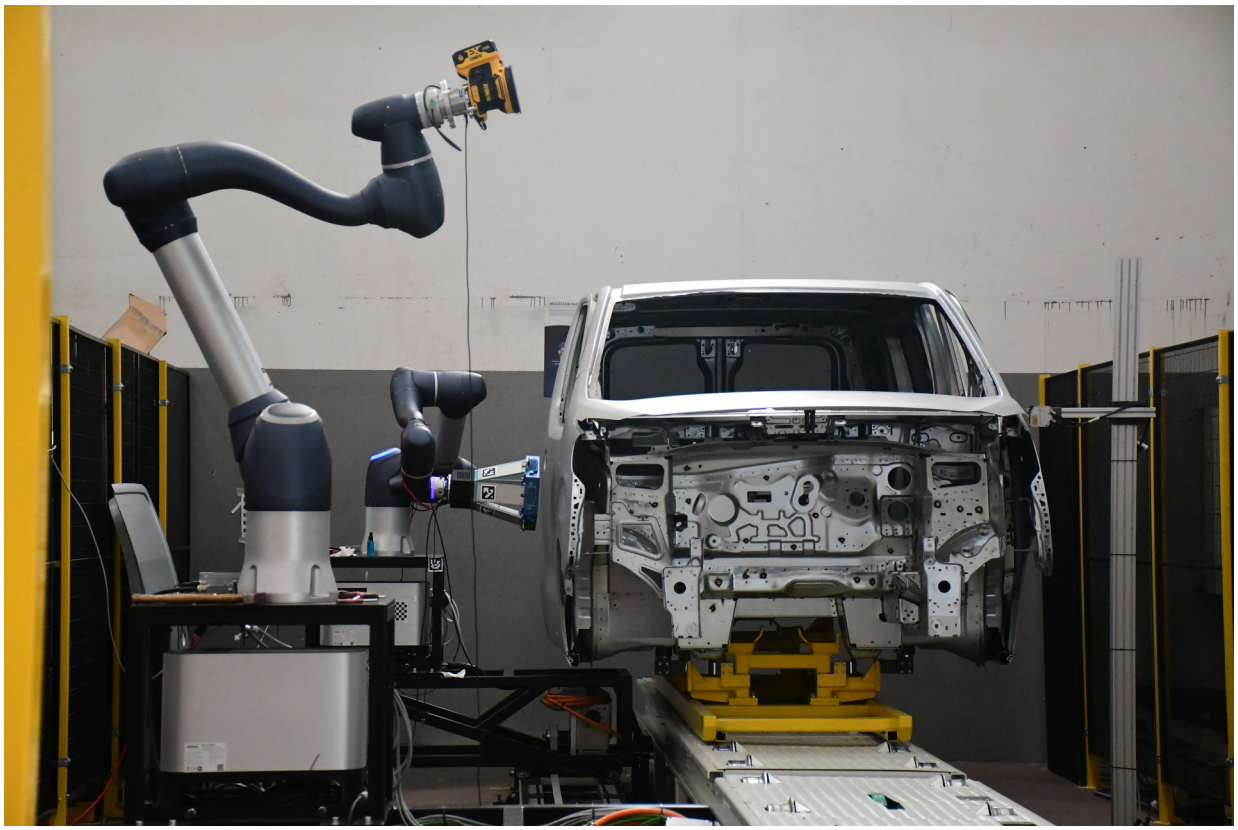


Figure 1: The Hardware & Software prototypes of the perception system integrated with the Doosan robot during the MAGICIAN 6th integration meeting in ALT premises.

1.1 PURPOSE AND SCOPE

The primary objective of this deliverable is to report the progress achieved in the development, integration, and validation of the perception systems within the MAGICIAN project during the third reporting period. Building upon the foundations established in D3.1 and D3.2, this document describes the evolution of the visual and tactile perception modules for imperfection detection, the human motion perception framework, and the methodologies for learning defect inspection skills from human operators.

Emphasis is placed on the transition from laboratory-level prototypes to integrated system deployments under conditions representative of the target industrial environment. The deliverable documents the consolidation of hardware and software components, improvements in sensing performance, the expansion of the acquired datasets, and the validation activities conducted during project integration events. In addition, it reports on the progress made toward the deployment of a complete perception pipeline capable of supporting autonomous defect detection, human-aware operation, and skill transfer within the MAGICIAN robotic platform.

The scope of this deliverable includes the description of the sensing hardware, data acquisition and annotation procedures, machine learning and perception

methodologies, experimental results, achieved performance with respect to the project KPIs, and the remaining challenges to be addressed in the final phase of the project. The presented results provide the basis for the continued integration of the perception modules with the robotic platform and for the final validation activities planned in the subsequent work packages.

1.2 CONTRIBUTION TO PROJECT OBJECTIVES

Similarly to D3.1 and D3.2, the developments described in this deliverable contribute to multiple scientific, technological, societal, and demonstration objectives of the MAGICIAN project. In particular, D3.3 supports:

Scientific and Technological Objectives:

- **O1:** A robotic perception module integrating visual and tactile sensors for defects analysis and classification. D3.3 reports the further maturation of the MAGICIAN perception module, including the integration of visual and tactile sensing, the consolidation of data acquisition and annotation procedures, and the improvement of defect classification methods. The work contributes directly to the development of the sensing robot, enabling more robust detection, classification, and interpretation of surface imperfections.
- **O2:** A robotic cleaning/reworking module attached to a robotic platform and equipped with specialised end-effectors to rework defects. Although WP3 focuses mainly on perception, the outputs described in this deliverable provide essential information for the subsequent reworking actions performed by the robotic platform. The defect localisation, classification, and multi-modal sensing results generated by the perception pipeline are necessary inputs for planning and executing repair strategies.
- **O3:** A human-aware robotic system able to work in collaboration with human operators. The human motion perception and prediction components contribute to the development of a robotic system that can perceive operators, understand their movements, and operate safely in shared industrial environments. These capabilities are essential for enabling collaborative inspection and repair scenarios.

Social Sciences and Humanities (SSH) Objectives:

- **O4:** Understanding, modelling, and transferring human expertise in defect inspection and repair. D3.3 contributes to the SSH dimension of MAGICIAN by addressing the observation and modelling of human inspection behaviours. The learning-from-demonstration activities support the capture of expert knowledge and its transfer to robotic systems, preserving human skills while enabling new forms of collaboration between workers and autonomous technologies.

Demonstration Objectives:

- **O5:** Demonstration of the MAGICIAN technologies in representative industrial scenarios. The work reported in this deliverable supports the project demonstration objectives by providing perception modules that have progressed from laboratory prototypes toward integrated deployment in realistic industrial conditions. The integration activities and validation tests described in D3.3 demonstrate the readiness of the perception framework to support the final MAGICIAN platform demonstrations and the assessment of project KPIs.

Overall, this deliverable demonstrates that the MAGICIAN perception technologies have reached a level of maturity that enables their operation as part of an integrated industrial robotic system. Through the combination of visual, tactile, and human-centred perception capabilities, the project has progressed beyond isolated sensing solutions toward a comprehensive perception framework capable of supporting autonomous defect detection, human-aware collaboration, and knowledge transfer from expert operators. The achievements reported in this deliverable represent a key milestone toward the final validation of the MAGICIAN platform and its vision of intelligent, trustworthy, and human-centred automation in industrial defect inspection and repair.

1.3 RELATION TO OTHER WORK PACKAGES

The perception components described in the present deliverable are at the heart of most of the project's activities, and they have an understandably strong relation with many of its work packages. The synthetic list of the most important relations remains identical to the one of D3.1 and D3.2 and is offered next for the purpose of document self-containment:

- **WP2 – Use case definition and platform design**: even if the perception solutions developed in the WP claim for a certain level of generality, the initial idea and the main design choices are connected to the specific requirements of the automotive use case identified as the main driver of the project's research activities. In particular, the unique combination requirements on the perception system's accuracy, on its integration within a robust robotic platform, and on the final cost of the solution pose several formidable challenges that we are facing in the activities in WP3 and that are succinctly reported in this report.
- **WP4 – Robotic platform and interfaces:** WP3 and WP4 are the two main pillars producing the technological assets at the heart of the system components. The activities of the two WP are deeply intertwined. Specifically, the planning and scheduling components (T4.3) take their decisions based on the results of the defect analysis and on the prediction of the possible motion of human operators. The motion control and active sensing component make a direct use of the Information processed through the perception pipeline. This information is also used for T4.5 (closed-loop defect analysis). On the other hand, also the Inverse Information flow (from WP4 to WP3) is extremely important. Knowing the

possible motion performance and strategies of the robot arm where the perception system is mounted sets the background for the development of perception strategies (e.g., the velocity of the motion and the accuracy of the distance between the perception system and the car plays an important role the decision of the optical system and of the visual processing pipeline).

- **WP5 - Integration and performance analysis**: The components developed in WP3 and described in this document will be integrated in the final platform (T5.1). Most of them will be part of the demonstrator (T5.2) and their performance will contribute substantially to the project's KPIs.
- **WP6 – Cascade funding management**: since the perception component will be used in the subprojects stemming from the cascade funding scheme, the project's findings will be crucial to offer support and technical assistance (T6.4).

1.4 STRUCTURE OF THE DOCUMENT

The document maintains the structure of D3.1 and D3.2 which comprises six chapters, addressing key components of the perception systems developed for the MAGICIAN project, with the last section devoted to the outlook and planning for the future integration of the modules and their fusion in a multi-modal system.

After this introductory chapter, Chapters 2 and 2 focus on the visual and tactile perception systems, respectively, for imperfections detection. Chapter 3 addresses human motion detection, emphasizing its importance for human-robot collaboration.

The document concludes in Chapter 5 summarizing the key outcomes and setting the stage for the next steps in the MAGICIAN project.

1.5 PERCEPTION SYSTEM OVERVIEW

This section outlines the key requirements and system architecture of the MAGICIAN perception module, which supports synchronized, real-time data acquisition across multiple sensing modalities.

1.5.1 REQUIREMENTS AND SPECIFICATIONS

The requirements for the perception sensor remain the same and are closely tied to the key performance indicators (KPIs) set by the project. The initial prototype version of the perception system proved a good initial approximation of the perception solution needed and provided the consortium with a very early platform to work, experiment and validate the project's required performance and accuracy analysis goals defined in the D2.1 "Use Case Definition" document.

1.5.2 SYSTEM ARCHITECTURE

The MAGICIAN perception system uses ROS (Robot Operating System) as its core development platform. This was a decision taken at a very early stage of the project that has in a large way shaped the software architecture of MAGICIAN and already proven very useful, especially with regards to the fast development cycle of the OC1 integration. The overall system architecture has had no drastic changes and is a mature and optimized aspect of the project.

Additions from the previous deliverable have to do either with minor modifications, for example extra fields in ROS messages or new ROS services or topics that allow faster debugging of the various MAGICIAN nodes. The most notable change to the system architecture is that the tactile data can now be also streamed using the zero-copy shared memory mechanism initially developed for the high-bandwidth camera data. Tactile data can now either be transmitted over ROS or through shared memory as a means to further offload ROS bandwidth so that it does not overhead seamless robot operation.

The dual-path design with ROS topics for control and low-rate telemetry and shared memory for high-bandwidth latency-sensitive data is now applied uniformly to both the visual and tactile modalities. This establishes a stable integration pattern for any high-rate sensor that may be added in subsequent OC1 or OC2, without requiring changes to the core ROS communication infrastructure.

2 VISUAL PERCEPTION FOR IMPERFECTIONS DETECTION

Visual perception is the primary sensing modality that factory workers rely on for imperfection detection. As seen during our TOFAS observations (Figure 58), due to the very large surface area of the car to complete the scan in the available time, workers need to quickly visually inspect the car and then use their touch to confirm their visual observations. The MAGICIAN system mirrors this by making vision the central defect detection path that quickly deals with the largest part of the car. The scanning speed and accuracy of the visual perception module therefore directly determine whether the project's Key Performance Indicators (KPIs) are met, making it one of the most consequential components for overall project success.

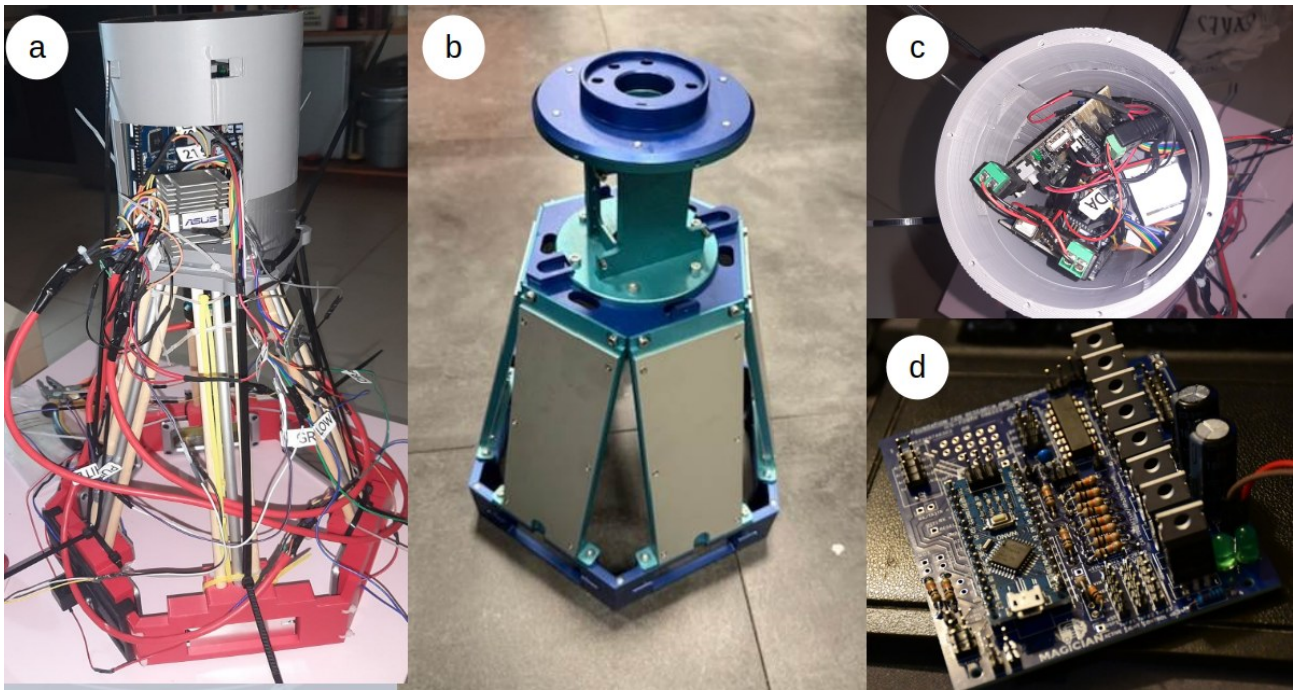


Figure 2: The mock-up prototype V1 (a) of the video sensor created to test various optics configurations and bootstrap the development of perception software completed its life and gave its place to (b) the CAD designed metal V2 sensor that is suitable for industrial deployment. The (c) prototype electronics where similarly updated to a (d) CAD designed PCB that is significantly more robust.

Immediately following the visit to the industrial plant of TOFAS in January 2024, during April 2024, FORTH received a large shipment of annotated defective material that was instrumental in forming a very early proof of concept prototype (Figure 2.a). After this initial prototype was demoed to the project partners during October of 2024, through much coordinated effort this first version of the visual perception system was first tested in the premises of Istituto Italiano di Tecnologia (IIT) during March of 2025 performing

detections in defective metal samples from the shipment sent to IIT that it had never seen before (since these samples were not part of the training dataset sent to FORTH). This initial version of the sensor was made of cheap and easily replaceable 3D printed parts, exposed all wiring to allow immediate experimentation with different light sources and electronic components allowing rapid development and changes to the prototype. Its electronics were based on a custom breadboard (Figure 2.c). The proof-of-concept sensor allowed us to gather data and develop the software stack of the perception system while simultaneously iterating on a robust physical design for the chassis and electronics of the sensor.

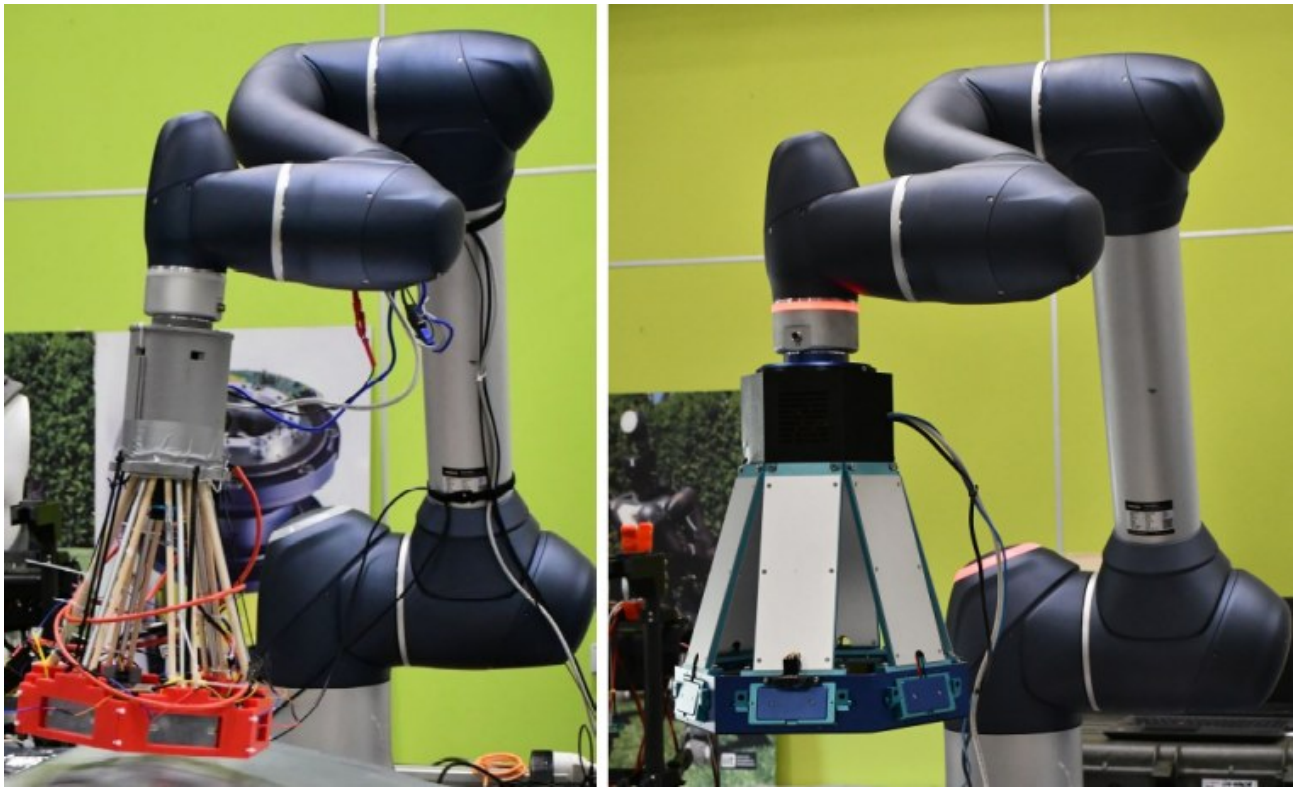


Figure 3: Left: V1 Camera mounted on the MAGICIAN robot in IIT premises. Right: V2 Camera mounted on the same robot during the December integration meeting in IIT. The new sensor weights 4.9Kg instead of 1.9Kg for the old plastic ones (The robot can handle 25Kg).

Overall, the visual perception system has converged and is now in a very mature state with high-quality and industrial grade materials and CAD design that allows easy replication of the sensor for the various partners without the need for custom modifications and device variability. The flimsy plastic shell of the initial mock-up sensor (Figure 2.a) was replaced with a sturdy robust metal shell (Figure 2.b). The various external electronics assembled on a hand-crafted breadboard (Figure 2.c) where unified and routed in a professional PCB designed using KiCAD (Figure 2.d). The number of external cables was minimized and cables where properly routed through dedicated channels of the CAD design protecting them from the environment (Figure 3). The fixed

plastic light orientation was superseded with tiltable lights that can be configured for other lighting scenarios. The initial software defined light/camera synchronization gave its place to hardware exposure synced lighting. The duty cycle of lights was thus modified from 4500 μsec of exposure per light to 650 μsec. This in turned allowed us to overvolt the LED COBs from 3 W continuous lighting to 100 W strobe flashes synced to the camera exposure. The incredibly low exposure both enabled incredibly crisp images (equivalent what would be expected from an off-the-shelf camera with 2222 Hz recording rate) and made ambient light much less prevalent compared to the emitted polarized light, boosting the quality of the obtained images. The software accompanying the sensor has similarly matured and received a lot of attention and polish to seamlessly support the integration meetings and continuous testing. All configuration of the sensor can be done using command-line parameters, there has been a lot of attention to computation optimizations, and all tasks that need to be done with the sensor can be performed using the same executables and the provided Github documentation.

2.1 INTRODUCTION

The project use cases described in D2.1 for automotive manufacturing present significant challenges. The MAGICIAN system will need to deal with various car chassis, each comprising hundreds of uniquely processed metal sheets with diverse shapes, edges and contours. Sensing is required to detect a broad variety of imperfections, namely:

- Positive/Negative dents
- Weld spatters
- Sealing material residuals,
- Deformations
- Material defects

each of which uniquely affects the involved surfaces and produces a unique identifying visual signature, as illustrated in Figure 4.



Figure 4: Left: Annotated defect examples from the shipment of TOFAS to FORTH being scanned to provide a training dataset. Right: the perception system deployed in ALT and tested against never seen defects.

OC2 applications are expected to extend the variety of data recorded and classified by the system, yielding an even bigger range of imperfections handled by the system. Even regardless of the novel defects targeted by the MAGICIAN sensing robot, the KPIs set for the Stellantis pilot (O1-KPI-SR1, O1-KPI-SR2, O1-KIP-SR3, Table 1) require exceptional optical clarity to be able to target the low-end 300 micron threshold set by O1-KPI-SR1. At the same time the very large surface area of the car does not allow us to use to use lens that overly magnify it to produce a sufficiently large and detectable signature on the sensor, since the framerate at which images are captured and processed is similarly constrained by the available bandwidth of the industrial camera GigE standard. The whole sensor design (see Figure 5) is thus a custom design working around these constraints and gracefully balancing, image resolution, optical clarity, surface covered by each frame and available image transport bandwidth.

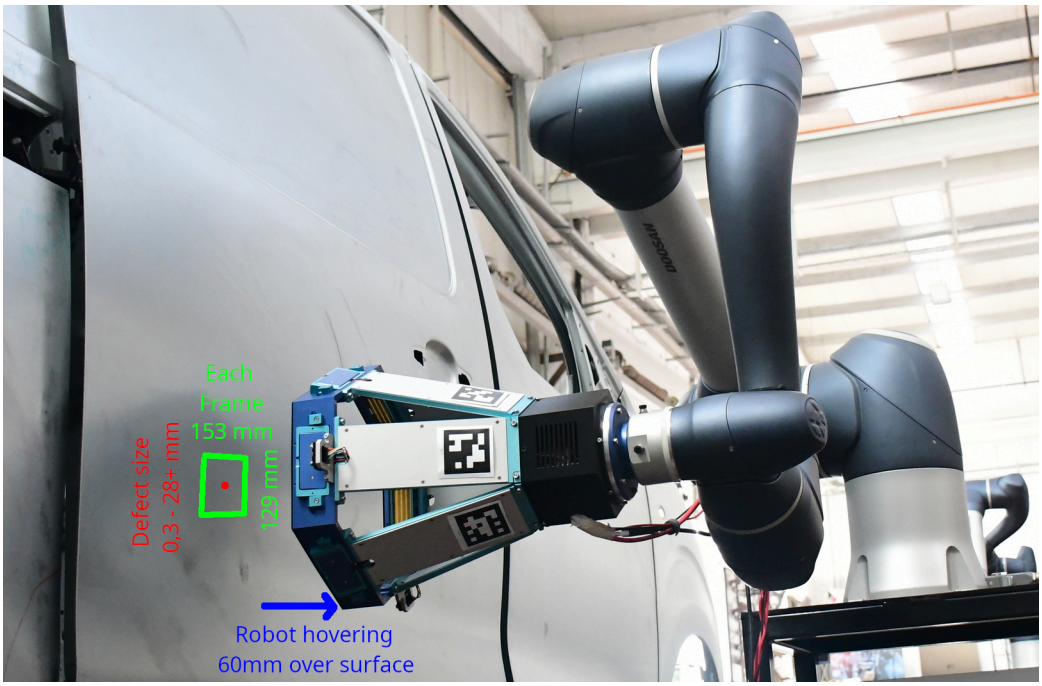


Figure 5: The MAGICIAN robot scanning surfaces and performing real-time defect classification during the 6th integration meeting in ALT premises. The image features an overlay of the various measurements of interest. Assuming a hovering distance of 60mm over the car surface from the bottom of the sensor each frame offers a 153 mm x 129 mm effective scanning area. The Gigabit ethernet standard allows reception of images at 2464x2056x4 resolution with a 23 Hz rate. The F5.6 setting of the 16mm zoom lens (Figure 9:) and expected defect range allows us to move even closer (e.g. 40mm) while staying in focus.

2.2 OBJECTIVES AND REQUIREMENTS

Hereafter we report the KPIs, as described in the Deliverable D2.1 - "Use Case Definition", related to the Visual Perception for Imperfections Detection. These KPIs (Table 1) define the constraints within which the developed visual perception solution must perform.

Scientific and technological objective	KPI ID	KPI definition	After MAGICIAN
(O1) A robotic perception module integrating visual and tactile sensors. The module will be embedded in a robotic sensor module (the SR, hereafter) and will be used for defects analysis and classification. The SR will replicate the skills of human workers through a learning scheme.	O1-KPI-SR1	Smallest size of defect that can be sensed/detected by the perception module.	≤0.3mm
	O1-KPI-SR2	Detection success rate vs humans.	False positives: ≤120% Skipped defects: ≤110%
	O1-KPI-SR3	Car-body scan time compared vs humans on a benchmark set.	≤110%

Table 1: KPIs related to the Visual Perception for Imperfections Detection.

2.3 CAMERA SYSTEM

The MAGICIAN camera system for imperfections detection is based on a SONY XCG-CP510 sensor equipped with a global shutter polarization 5.1 Megapixel CMOS sensor and GigE interface. The sensor captures a polarized image where each individual pixel of each captured frame features one of four different linear polarization filters. This enables simultaneous capture of four polarization images at 0°, 45°, 90° and 135° with perfect synchronization. The camera is based on the IMX250 sensor and offers excellent characteristics with MTBF of 62042 hours (7.1 years) robust vibration, operating temperature, humidity and shock resistance and a C-Mount for lens making it plug & play compatible with a wide variety of optics. Of course, correct lighting matters for the signal captured by even a great camera sensor and is also very important for the camera system to provide crisp, low noise high-fidelity images. The 6 LED illuminators featured on the MAGICIAN sensor emit light with a specific polarization. Defective areas cause light to be reflected in a non-uniform manner. The neural network thus can work with 4x different readings for each pixel (instead of just one luminosity value compared to a regular monochrome camera) for more robust defect detection. After multiple experiments the LED illuminators initially selected with a nominal wattage of 10W for the V1 Camera (Figure 3) were found to under-perform. Testing the initially chosen lights of the first iteration of the camera system in different settings such as in IIT premises, FORTH premises and the Trento Pergine facility revealed captured images with different illumination profiles. While images collected at a lab setting using only artificial lights of low intensity featured much stronger polarization cues compared to images from the Pergine facility (Figure 6) that was exposed to very strong sunlight during recording experiments. Further lab experiments revealed the real wattage of the "advertised nominally as 10W" LED COBs to be around 3W explaining the discrepancy.

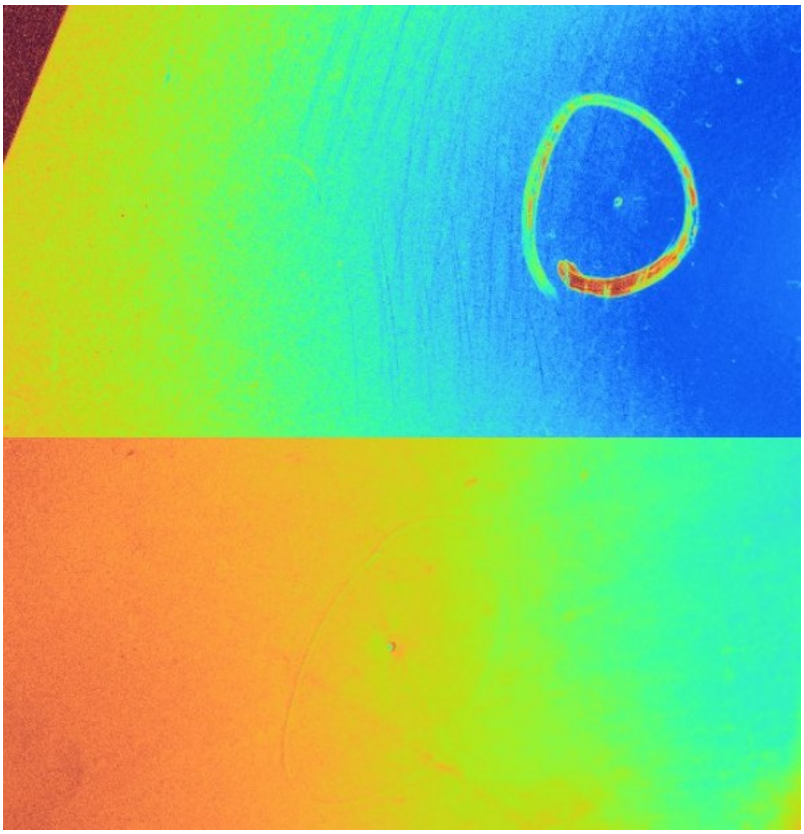


Figure 6: Degree of Linear Polarization visualizations of a Class A Negative Dent comparing the improved V2 vs mock-up V1 lighting system in adverse conditions. Up: The V2 lighting system produces a much cleaner polarization signal that is not affected by ambient light. Down: Extreme ambient sunlight (e.g. during the Trento/Pergine experiments) dominated the recorded signal provided by the V1 lighting system.

Another architectural flaw of the initial V1 lighting system was the complete reliance on software defined light control that let each of the 6 lights on continuously until an image was captured using the camera. Given that an exposure took approximately 4500 microseconds and the relative imbalance of the very strong sunlight compared to the under-powered LED COBs this also prompted us for a move towards hardware light synchronization via strobing. To make the recorded images consistent across any deployment and reduce the generalization gap of training data we thus opted to radically improve the lighting system. Instead of relying again on a LED torch with potentially dubious illumination specifications, we instead directly sourced LED COBs from a LED COB manufacturer (Figure 7). We also selected COBs with smaller dimensions and where able to incorporate 4x of them in the same space previously occupied by one. Furthermore, by utilizing the Camera GPIO headers and incorporating them in the LED control circuit we were able to drastically reduce the duty cycle of each COB allowing us to precisely align the light pulses with the camera exposure. This allowed us to use 100W light pulses without damaging the LED COBs. Figure 7 shows the upgraded lighting system which successfully minimizes ambient light and provides

much stronger and cleaner polarization signal which is a big and fundamental upgrade of the current V2 version of the MAGICIAN Camera System.

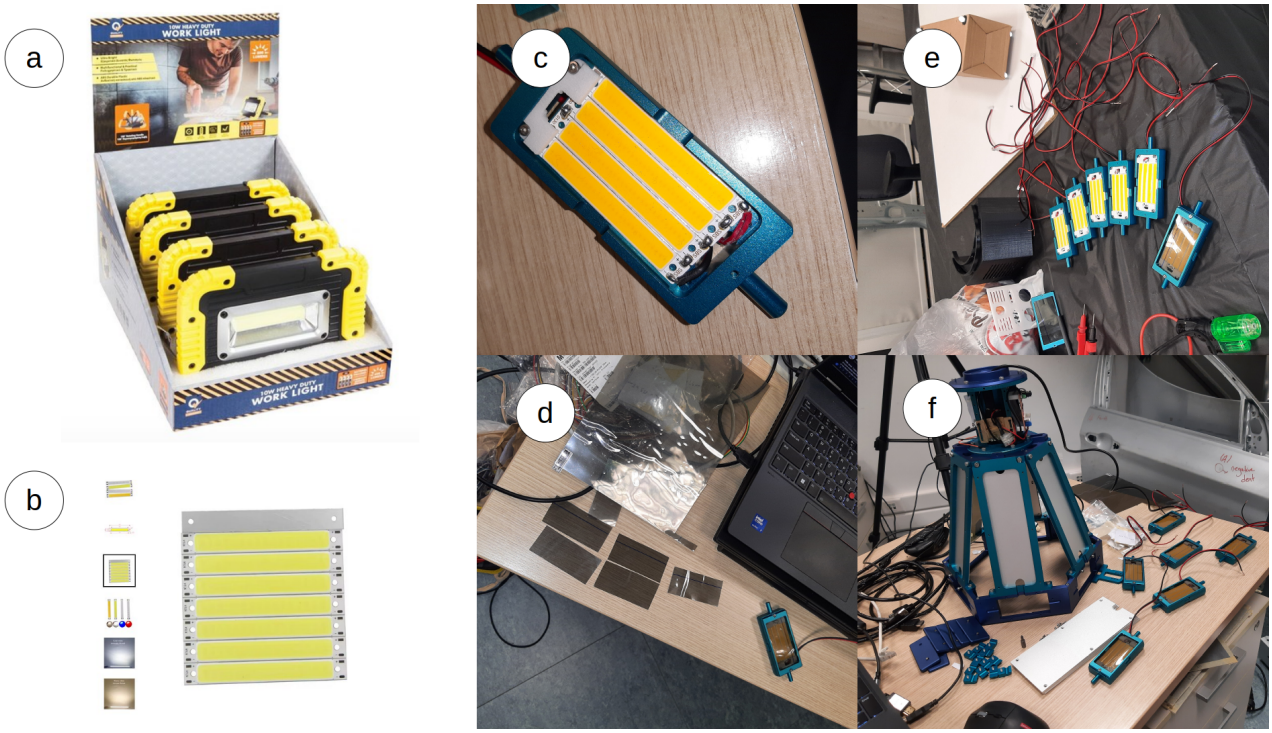


Figure 7: (a) The initial camera lighting system was based on LED COBs originating from workbench torches with a nominal power of 10W lab measurements however revealed they operated at approximately 3W. (b) After careful review they were superseded by 4x 3.7W LED COBs, driven with 100W of power for ~650 microseconds. (c) Custom cases were developed to house them, (d) polarizers were adjusted to their size, and they were assembled (e) and installed in the camera system (f) to provide better lighting.

The initial camera optics were chosen using a camera configuration simulator developed to calculate an effective selection of lens to accommodate the requirements of our project. The initial 12mm lens was selected to provide a balance between accuracy and scanning speed since a less zoomed-in view requires less time to complete the scanning operation helping us to comply with O1-KPI-SR3. After deploying this configuration however and despite the polarization camera that effectively provides 4 measurements for each pixel it became apparent that in order to accurately classify detects in near the minimum size limit (O1-KPI-SR1) better optics would provide a clearer view of the defects and thus allow the deploy neural networks to learn cleaner features and perform classification with higher accuracy. The lens of the camera system was thus updated to a 16mm lens that reduced the effective view area to 14 cm x 10 cm when hovering 4.5cm over a surface. For the Class C 300x300 micron defects we get 3.2p x 3.2p ~= 9 pixels coinciding with the 3x3 Convolution size minimum feature of the neural network that performed classification.

However, the miniaturization of the onboard electronics that initially were based on an Arduino Mega, to the Arduino Nano as well as the design of a new PCB drastically

reduced the size needed for electronics. Similarly, the miniaturization of the LED COBs allowed reconfiguring the cone shape of the sensor. Thus, even when hovering 60mm over the car (Figure 5) with the new CAD designed camera sensor we now get 4.3 x 4.3 ~= 16 pixels on the minimum size defects. The initial microcontroller selected for controlling the low-level peripherals such as Buttons, Accelerometers, ToF sensors and LED lights was Arduino MEGA 2560 (Figure 8).

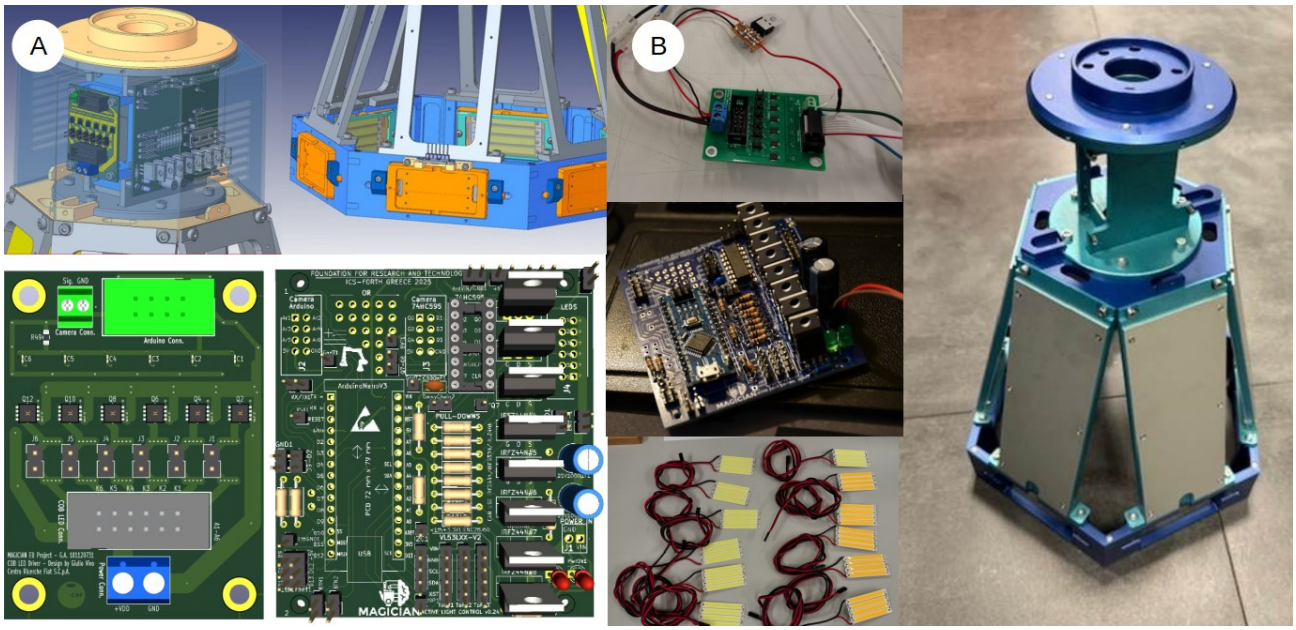


Figure 8: (A): All aspects of the hardware design are now CAD defined and made from high quality and durable materials suitable for industrial applications. (B) The physical components we designed and sourced o manufactured.

The main reason for its choice was its 4096-byte EEPROM (program memory) which was needed to populate data structures to interface with 3x instances of the vl53lxx-v2 range sensor code using the official API for the sensor provided by STMicroelectronics. Trying to use an Arduino NANO (1024-byte EEPROM) initially seemed infeasible. After optimization of the STMicro code and omission of API features that where not used compilation became possible on an Arduino NANO, however available SRAM (2KB in Arduino Nano vs 8KB in Arduino Mega) became the bottleneck resulting in situations where during execution and depending on the load of the microcontroller the stack pointer would surpass the heap pointer leading to memory corruption and a restart of the microcontroller. A written from scratch implementation for VL53LXX-V2 (https://github.com/pololu/vl53l0x-arduino) coupled with significant effort made Arduino NANO a viable choice for the camera system freeing up valuable physical space (Figure 9) allowing further miniaturization of the whole sensor. Moving from the Arduino Mega (101.52 mm length, 53.3 mm width) to a Nano (45 mm length, 18 mm width) and taking into account the thermal properties of the camera, the 3D structure of the connecting bracket between the camera and the robot arm, we thus went on with a split 3x board design, that consists of the MAGICIAN Main board which is the biggest one at

72x79 mm and featuring two smaller boards one for the dedicated exposure synced LED strobing, and one for ethernet connectivity (instead of USB) for the magician main board. The design is flexible since the presence of the two additional boards is optional and enables H/W syncing to the camera exposure and ethernet operation.

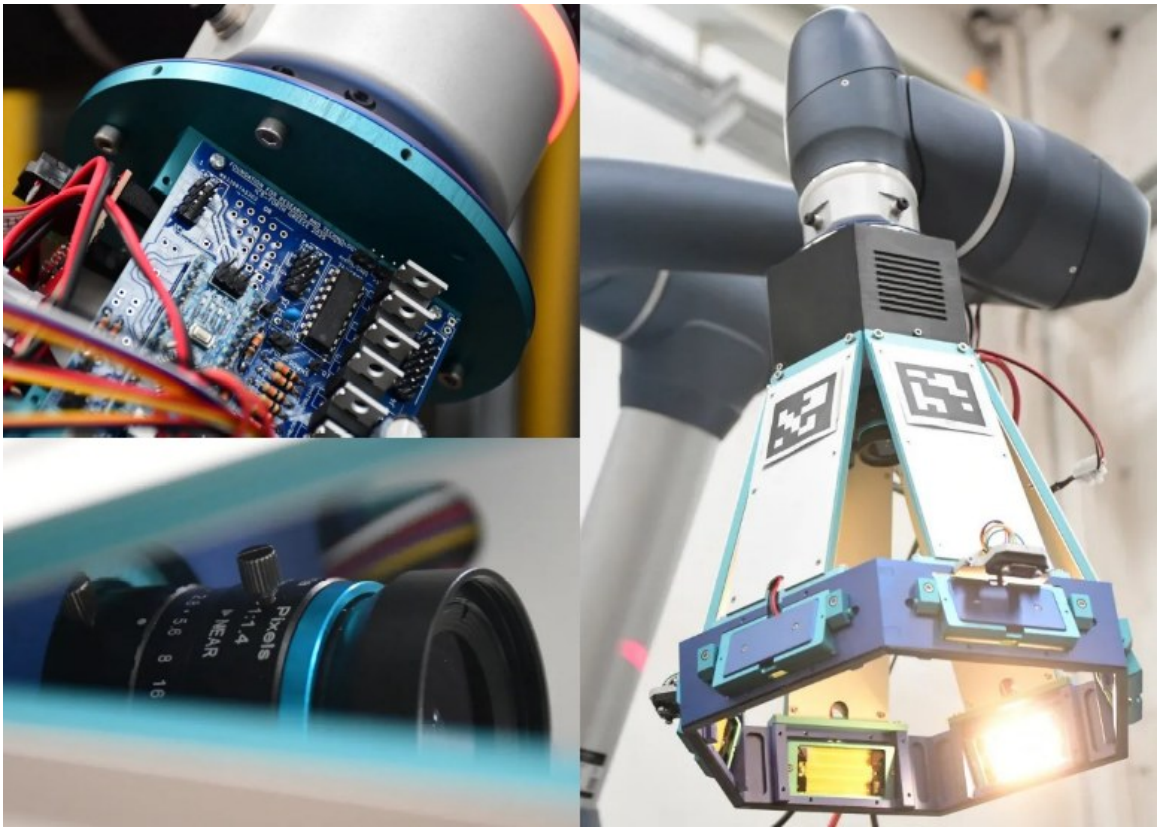


Figure 9: Top left: the MAGICIAN Main Board PCB. Bottom left: the camera lens configuration that uses an F stop of 5.6. Right: The V2 camera assembled mounted and working on the robot.

A total of 8x PCB boards (Figure 10) where manufactured and distributed between the project partners provisioning one board for each partner as well as a backup that could quickly and easily substitute one if needed. After completing the assembly of the V2 Camera Sensor (Figure 11) a one-time calibration step needs to be performed (Figure 12) to focus the optics and confirm that the lights and sensors are working as intended. Moreover, an ArUco-based calibration has been also implemented for the mesh registration (Figure 14). After this procedure concludes the sensor becomes immediately and transparently accessible to the MAGICIAN ROS stack using the *magician_grabber* package and data can be immediately viewed, recorded and modified digitally from the connected computers (Figure 13) without needing physical access to the camera.

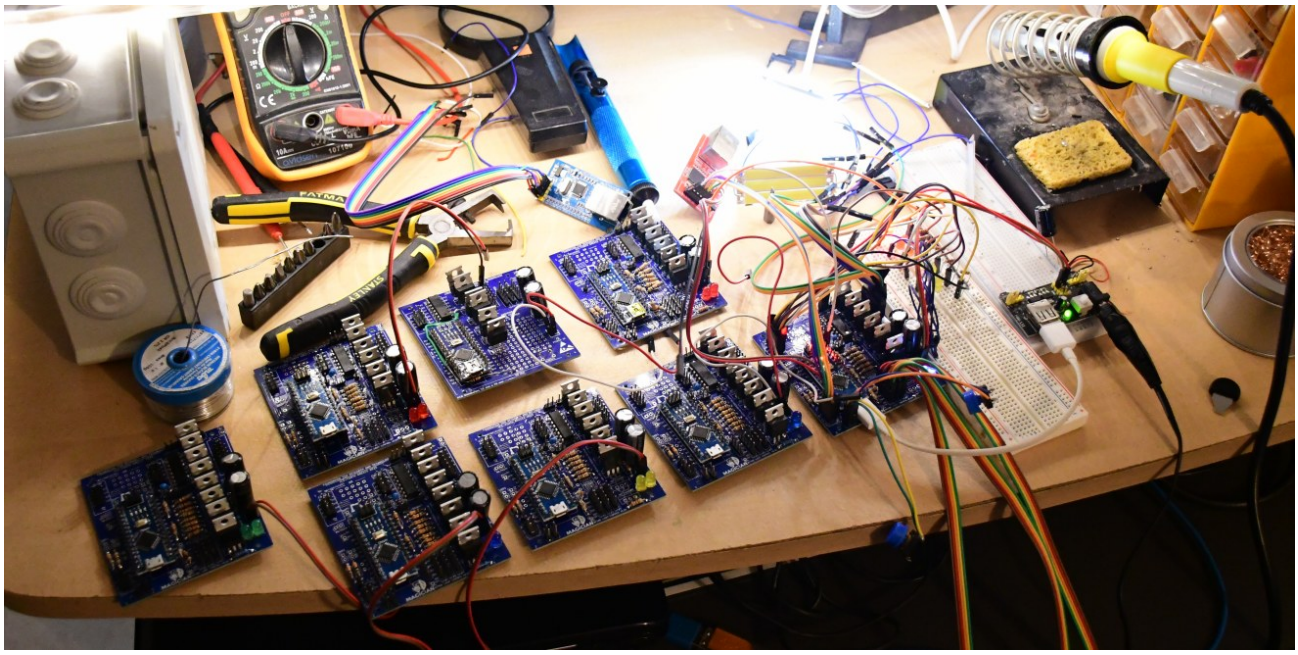


Figure 10: To provide enough boards for all partners (including backups) 8x PCB main-boards where sourced, hand-soldered, stress tested and distributed to the partners of the project.

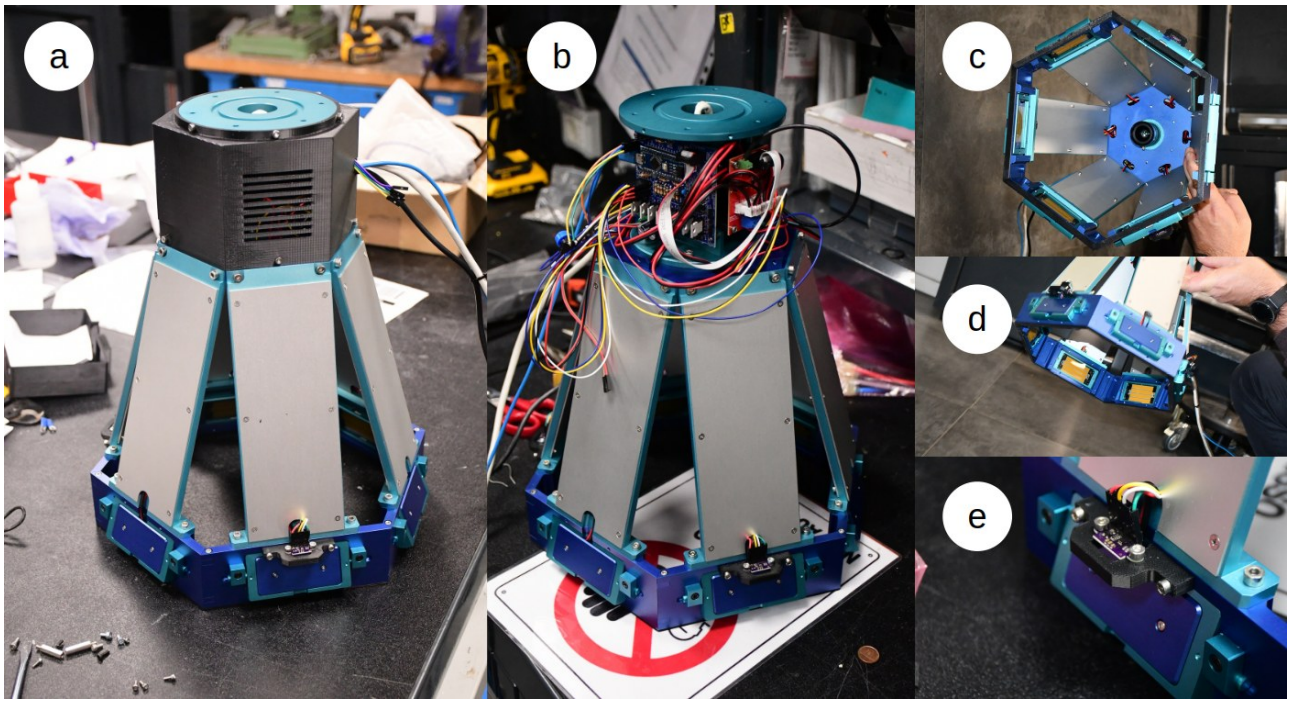


Figure 11: (a) The V2 camera assembled and ready to be mounted on the robot. (b) The camera electronics compartment open, and the wire harnesses removed exposing the internal wiring. (c) The SONY Polarsense camera lens sitting in the middle of the sensor. (d) The 6x LED COBs with polarizers that provide illumination. (e) Close up on one of the 3x ToF distance sensors that passively provide a distance to the scanned surface.

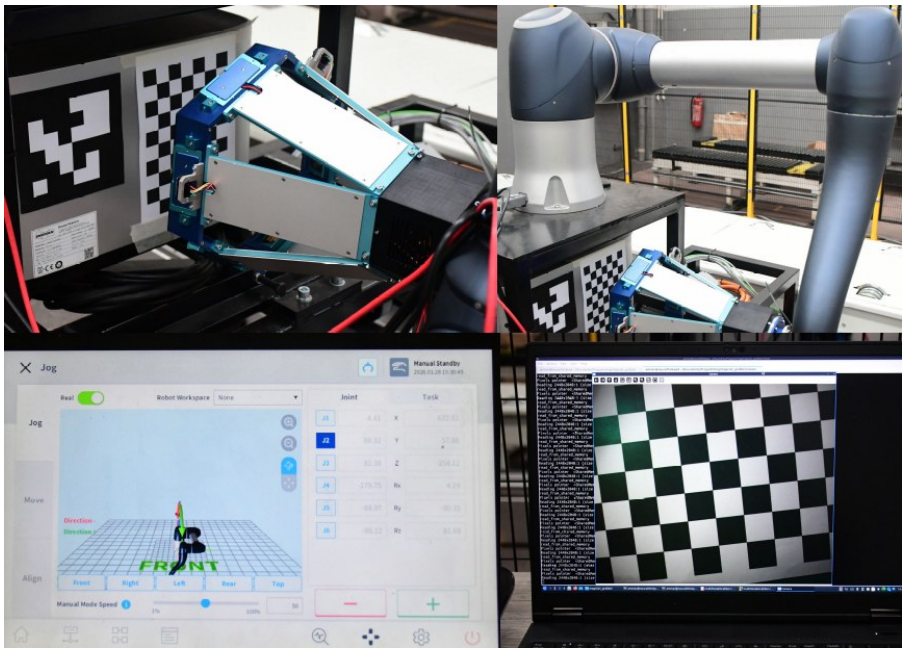


Figure 12: A newly assembled vision sensor needs a calibration session to focus the optics and correct/account for possible lens aberrations. This typically needs to be performed once. However, to automate the intrinsic calibration of the vision sensor and make optics focusing easier for adopters of the plat a calibration pattern was included in the base of the robot. We also recorded a set of robot joint configurations that provides an easy calibration procedure.

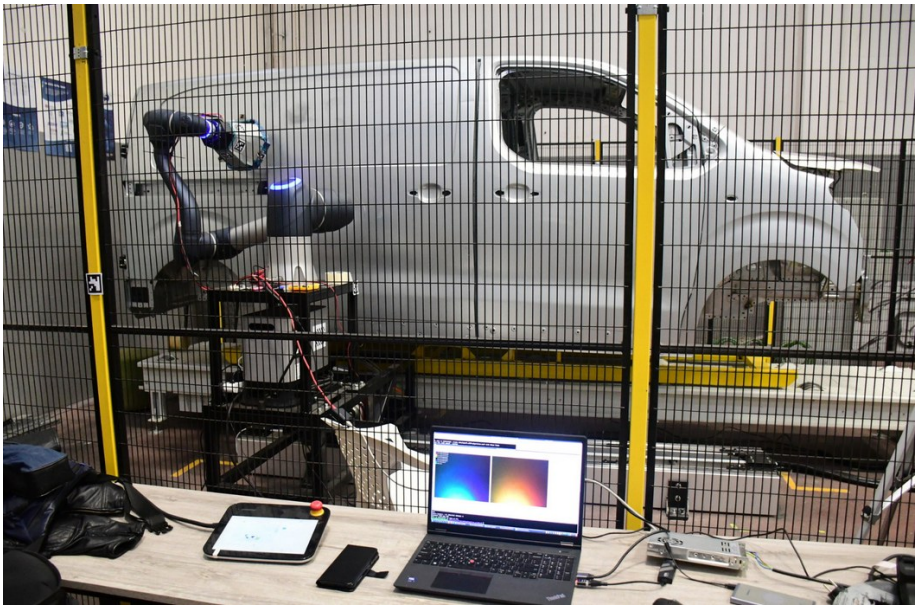


Figure 13: The sensing robot active on the ALT premises working cell with a laptop connected and viewing the last recorded dataset.

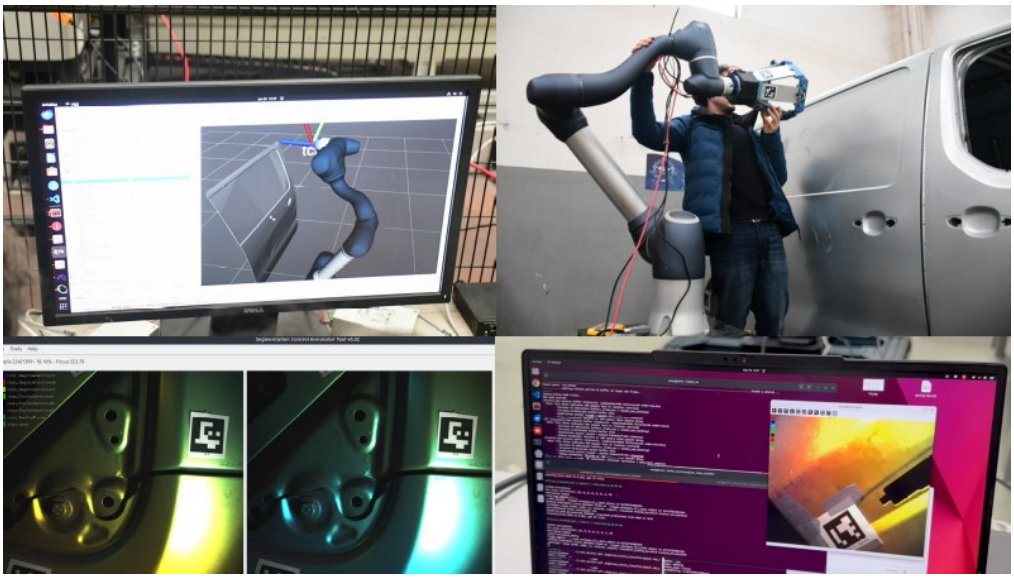


Figure 14: ArUco marker detection was included in the camera system stack to more easily provide passive 6D mesh registration.

2.4 DATA ACQUISITION AND ANNOTATION

2.4.1 DATA ACQUISITION

The MAGICIAN Grabber utility (Figure 15) has gradually evolved as a mature, centralized software interface that can simultaneously handle all the data streams of the MAGICIAN sensor in a synchronised, transparent and easily programmable way. The utility has been used to successfully capture 129199 frames of V1 camera data and at the time of writing 73305 frames of V2 camera data. Since the grabber has been a vital part of the project since its beginning, the interfaces and main design of the data acquisition pipeline have remained largely the same. However, the need of annotating a huge number of captured data in a timely manner motivated many quality-of-life improvements to the annotator tool that has been radically improved since the previous version. In this chapter we mainly focus on the new additions of data acquisition and annotation tools, however also keeping relevant base information about the various utilities to make the deliverable more easily accessible without constant cross references to previous deliverables.

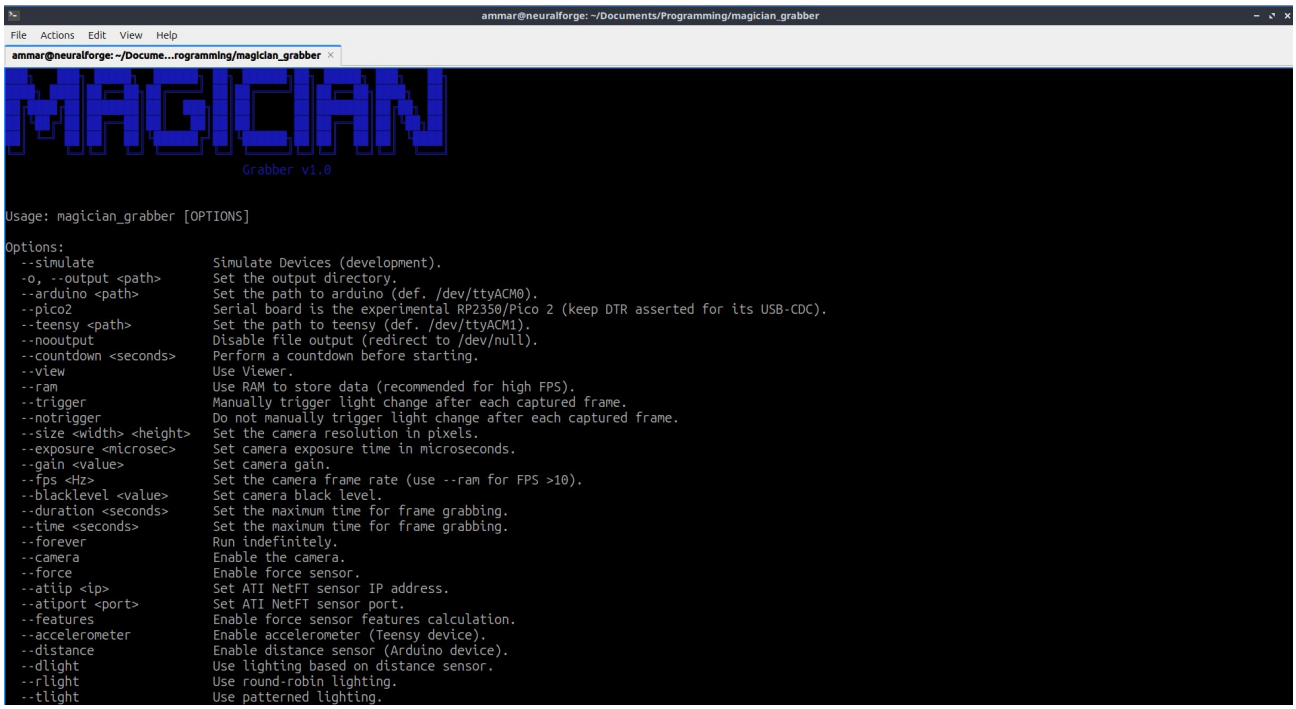


Figure 15: The MAGICIAN 1.0 grabber utility listing all available commands when started using *–help*.

With respect to sensing modalities, the grabber utility can either record data to disk and/or stream data to client applications. It is written from scratch with a multi-threaded approach in mind, dedicating a thread (possibly pinned to a distinct CPU core) for each data source. It can also be compiled in 3 different configurations. Its first and simplest binary form is the *magician_grabber* which is entirely written in C and handles all input streams (including tactile), however without performing any processing of the input streams. Its second and more complicated C++ binary is *magician_grabber_tactile* which also includes bindings to libfftw3 and the IIR library (https://github.com/berndporr/iir1). This version of the library processes tactile force and accelerometer data in real-time calculating friction, acceleration spikiness, and power spectral density (PSD) for the exerted forces and accelerations experienced by the sensor. Calculating these representations of the incoming force data simplifies the requirements of the tactile classifier by locally performing computations in a context where the data is closely synchronized. The final C++ binary generated is the *ros_magician_grabber* which includes the dependencies to the ROS library and can interface the data streams using ROS topics and services. All topics broadcasted by the ROS version of the *magician_grabber* are published under the */magician_grabber/* prefix to make them easily distinguishable from other ROS topics. In more detail: Force data is broadcast under */magician_grabber/wrench_sens*ed using a *WrenchStamp*ed geometry message. This allows force data to be used by the robot control nodes to also estimate contact forces during manipulation. Accelerometer data are broadcast at the *magician_grabber/accel_sensed* topic using an *AccelStamped* message. It is worth noting that *AccelStamped* features both a linear and angular component, however the

data produced by the MAGICIAN sensor does not contain angular information, so this component is set to zero. Distance values from the time-of-flight sensors are emitted as Float32 values in millimetres in the */magician_grabber/distance1-3* topics. Active LED lights of the sensor can be intercepted using the */magician_grabber/light1-6 topics* and */magician_grabber/button* broadcasts the state of the physical button that is used for the annotation of tactile data.

All versions of the grabber accept the same command line parameters and can be configured using the same commands. Issuing the command line parameter *–help* displays a list of available commands and parameters as seen in Figure 15.

A fully manual data capture session for the Camera V2 can be accommodated using only the *magician_grabber* and the following command: *./magician_grabber_tactile --all --size 2448 2048 --exposure 650 --fps 23 --time 45 --countdown 10 --output* ***ExperimentName***

This will provide 10 seconds of lead time for the user to take the sensor in his hands and position it over the area with defective samples, grab using all available streams (camera, force sensor, accelerometer, etc.) with an exposure of 650 microseconds acquiring a video stream of 23 Hz for 45 seconds. This is a typical configuration when grabbing datasets. The main difference of executing the grabber between the old V1 and V2 camera is changing the exposure to 650, since the old 4500 microsecond will produce so bright lights that it will saturate the captured image.

A ROS broadcasting session involves just running *./ros_magician_grabber* that automatically enables the *--all, --nooutput* and *--stream* commands.

A dataset generated on disk features files of three different file formats, JSON, CSV and PNM. **ExperimentName/info.json** is the first file generated that contains the grabber configuration for the specific dataset. **ExperimentName/colorFrame_0_xxxxx.pnm** files contain the recorded images using a lossless representation, with each image occupying 4.8 MB on disk. **ExperimentName/controller.csv** contains information about peripherals connected on the Arduino NANO (Figure 12). The CSV header contains System Timestamp (Unix Time), *dev_timestamp* (Arduino loop time in msec), Button state, Distance1-3 (VL53LXX-V2 ToF laser range sensor reading in millimetres), Light1-6 LED power switch state with values 1 or 0. Values follow the capture framerate of the Polarized Camera to align LEDs and distances to the incoming frames. Sampling rate for messages is typically 10 - 23 Hz depending on how fast frame acquisition is set up (command line argument -- *fps XX*). One of the improvements implemented to improve the storing efficiency of the grabber was using .PNG file compression to reduce disk size and make the transport of data over network faster. The 4.8 MB of an uncompressed image amount to approximately a 2.6 MB PNG on average. The images are also de-bayered and the 4 polarization channels are stored as RGBA channels fitting the PNG specification. Further experiments to reduce size and compression/decompression overheads led to an experimental PZP (Portable Zipped PNM) format compatible with ZSTD or LZ4 compression codecs. This typically is 2.28x faster to load than PNG and may

produce 52% smaller on disk (for example when storing segmentation masks). However, in the interest of making the recorded data available to a wide audience we opted for the universally decodable without extra dependencies PNG format that makes data more accessible.

ExperimentName/tactile/force.csv records the force sensor readings from the ATI Net F/T force sensor recording timestamp (Unix timestamp when the message was received in microseconds), *fX, fY, fZ* force data and *tX, tY, tZ* torque data. Sampling rate is 7 kHz but can be reduced using the ATI Web interface.

ExperimentName/tactile/accelerometer.csv records timestamp (Unix timestamp when the message was received by the host system in microseconds), *dev_timestamp* (Teensy microcontroller loop time in milliseconds) and *accX, accY, accZ* acceleration measurements. Sampling rate is 4 kHz but can be reduced by re-flashing the teensy microcontroller. Finally, if the *magician_grabber_tactile* binary is the one used, it will automatically emit the following files:

ExperimentName/tactile/acceleration_psd.csv

ExperimentName/tactile/acceleration_spikeness.csv

ExperimentName/tactile/force_psd.csv
ExperimentName/tactile/friction.csv

Each of these files contains a pair of values, the first is a timestamp which has been synchronized between the acceleration and force data (even if they operate at different rates), and the second column of data contains the processed value for each specific measurement. These measurements can be directly used by the tactile classification module to simplify calculations. Other notable additions to the grabber are command-line definition of the various peripherals to connect to, allowing to distinguish between multiple devices MAGICIAN Sensor devices simultaneously connected on the same system. The support of a richer serial command interface between the grabber and the main board to help diagnose potential connectivity problems with informative CLI messages. The timestamp synchronization happens on microsecond level basis, and the shared memory has been also improved with provisions that ensure immediate access to the recorded camera framebuffer while also considering the graceful mutex locking.

2.4.2 DATA ANNOTATION

To annotate the data, we developed a tool based on the cross-platform *WxWidgets* library, specifically its *wxPython* wrapper. Upon execution the tool opens a GUI window that accepts a path to a dataset and proceeds to open it and allow visual inspection of recorded samples and given a few clicks highlight the defect type and classification which will in turn be used by the neural network stack to perform learning and successfully tackle the task. The original design of the Annotation tool was intended for use within a single organism (e.g. FORTH, ALT, or TOFAS) with local storage for the various datasets. Having multiple sensors recording data across different project partners and locations and needing to coordinate the dataset archival and annotation

lead to the overhaul of the tool adding full network capabilities for the transport and retrieval of the data. The network functionality of the annotator is inspired by the Git workflow but with a simplified and GUI oriented implementation. Each user of the annotator has a username and password acquired after applying for access through the MAGICIAN Data Access Committee (DAC). Having access, the user can then list the available datasets through the menus of the application and stream or download them locally. Every downloaded frame also automatically carries its latest accepted annotations. The user can then click on the screen to add more annotations, and these are applied to his local copy of the datasets. Finally, the user can select File -> Upload Annotations and "commit" the local annotations to the FORTH server where the administrators can accept them and merge them to the source datasets for all users. At the time of speaking the MAGICIAN consortium users of the application include FORTH, UNITN, IIT and PIPPLE while from the OC1 Cascade funding teams YongaTek, Multiverse and Introsys also requested access and were approved by the DAC (Figure 16).

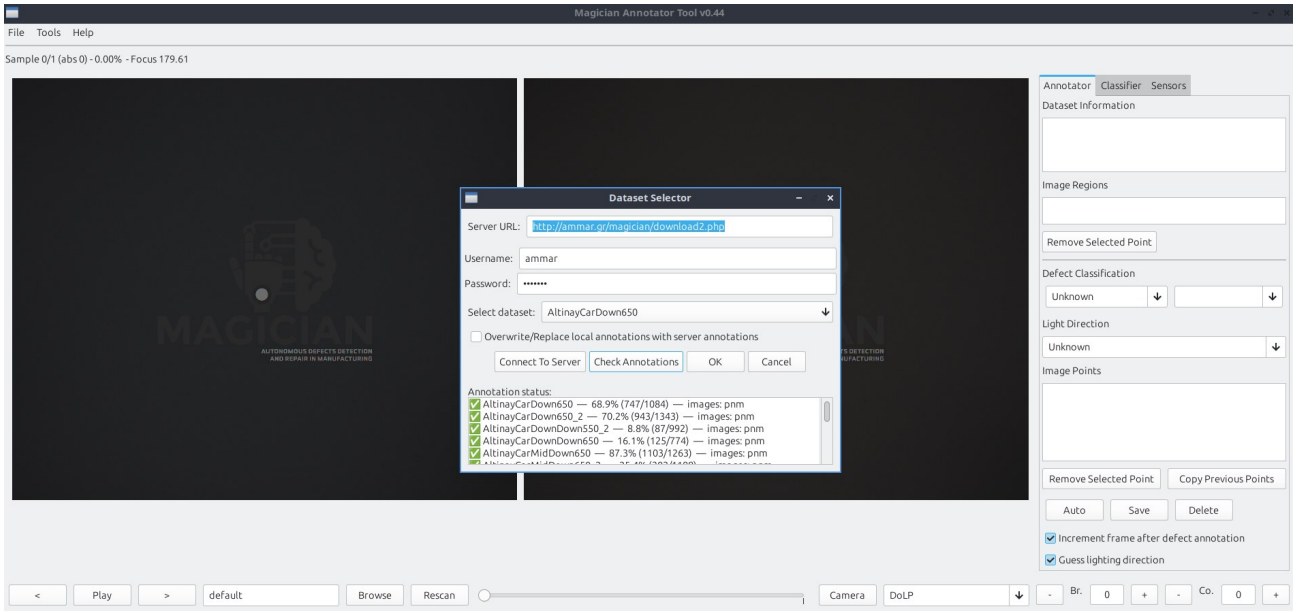


Figure 16: The partners of the project as well as third parties (OC partners etc.) that sign the data sharing agreement receive a username and password with which they can access recorded defect detection datasets through the annotation GUI.

Data annotation is a time consuming and repetitive task that consumes a lot of human resources. Furthermore, physical details in the lighting system, camera lens, focus, and physical geometry of the sensor invalidate previous training efforts giving rise to new patterns and requiring new data acquisition and data annotation to be properly supported. Since the start of the project, we have consistently tried to improve this aspect and facilitate data annotation in a way that yields time savings with a cascading positive effect to the rest of the project. Experiments were also conducted with additional VLMs, namely Gemma 4 and Qwen 3, following the same prompt-and-parse methodology. None of the evaluated models reliably generalised to polarisation-sensor

imagery, which differs strongly from the RGB natural-image distributions these models were trained on. Experiments with Segment Anything 3 (SAM3) for zero-shot segmentation where also conducted. After converting the polarised images to greyscale by averaging the four polarisation channels, SAM3 was applied in both automatic and semi-automatic modes. Despite SAM3 receiving an Honourable Mention for Best Paper at CVPR 2026 and being considered the state of the art for zero-shot segmentation, it was not successful in detecting defects, even Class A instances with a relatively large magnitude. SAM3's segmentation proposals are trained on natural image statistics and seem to be insensitive to the subtle intensity gradients that distinguish defects in our greyscale images. The annotation workflow accordingly relies on point/region annotation with optional VLM assistance rather than automatic segmentation, though this automation capability may prove useful in the OC2 cascade funding round as more use cases are introduced.

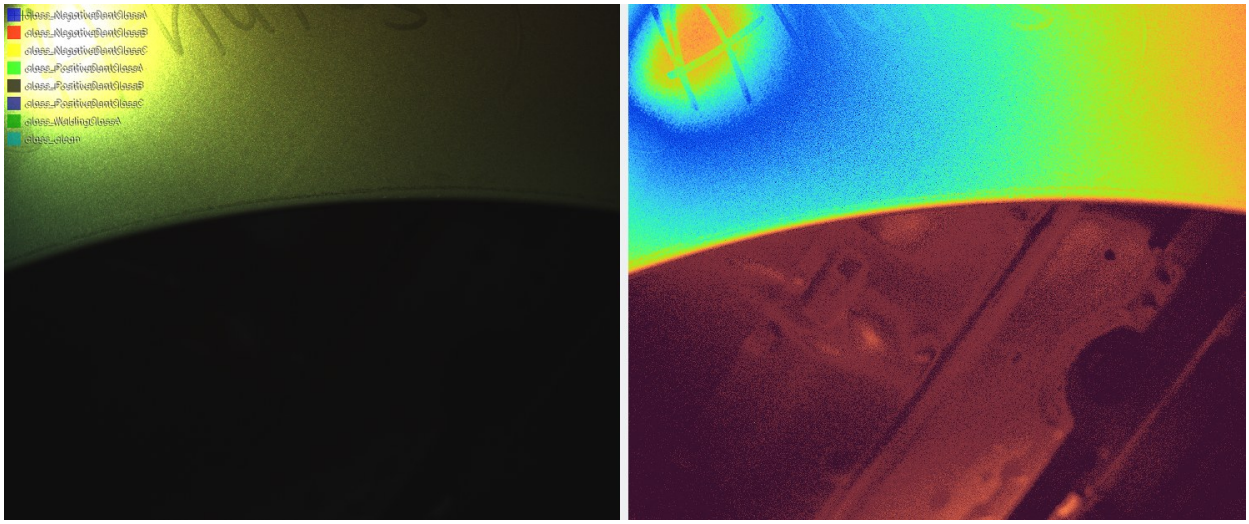


Figure 17: Even areas that look "relatively" dark (Left image) encode information as seen when using the Degree of Linear Polarization (DoLP) visualization.

The list of data visualization methods was extended to include Degree of Linear Polarization (DoLP), Angle of Linear Polarization (AoLP), s0, s1, s2 and s3 Stokes Parameters, Degree of Polarization (DoP), Degree of Circular Polarization (DoCP), Total Polarization (ToP), Coherence of Polarization (CoP), Max/Min/Average Polarizations, Surface Normals and Sobel edge detection. This extended palette of side-by-side visualizations was motivated by the observation (illustrated in Figure 17) that frames which appear uniformly dark in the conventional intensity view can still carry rich structural information when rendered through the DoLP view. The right image pane consists of ground truth, while the left one also incorporates classification results from the various trained classifiers, with small crosses of various colours representing the different defect type and severity combinations detected as seen in Figure 18.

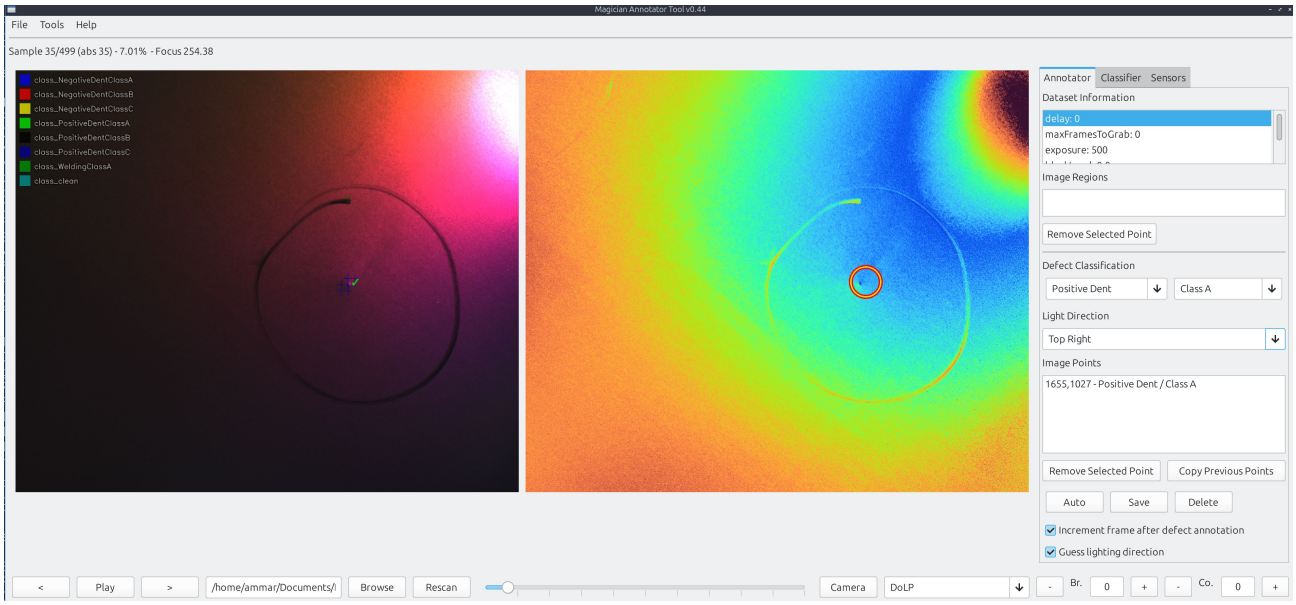


Figure 18: The left image visualization shows 0,45 and 90 degrees polarizations mapped to R G and B channels as well as the output of the vision classifier. The right image visualization shows the Degree of Linear Polarization view of the data with the red circle marking ground truth. Since there is ground truth and the neural network has correctly responded there is a green tick on the dent on the left image.

The dual-panel layout of the annotator exploits this directly. The left panel renders the 0°/45°/90° polarisation channels mapped to R/G/B together with the live inference overlay coloured crosses marking each detected defect class and severity combination. The right panel simultaneously displays the DoLP view alongside a red circle indicating the ground-truth annotation. When the classifier correctly responds, a green tick is superimposed on the left panel at the same location (Figure 18). This side-by-side layout allows the annotator to compare the model's response against the polarimetric signature of the defect in a single glance, substantially accelerating the verification workflow.

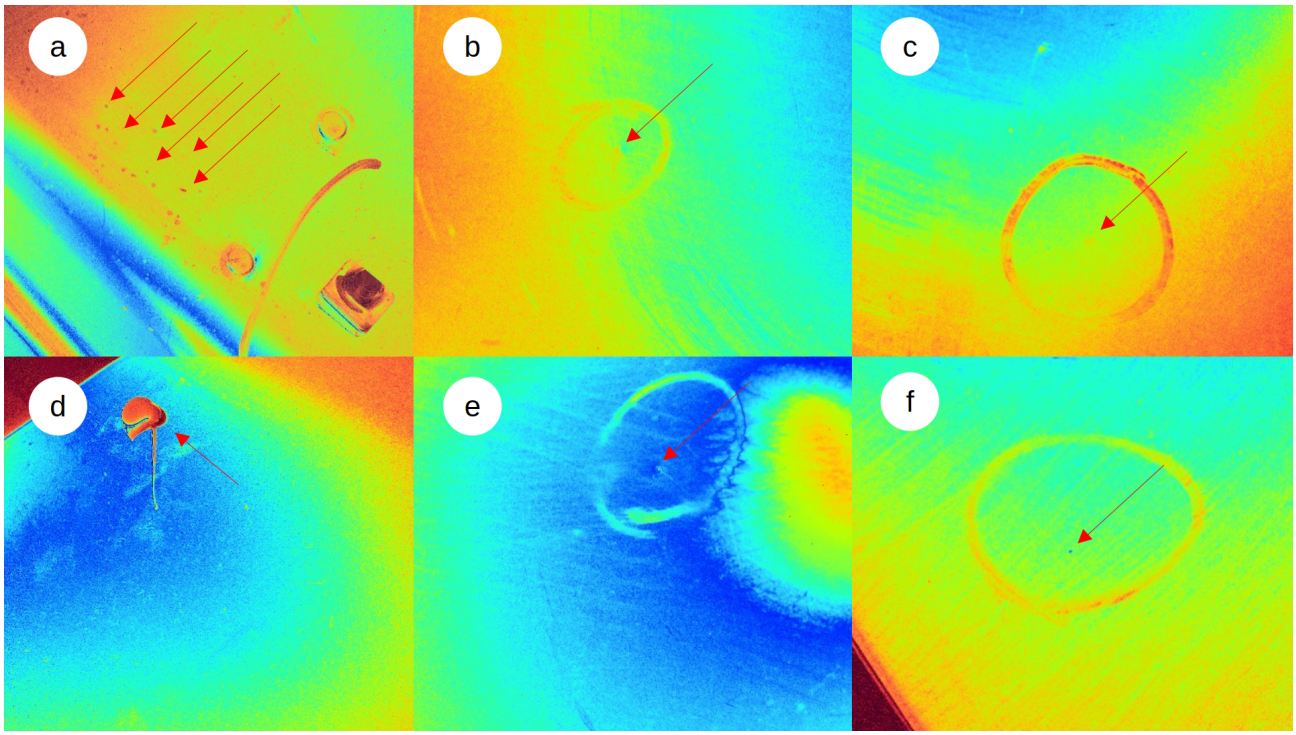


Figure 19: Degree of Linear Polarization (DoLP) visualization from the Magician Camera Sensor. Red arrows point to the various defects. a) Welding Spatters, b) Class A positive dent, c) Class C positive dent, d) Sealing residual, e) Class A negative dent, f) Class C negative dent.

The DoLP visualisation is particularly powerful for survey purposes because it produces a consistent, strong visual signature for every defect category regardless of the illumination direction or surface finish. Figure (Figure 19) shows representative crops across all defect types annotated in the MAGICIAN corpus: (a) Welding spatters, (b) Class A positive dent, (c) Class C positive dent, (d) sealing residual, (e) Class A negative dent, and (f) Class C negative dent. In each case the defect boundary is clearly delineated in the DoLP channel even when it is ambiguous in the intensity view, confirming that the polarimetric modality is the primary discriminating cue for the tile classifier. After downloading (and if needed annotating) all datasets an operator wants to have locally available the functionality of the annotator tool does not end. It also has provisions for Exploratory Data Analysis (EDA) and direct generation of training data. By selecting the Dataset Creator from Tools, the user can select the tile size, step size and various configuration parameters and thresholds (Figure 20) to convert the annotated frames to a dataset ready to be loaded and be used for training or validation.

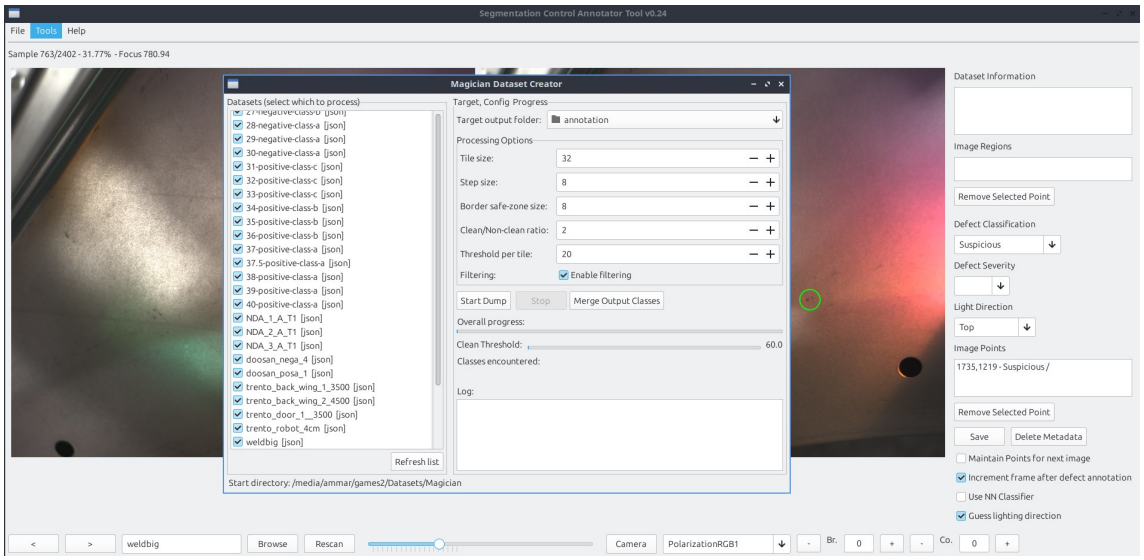


Figure 20: The Dataset Creator dialog (`Tools → Create Dataset`) allowing the operator to configure tile size, step size, class inclusion, and focus threshold before converting locally stored annotations into an archive ready for classifier training.

The annotator includes a Tile Viewer panel that displays all annotated tile patches from a dataset in a two-dimensional embedding computed via Principal Component Analysis (PCA) (Figure 21). PCA is applied to the raw by default 48 × 48 × 4 pixel values of each image tile, reducing the high-dimensional representation to two components that are displayed as a scatter plot. The resulting map reveals the internal structure of the annotated tile corpus: tiles of the same defect class cluster together, while outlier tiles (atypical illumination, unusual defect geometry, misannotated frames) appear as isolated points far from their class cluster. A K-Nearest Neighbours (KNN) search widget (Figure 22) allows the annotator to select any tile and instantly retrieve the K most visually similar tiles across all datasets. This feature was introduced to support exploratory data analysis and to identify underrepresented defect sub-types for example, Class C dents at the boundary between the positive dent and “clean” categories which may require targeted data collection to ensure balanced training.

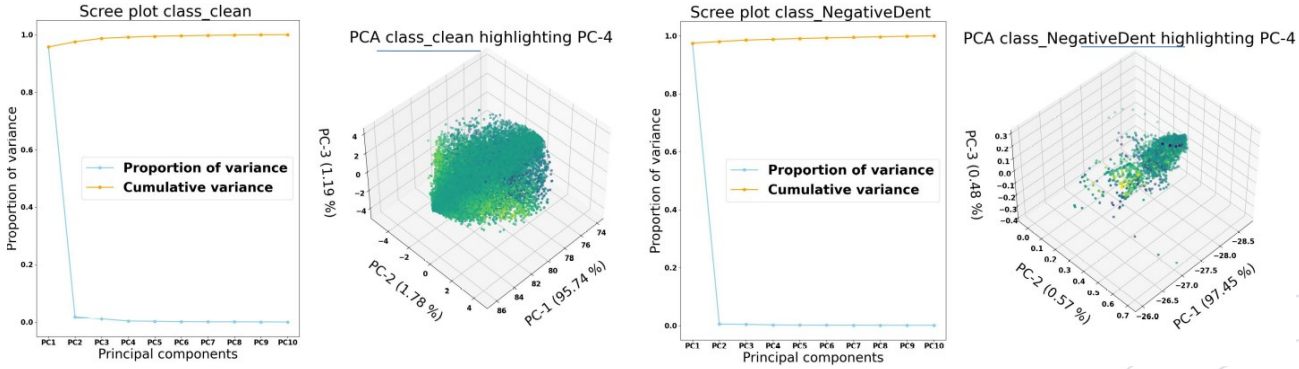


Figure 21: The tile viewer allows Principal Component Analysis (PCA) allowing easy visualization of defect types and identifying variances and sample outliers.

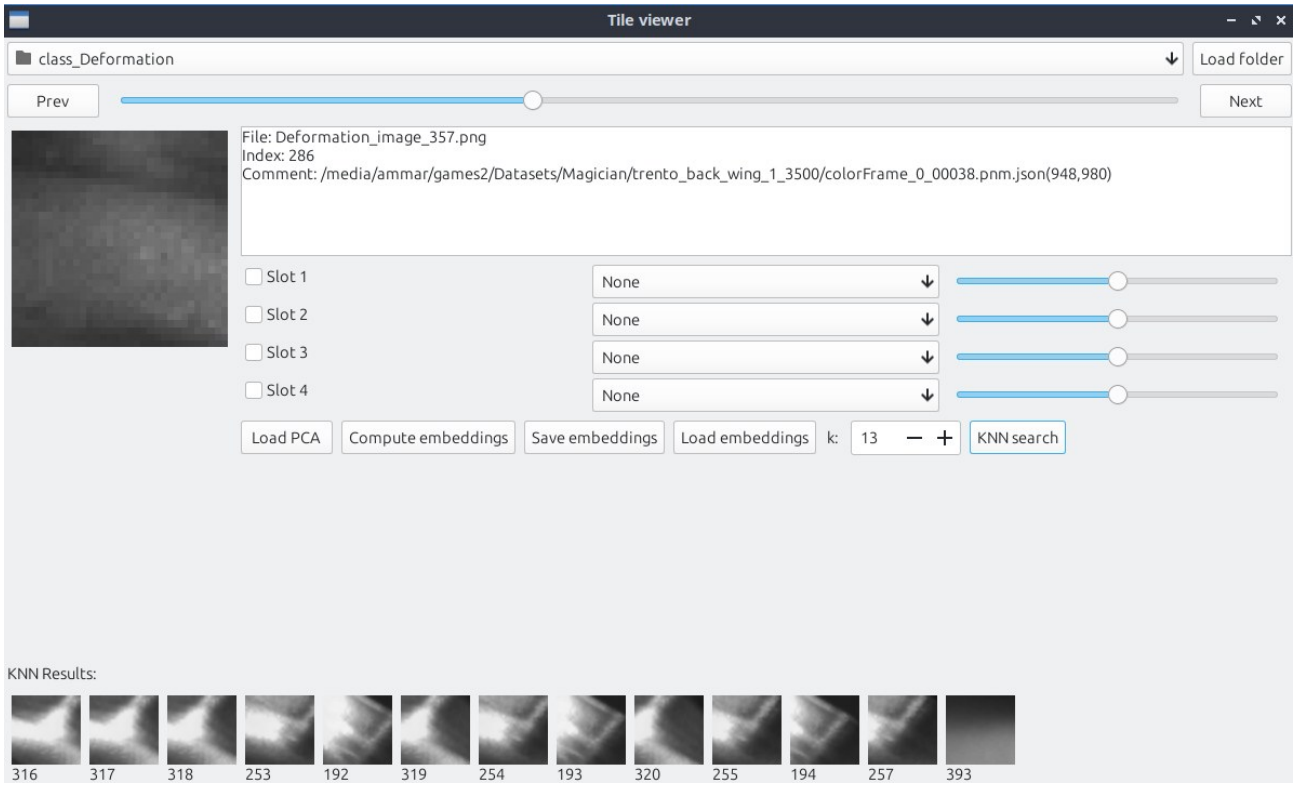


Figure 22: KNN search results in the Tile Viewer: a query tile (top left) with the K most visually similar tiles retrieved from all loaded datasets, supporting exploratory identification of underrepresented defect sub-types

The annotator includes several quality-of-life tools that accelerate the annotation workflow beyond the core point/region labelling. A magnifier panel renders a zoomed view of the region under the cursor, aiding precise placement of annotations on small Class C defects. Per-frame brightness and contrast sliders apply independent tone adjustments to both display panels without altering the underlying data. An auto-advance checkbox automatically steps to the next frame after each annotation is saved, enabling single-click-per-frame throughput for sessions where only one defect is expected per frame. Full UI controls over the loaded classifiers threshold, ensemble composition, majority-vote radius are exposed in the Classifier tab, and one-click benchmark buttons run precision/recall evaluation across the entire loaded dataset without leaving the GUI. Using the camera intrinsic calibration (Figure 12) together with the millimetre-accurate ToF distance sensor reading recorded in each frame's metadata, the annotator can compute real-world distances from pixel coordinates. The operator draws a measurement line (Figure 23) on the image, and the tool reports the corresponding physical dimension in millimetres, providing a sanity check on defect size classification and supporting quality-control reporting.

Finally, the Annotator tool also includes bindings to the same shared memory mechanism of the *magician_grabber* and can be used as a virtual "software defined" camera that provides measurements to the rest of the ROS nodes of the project (Figure

24). This capability allows easy development of ROS packages that rely on the MAGICIAN sensor data. Instead of needing a camera as well as a robot or camera operator during the software development effort ROS node development becomes significantly developer friendly something that was found to considerably speed up development, such as in the case of OC1 Team Multiverse classifier integration.

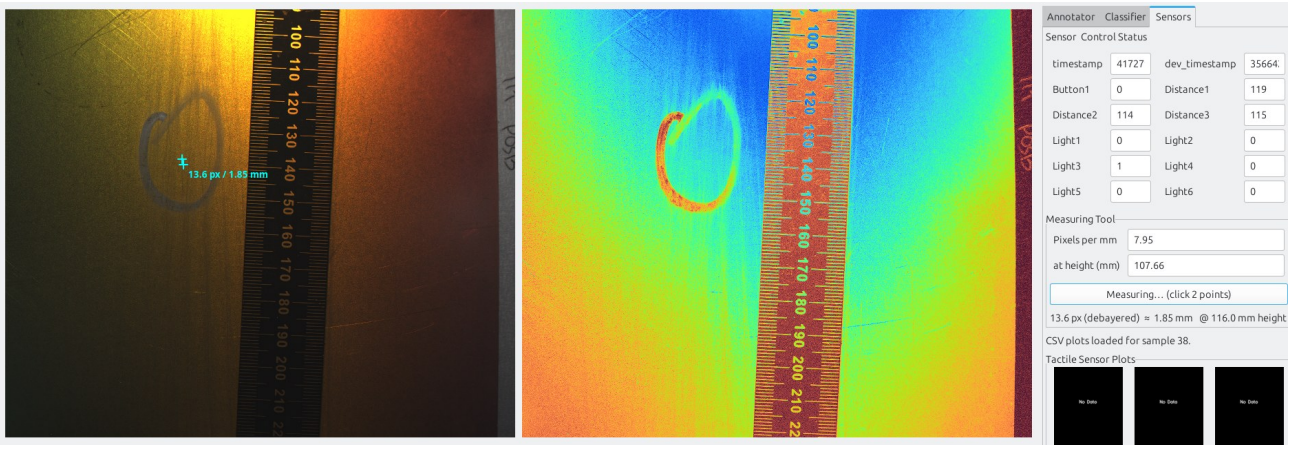


Figure 23: On-image distance measurement in the annotator: the operator clicks on two points on the image and the system draws a line between two points on the polarisation image, and reports the physical distance in millimetres using the stored camera calibration and ToF stand-off distance

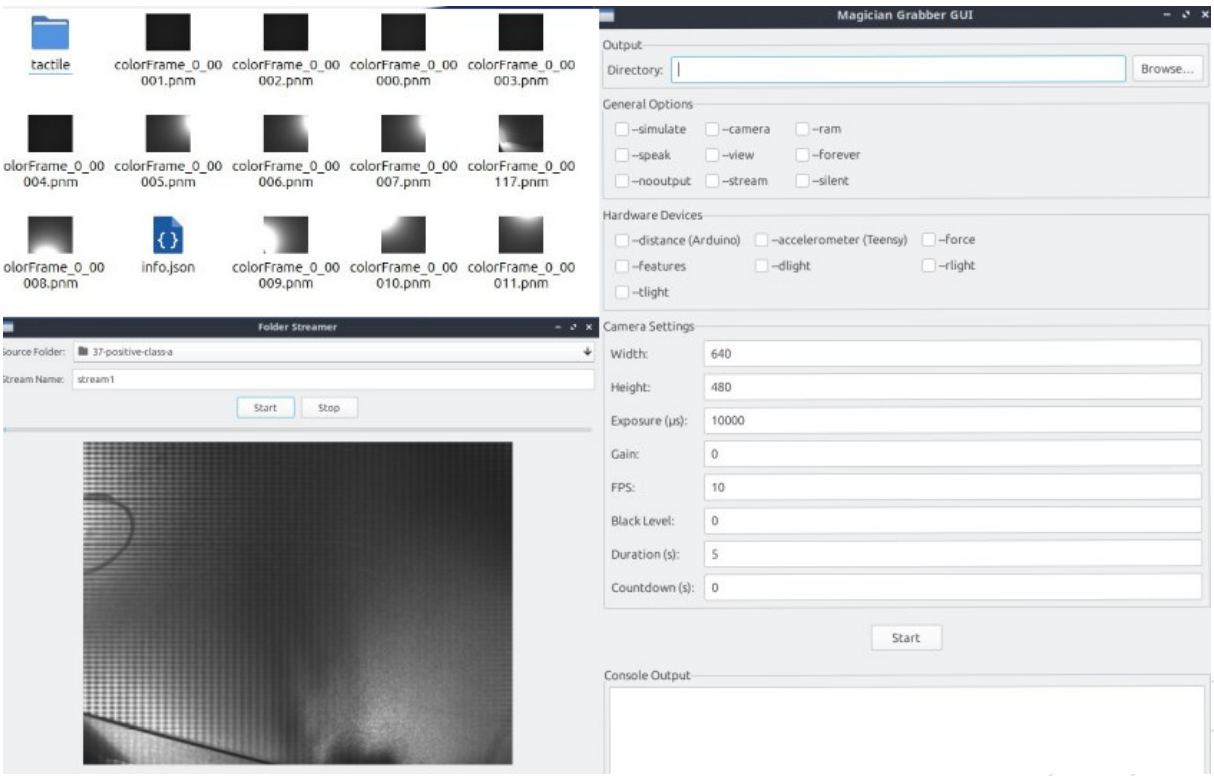


Figure 24: The "Stream To Shared Memory" dialog in the annotator, configured to replay a stored dataset into the shared-memory ring buffer. Downstream ROS nodes receive the replayed frames indistinguishably from a live sensor.

2.5 EMPLOYED METHODOLOGIES

The *magician_vision_classifier* repository (approximately 10k lines of Python, built on PyTorch Lightning) implements the complete training and deployment pipeline for the tile-based CNN defect classifier. The pipeline is designed for the specific characteristics of polarisation camera defect imaging. Spatially localised defect signatures on a large background of clean surface, moderate intra-class variation due to illumination-direction dependence, and strongly imbalanced class distributions (typically < 2% of tiles in any frame contain a defect signature).

The pipeline operates on custom tile sizes extracted from each camera frame which are currently configured to 48 × 48 x 4 pixel tiles but can be trivially altered using the provided GUI (Figure 20). At the nominal de-bayered resolution (1224 × 1024 px), non-overlapping extraction (step = 48 px) yields approximately 525 tiles per frame. In dense-coverage mode (step = 14 px) this increases to 5,950 tiles per frame, and at the standard deployment step size (step = 18 px) to 3,630 tiles per frame.

At the 4.5 - 6 cm robot stand-off distance with the 16 mm lens, a single Class C defect (smallest severity, approximately 300 × 300 µm) occupies 1 tile projecting to approximately 3.2 × 3.2 ≈ 9 de-bayered pixels to 4.3 × 4.3 ≈ 16 pixels. The optics design mirrors the classifier design. The minimum feature size the classifier can respond to is 3 × 3 (Conv2D convolutional receptive field), with larger defects (Class B and Class A, progressively deeper and wider) potentially spanning more than one tile. The step size is the primary parameter governing the accuracy/throughput trade-off at runtime. A smaller step produces overlapping tiles and therefore multiple independent classifier votes over the same physical surface region, statistically increasing the chance of detecting a defect that falls near a tile boundary, at the cost of proportionally more inference calls per frame. A step of 14 px yields 11× more tiles than non-overlapping extraction (5,950 vs. 525), meaning 11× more CNN forward passes and a corresponding increase in GPU load. This makes step size the dominant factor in whether a given backbone can sustain 23 Hz real-time operation: architectures that comfortably exceed 23 Hz at step = 18 may fall below the threshold at step = 14, as shown in (Figure 25).

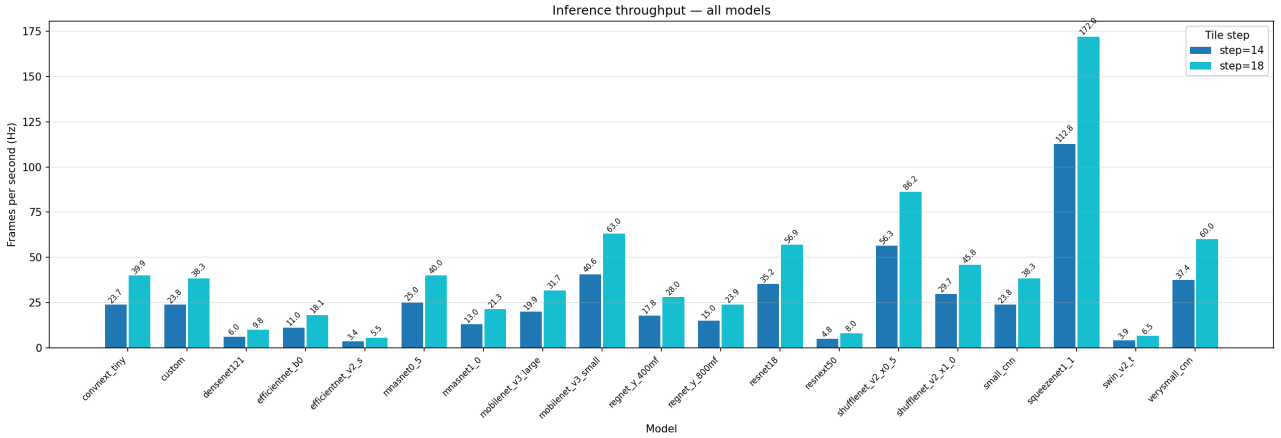


Figure 25: Performance graphs of the various classifiers measured while running on the Magician Pilot PC in ALT premises during the 6th integration meeting. All classifiers in this experiment handle all defect types/severities (one-pass configuration) for two different step sizes (classification densities) applied flatly over the whole image. Framerates above 23 Hz can process the incoming input feed without dropping any frames. For networks beneath the 23 Hz threshold, a larger step size can bring them to an acceptable framerate. The maximum theoretical step size at which all pixels in the image are accounted for exactly once is 48 (non-overlapping). Step sizes 14 and 18 oversample the image by approximately 11× and 7× respectively relative to that baseline, providing multiple independent classifier votes per surface location at the cost of proportionally higher compute load.

The step can therefore be tuned independently per deployment site to balance coverage density against available compute. The overlapping-tile redundancy created by small step sizes also interacts directly with a 2D spatial majority-vote post-processing filter. A physical defect at a tile boundary produces positive responses in multiple adjacent tiles rather than an ambiguous single-tile result. The majority-vote filter which suppresses any positive tile not adjacent to at least one other positive tile becomes statistically more reliable. Sensor noise and specular reflection artifacts can at times produce isolated single-tile activations, while a genuine defect at high coverage density produces a compact cluster of multiple adjacent positive tiles. The combination of overlapping tiling and spatial majority voting thus acts as a spatial consistency check that substantially reduces the false positive rate without sacrificing recall on multi-tile defects.

At the time of writing this deliverable the main volume of data captured and annotated (~130K samples) comes from the V1 camera. From the V2 Camera the main volume of data that has been annotated is captured from ALT Integration meetings and intended to be used as a validation set. The effort required to annotate the remaining (~73K) Camera V2 samples will still require an additional month of work and is expected to be delivered early August, leaving time for training during August. Classifiers trained on Camera V1 data and tested on the validation set of Camera V2 (Figure 26) perform differently than datasets trained on a train/validation split of V2 data (Figure 27). All inputs use the full four-channel (RGBA) polarisation representation (R = 0°, G = 45°, B = 90°, A = 135°). Intensity-only and three-channel variants are supported to simulate monochrome camera configurations with ablation experiments planned to confirm the contribution of the polarisation modality for shallow-defect detection when the new V2

datasets are annotated and available.

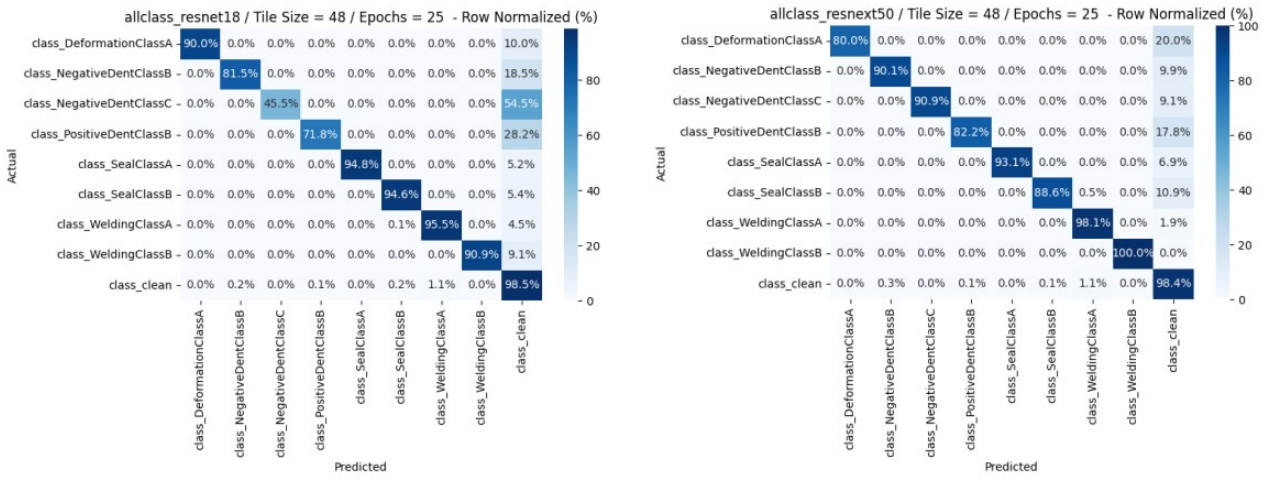


Figure 26: When training on Camera V1 and Evaluating on Camera V2, Resnet 18 and ResNeXT 50 are the top performers. It is worth noting that the vehicle present in the ALT presences features only a subset of defects, for example no Positive Dents of Class C. High visibility defects such as Class A defects and welding spatters seem to be reliably detected, however smaller defects seem to not perform well on this cross-camera experiments.

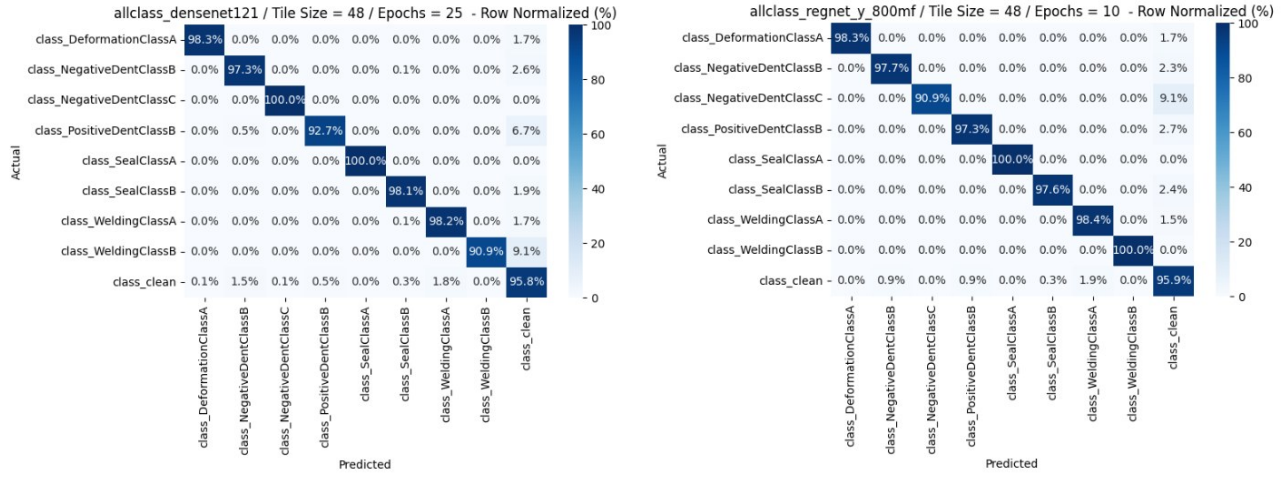


Figure 27: We perform a 5% train/test validation split on V2 Camera data from ALT integration meetings since at the time of writing this deliverable the recently recorded V2 datasets with the newly assembled cameras from FORTH, UNITN and IIT have not yet been annotated. We observe different classifiers taking the lead in accuracy, with Densenet 121 and Regnet 800 mf taking the lead. We also observe much more robust results when comparing to V1 -> V2 cross validation experiments (Figure 26).

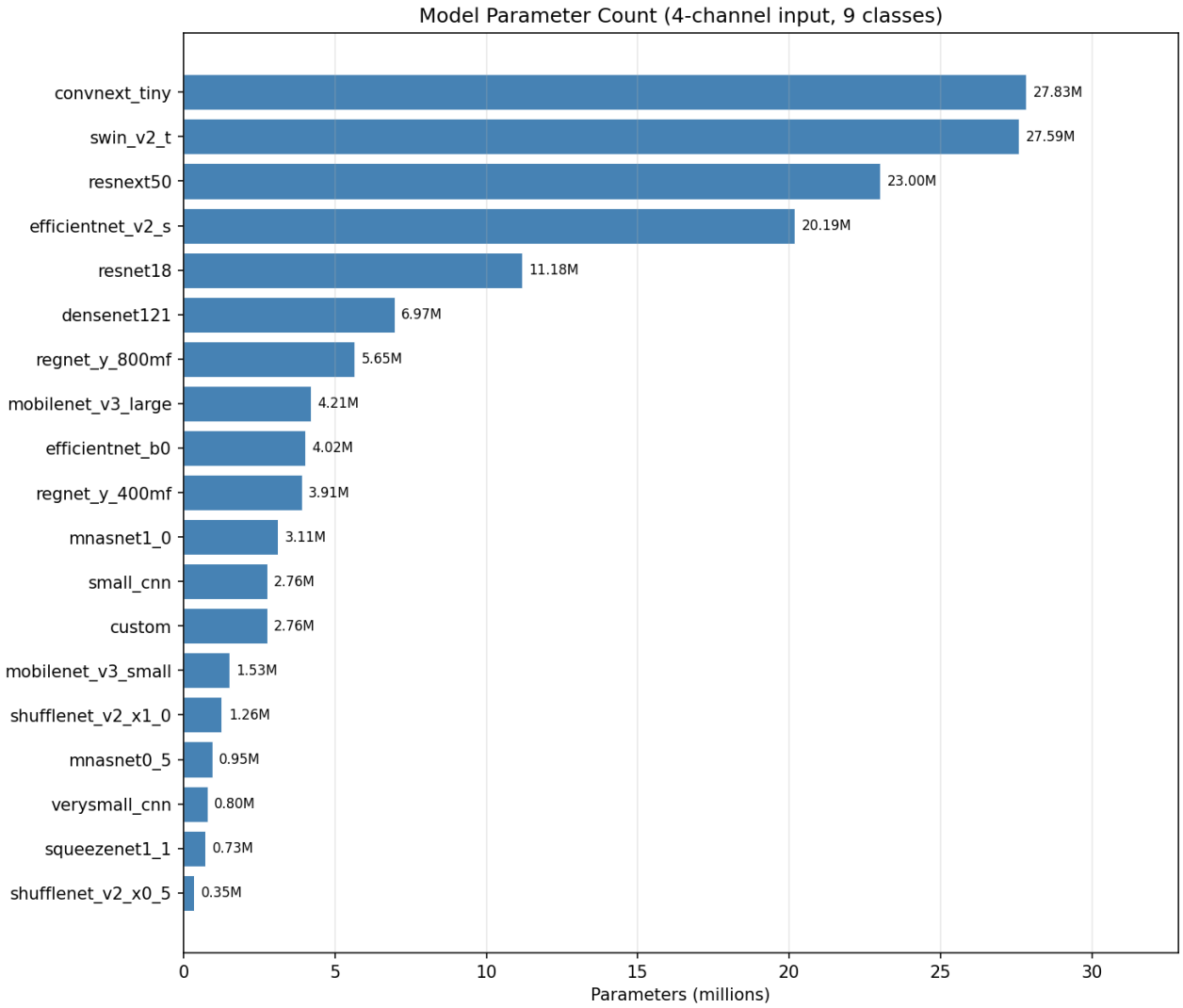


Figure 28: The vision classification neural networks employed and tested for the MAGICIAN pilot use-case. Each of them is currently trained to handle 48 × 48 × 4 inputs in several thousands of training samples and output soft-max classification scores identifying the various defects.

All backbone architectures (Figure 28) are adapted for four-channel input by replacing the first convolutional layer with an initialised four-channel equivalent. The output layer is configured to the number of target classes for the given training mode: nine classes in the full multi-class configuration (Negative Dent Class A/B/C, Positive Dent Class A/B/C, Welding Spatter, Sealing Residual, Clean).

A key design decision in the tiling pipeline is the potential deployment of two complementary classifiers rather than a single one-pass multi-class model. Running a "big-model" for full nine-class (all types and severities) inference on every tile in every frame imposes a substantial GPU load, particularly at high coverage density (step = 14, ~5,950 tiles per frame): even a lightweight backbone requires thousands of forward passes per second, and at 23 Hz the total tile throughput exceeds 100K tiles/second for

the densest configurations. Furthermore, the class distribution is heavily skewed, typically fewer than 2% of tiles contain any defect meaning the vast majority of inference compute would be spent on clean tiles that contribute nothing to the detection output.

Our tiled based approach offers an elegant solution to this problem by employing a two-class binary prefilter trained with a high-recall, low-precision objective (Clean / Non-Clean) that discards 80 - 90% of tiles per frame before any fine-grained classification is attempted. Only the small fraction of tiles flagged as candidate defect regions are then passed to the full "all class/severity ensembles", which performs the detailed defect-type and severity classification presented as the output to the rest of the system. This cascade structure concentrates compute where it is needed, makes the system scalable to denser tiling and larger ensemble sizes, and allows the Stage 1 and Stage 2 models to be trained and tuned independently for example, maximising recall in Stage 1 (accepting more false positives) while optimising precision/recall balance in Stage 2. Having an efficient way to narrow down an image to, e.g., 100 candidate defect tiles leaves us with enough computing resources to even run all trained classifiers listed in Figure 28 for a very well-informed classification decision.

The two-stage cascade described above (binary → all class) is not the only decomposition the architecture supports. A natural extension (not currently prioritised but straightforward to implement) is an even more elaborate three-stage hierarchy that inserts an intermediate class-level discriminator between the binary (clean/defect) prefilter and the fine-grained (all class/severity) classifier. Rather than asking a single nine-class model to distinguish both defect type and severity simultaneously, the intermediate stage would first route each candidate tile to its defect family Negative Dent, Positive Dent, Welding Spatter, or Sealing Residual using a small four-class model. Each family would then be handed off to a dedicated expert network trained exclusively on that class, whose sole task is severity grading within the family (e.g., Class A / B / C for dents). This hierarchy is motivated by the observation that genuine defects are rare events. With stage 1 already eliminating 80 to 90% of tiles, only a handful of candidates per frame would reach the downstream stages. Once a tile has cleared the prefilter, the cost of running one additional lightweight routing step is negligible, yet it unlocks substantially richer specialist capacity. An expert network for Positive Dents, for example, can devote all its representational budget to the specific surface-texture and curvature signatures that separate Class A from Class B from Class C severity without being forced to simultaneously distinguish welding spatters. This separation of concerns is already supported by the developed framework allowing models with different class subsets to be evaluated interchangeably in the same pipeline. The cascade also reduces the effective training data requirement per expert: a model that sees only Positive Dent tiles during training faces a much more balanced class distribution (three severity classes in roughly equal proportions) than a model that must handle the full nine-class imbalance. The modular design means each expert can be retrained or swapped independently as new annotated data accumulates for a given defect family. The full training and inference pipeline is packaged as a Docker container that packs a Python Virtual Environment dependency system making deployment reproducible across operating

systems and hardware configurations without requiring manual dependency installation. The code can thus be executed with or without Docker for maximum ease of use. The container bundles PyTorch, PyTorch Lightning, the MAGICIAN classifier source, and all runtime dependencies. This was particularly relevant for deployment at partner sites (ALT Pilot PC, YongaTek, Multiverse OC1 winners etc.) across different host operating system (e.g. Macs for Team Multiverse) and GPU driver and CUDA/CUDNN versions that are different from the FORTH development environment.

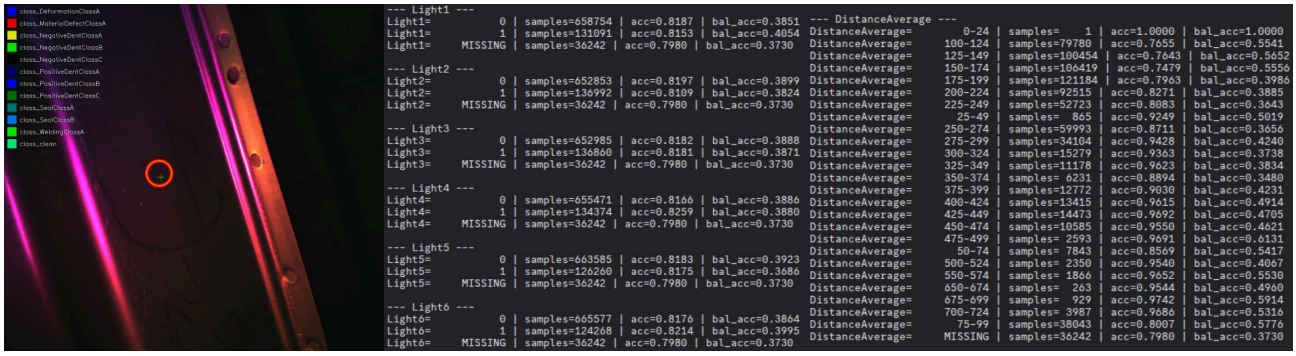


Figure 29: All available data from the annotator tool gets infused into the training .PNG tiles and allows fine-grained analysis of results based on criteria such as the active light, distance from defects etc.

Each exported 48 × 48 tile carries an embedded PNG metadata field encoding the source image path, pixel coordinates, annotation provenance and camera status. The classifier toolchain can parse these metadata fields to extract per-dataset, per-defect-class, and per-severity statistics (Figure 29) across the full tile corpus without requiring access to the original full-resolution images. This enables dataset coverage analysis e.g. identifying which defect severity/class combinations are underrepresented in the training set or validating uniform operation across different lights, distances that work better for various defect type classifications etc.

2.6 RESULTS AND FINDINGS

The successful implementation of the MAGICIAN visual perception classifier builds on multiple complementary design pillars. The optics are carefully selected to deliver image features whose physical scale at the KPI level corresponds to the minimum detectable feature for a 3×3 Conv2D operation. The LED illumination is specifically designed to suppress ambient light while delivering a strong, repeatable polarimetric signal. The global shutter sensor operating at 650 µs exposures provides crisp frames that are robust to the fast arm motions and workshop vibrations encountered in an industrial deployment. The sensor field of view covers as much panel area as possible per scan pass, and the selected sensor saturates the bandwidth provided by the GigE standard for industrial cameras. The four polarisation channels recorded by the sensor provide fine-grained per-pixel evidence about how light reflects over surfaces a discriminative modality for the shallow surface deformations targeted in this project. The tile-based classification approach avoids overfitting on the prominent pen markings that workers

apply to metal sheets, while simultaneously enabling the accumulation of enough labelled samples to satisfy the data-hungry deep learning methods employed. Massively parallel batched tile classification potentially across multiple architectures running concurrently on the same frame ensures throughput adjustable to available computational resources. The careful annotation and data analysis conducted using the annotator tool ensures a user-friendly framework for producing the high-quality labelled data required by the rest of the system.

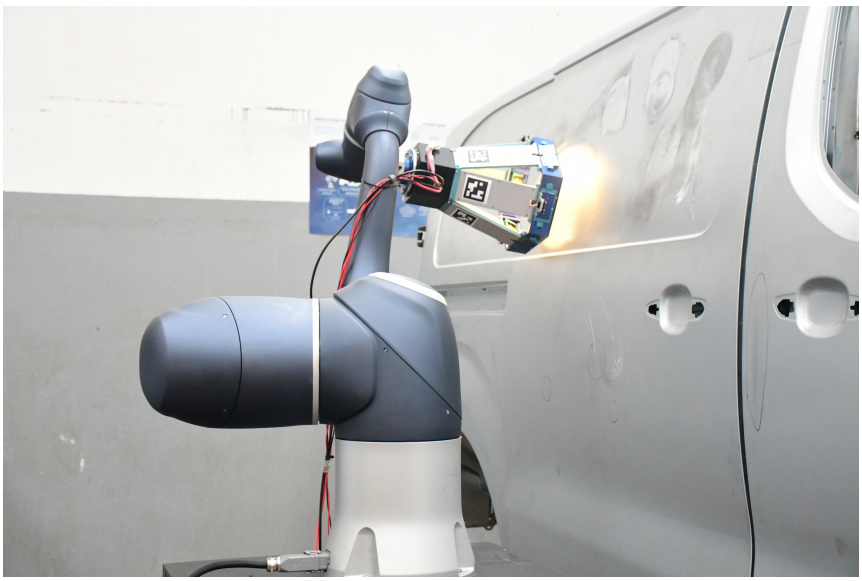


Figure 30: Defect detection running in real-time on the MAGICIAN platform tested on ALT premises during the 6th integration meeting.

Our preliminary results on the integrated robot (Figure 30) show that the design and implementation of the camera sensor, annotator and vision classifier are on track to tackle the very challenging detection problem and the demanding KPIs set for the MAGICIAN project. That said the complications in fabricating the new Camera V2, delays in shipment to and from the various partners (e.g. ALT, IIT and FORTH), the postponement of scheduled integration meetings in Turkey due to safety concerns stemming from the recent Iran war have compounded in delayed dataset capture which prevented us from being able to annotate the data in time. We thus have not yet been able to train on a complete training dataset based on the available metal samples provided by the consortium from TOFAS. The annotation and finalization of the V2 dataset and a full retraining campaign are scheduled after the submission of this deliverable.

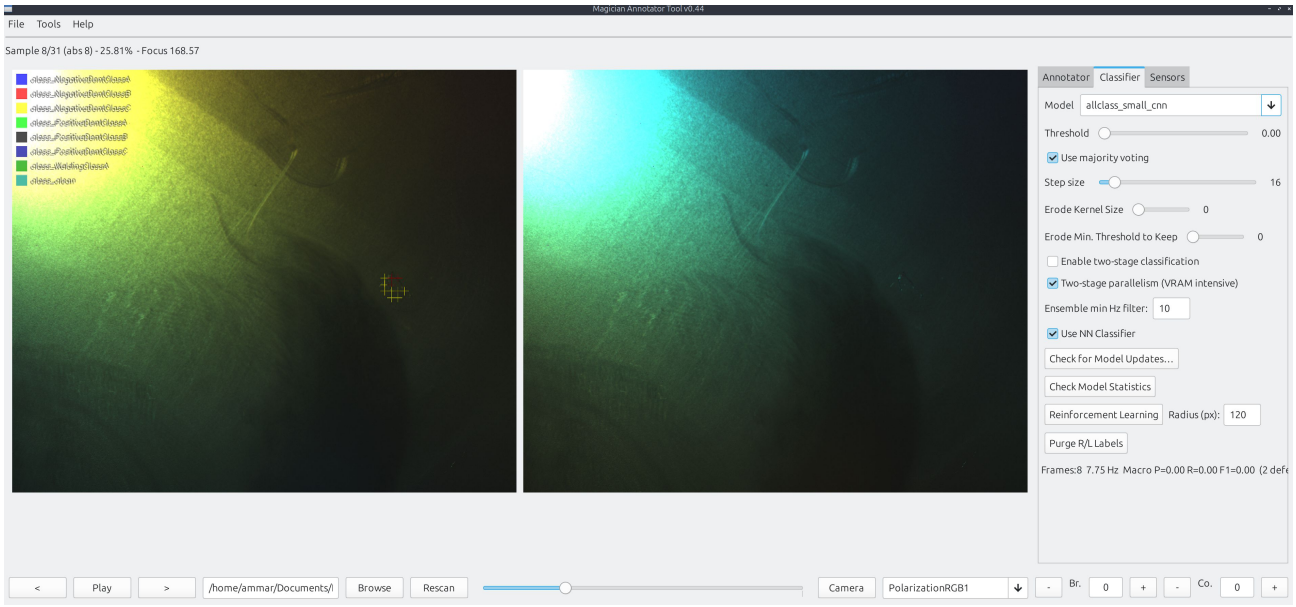


Figure 31: Detection of a novel defect (created using a screwdriver and a hammer by an expert technician from TOFAS) during ALT trials during the 6th integration meeting. The defect is not part of the training corpus, and we can see the network classifiers agreeing on the Defect Type (Negative Dent) but slight disagreement between different tiles, with some detecting it as Class B and some as Class C.

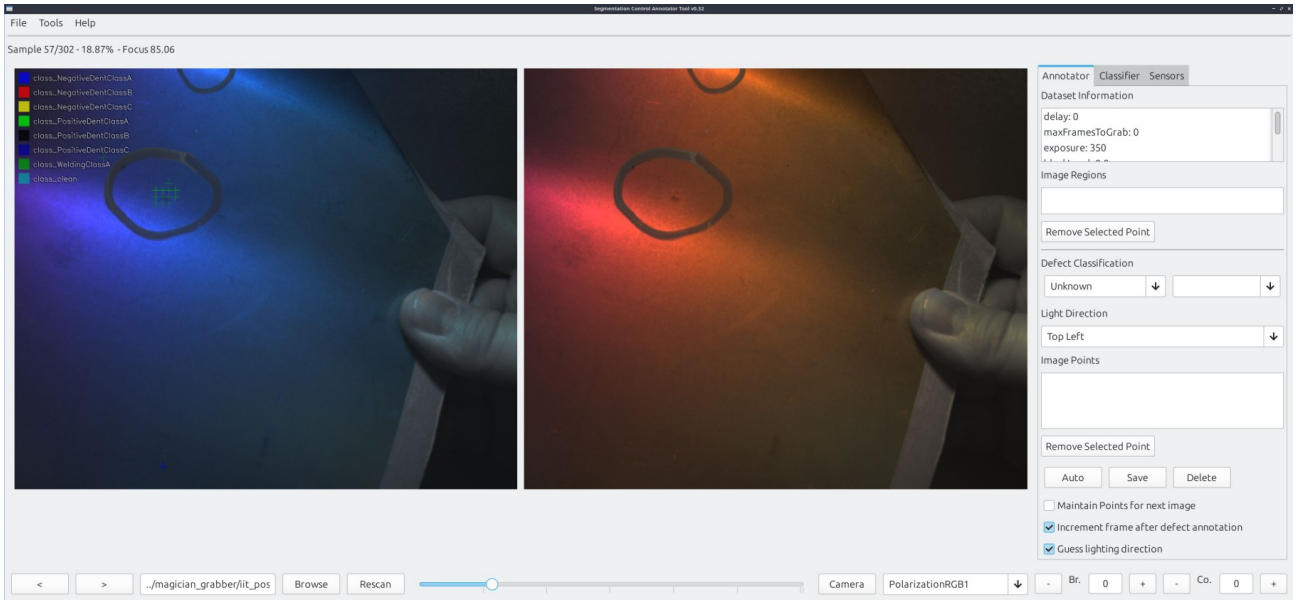


Figure 32: Detection of novel unannotated defects from metal samples shipped to IIT from TOFAS. Using a step size of 16, typical detections output for "large" Class A, defects look like this example with multiple detections finely covering the area of the defect offering a dense localized classification that can be used to perform the repair.

During the 6th integration meeting at ALT, an expert technician from TOFAS created novel defects on a panel using a screwdriver and hammer, a creation technique not

represented in the available training corpus and produced a deformation geometry different from the standardised dents used during dataset collection. Figure 31 shows the classifier output for this defect. The ensemble agrees on the defect type (Negative Dent), demonstrating that the learned polarimetric signature for surface deformation generalises to novel creation techniques (Figure 32). Minor disagreement on severity some tiles classifying the defect as Class B, others as Class C is consistent with the irregular geometry produced by a non-standardised defect created with a tool, and with the limited number of Class C training examples in the current corpus. Despite the severity ambiguity, the spatial cluster of positive tiles localises the defect very accurately allowing ergodic repair trajectory guidance.

The classifier evaluation framework was applied to the Camera V2 dataset using a 5% train/validation split on the ALT integration data. Figure 33 shows per-model precision over the training run: all architectures achieve above 90% precision from the first training epochs, confirming rapid convergence even with the limited V2 corpus available at this stage.

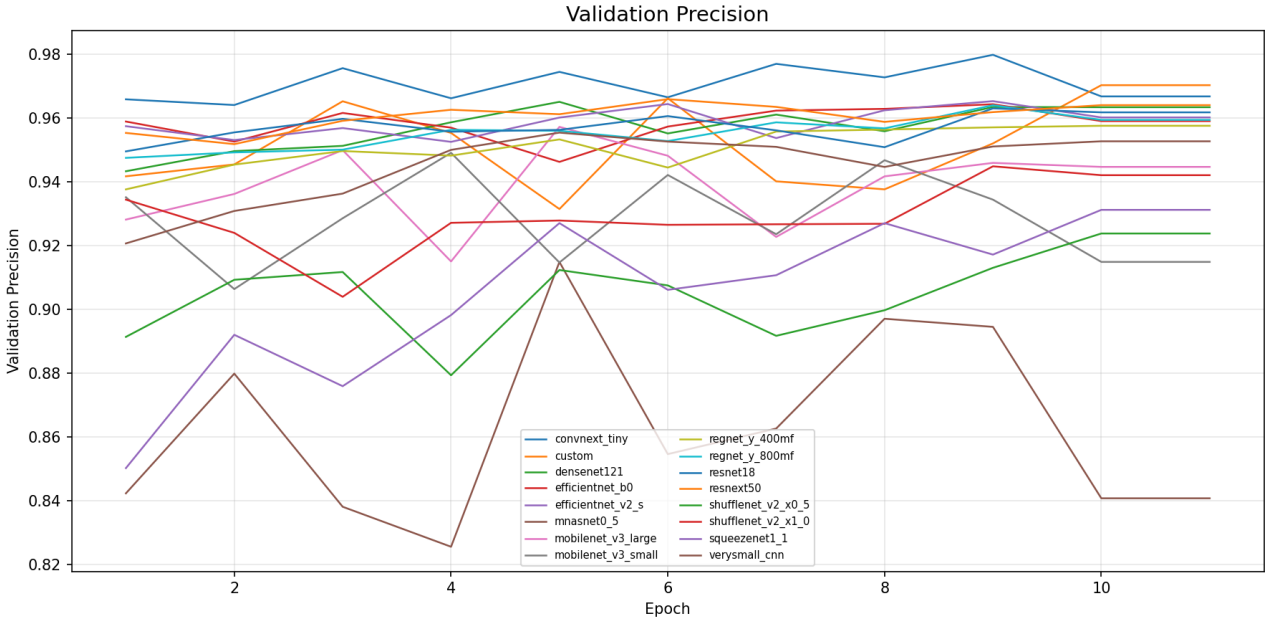


Figure 33: The plot shows the precision of each model in the validation set when doing a 0.05 train/validation split on available Camera V2 data.

Figure 34 shows per-model recall: all architectures maintain low false-negative rates throughout training, meaning missed positive and negative dents on the metal samples are rare. Figure 35 shows accuracy over the training run, with all architectures exceeding 90% even after the first epoch. It is worth noting that the current V2 training and validation set is relatively small. To avoid overfitting, a deliberately short 12-epoch training schedule is used. As more annotated data become available through the planned August annotation campaign, the number of epochs will be increased to match the

longer V1 training schedule.

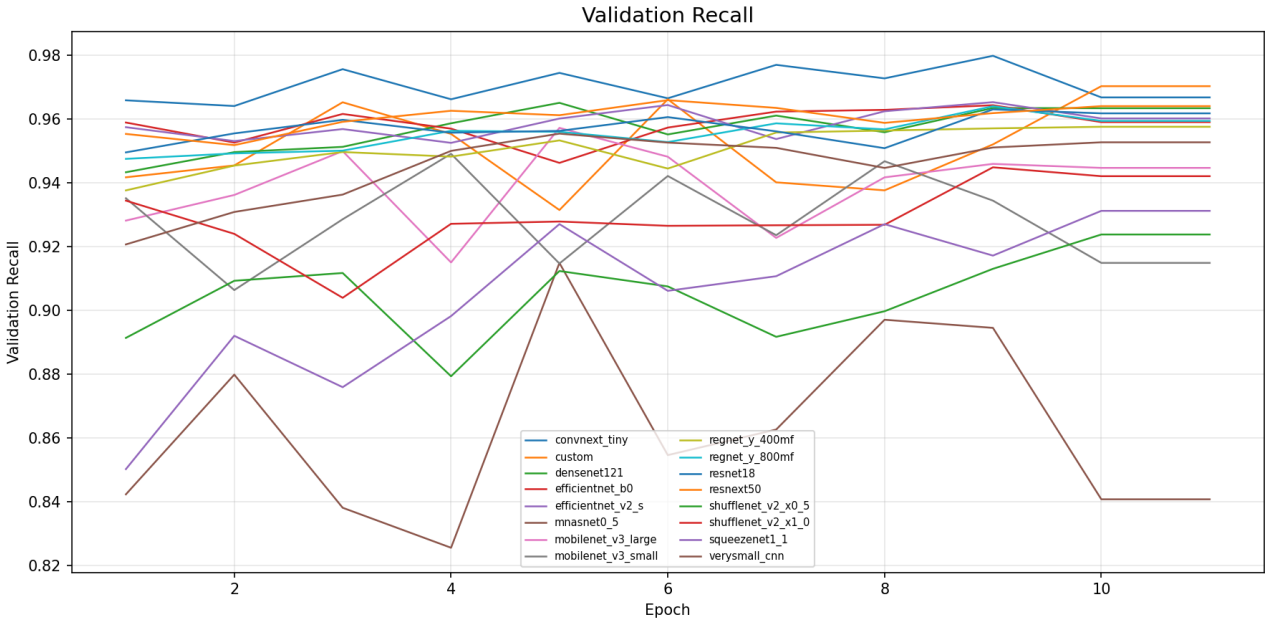


Figure 34: The plot shows the recall of each model in the validation set when doing a 0.05 train/validation split on available Camera V2 data.

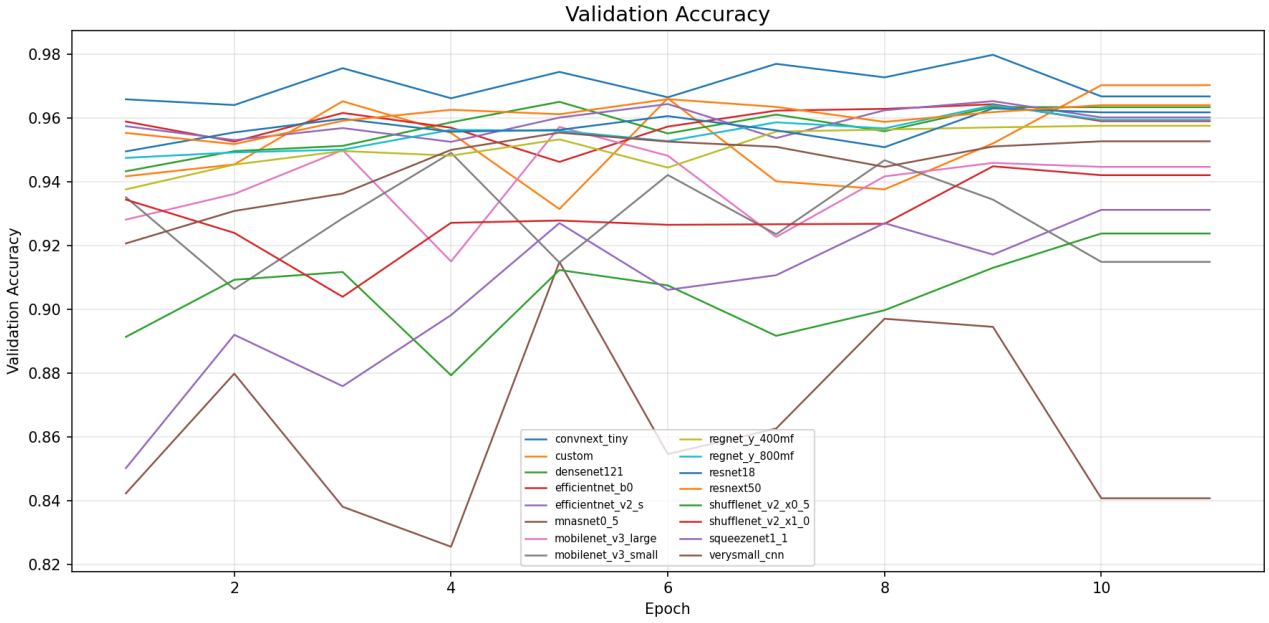


Figure 35: The plot shows the accuracy of each model in the validation set when doing a 0.05 train/validation split on available Camera V2 data.

The classifier framework now supports 18 backbone architectures, with the expanded set reflecting architectures identified as strong performers in the cross-camera evaluation, in particular RegNet-Y-800MF and DenseNet-121, which showed leading

performance on V2-only data (Figure 27). Improvements to batch and class balancing, included weighted sampling and focal loss tuning and were implemented during this reporting period to reduce the effect of the severe Clean/Non-Clean imbalance on training dynamics. The metadata-based statistics extraction capability (Figure 29) was used to monitor per-class sample counts and offer per-dataset quality metrics.

We performed extensive benchmarks (Figure 36) plotting accuracy as a function of model size (millions of trainable parameters) for all 18 evaluated architectures. The figure reveals a wide spread of efficiency operating points: some architectures achieve balanced accuracy comparable to the largest models while consuming a fraction of the parameters. In the MAGICIAN use case where inference must remain real-time at 23 Hz on the Pilot PC while a complementary pose estimation node and environment sensing node also occupies the GPU (among the nodes for motion prediction, trajectory planning etc.) this efficiency dimension can be regarded as comparably important as the raw accuracy of the models. Architectures on the Pareto frontier of the balanced-accuracy/parameter-count trade-off are the natural candidates for ensemble members in resource-constrained deployments and for the fast stages of the binary/all-class cascades.

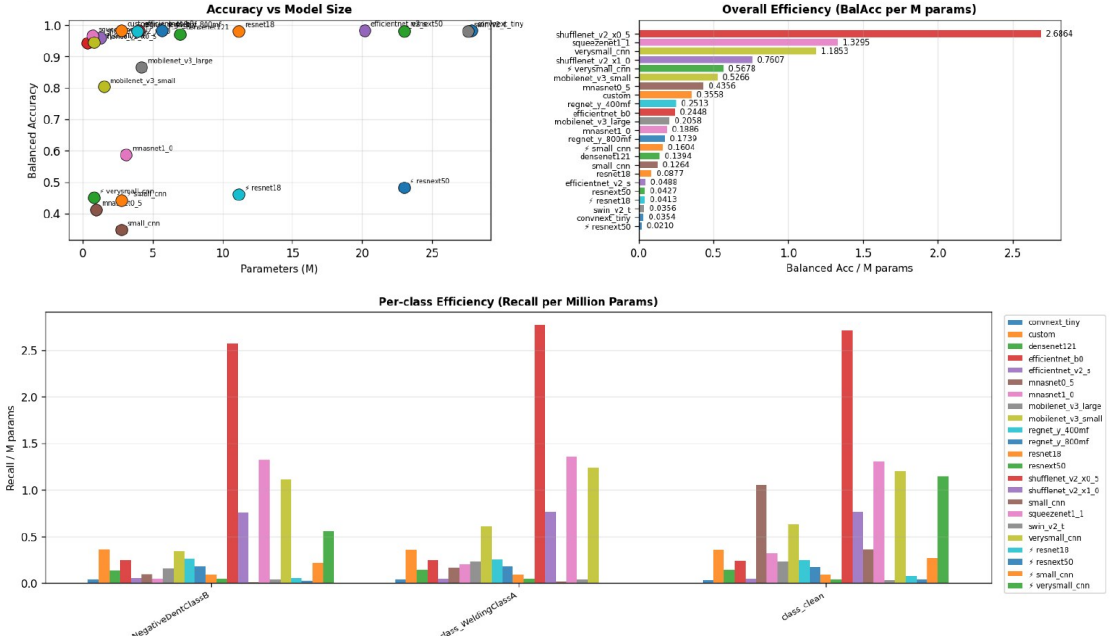


Figure 36: Efficiency graphs showing balanced accuracy when also considering the model size (Millions of parameters). We observe different efficiencies (ideally, we would want a network that offers very good accuracy without needing a lot of computation) that hint at the architectures that can competently handle our classification tasks without wasted computations.

Experiments, compared the achieved framerate of all classifier architectures at step sizes 14 and 18, measured on the Pilot PC during the 6th ALT integration meeting. Architectures that comfortably exceed 23 Hz at step = 14 can be deployed in the dense-

coverage configuration; those that fall below the threshold can recover real-time operation by switching to step = 18, with only a moderate reduction in spatial coverage density.

Figure 33, Figure 34 and Figure 35 shows strong performance in the precision, recall and accuracy when training on the currently available Camera V2 dataset. In our case indicates that all models have a low false negative percentage which means that they do not miss the positive and negative dents on the metal samples.

Figure 37 provides a richer view of the throughput landscape by simultaneously varying batch size (x-axis), step size (z-axis), effective frames per second (y-axis) and achieved accuracy (colour). Even a very fast network invoked serially (batch size 1) performs poorly compared to a slower network that can fill the GPU and execute as a single large batch, utilising all available CUDA cores. The figure captures both how parallelisable each model is and the contribution of step size to boosting classification quality through multiple overlapping votes. It is therefore used to select batch and step configurations per target deployment site.

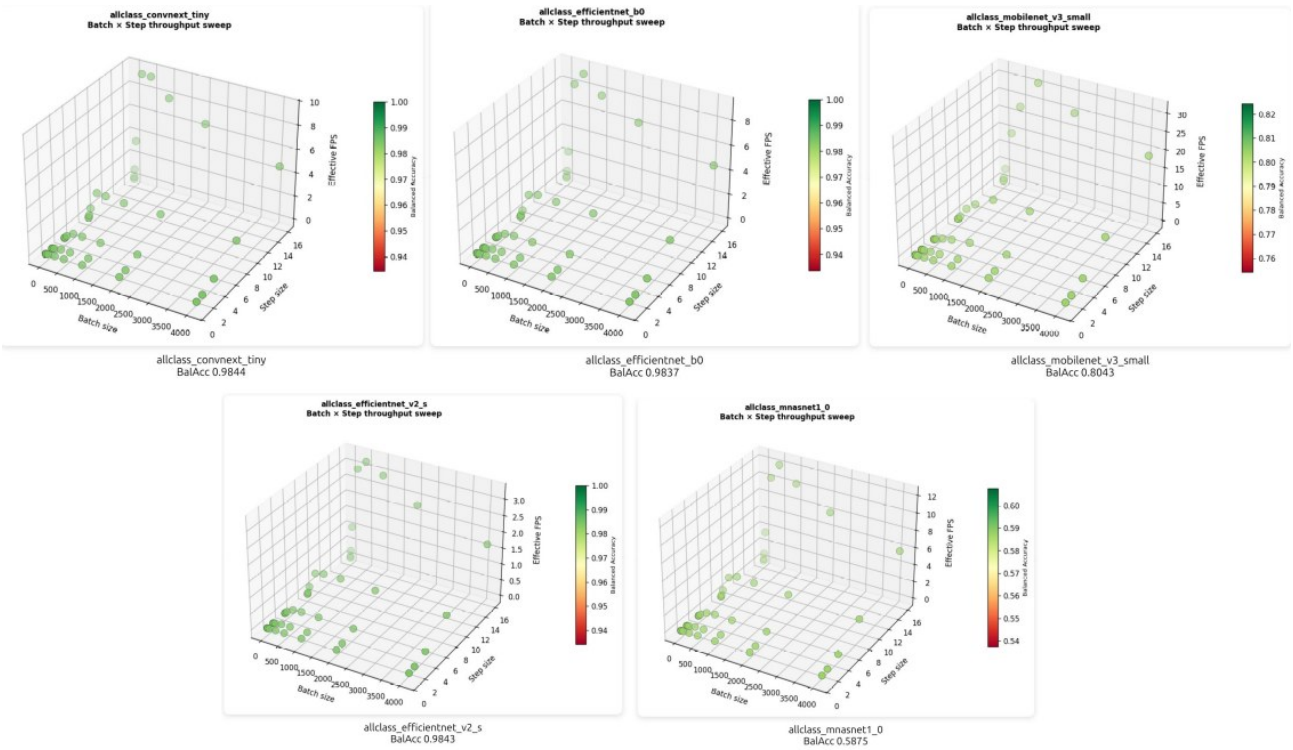


Figure 37: 4D throughput characterisation: batch size (x-axis), step size (z-axis), effective FPS (y-axis), and achieved accuracy (colour channel) for all evaluated architectures. The plot captures how well each model parallelises on the available GPU and quantifies the accuracy contribution of step size

Figure 38 shows our best-performing model configurations experimentally derived from 18 candidate model/hyperparameter combinations evaluated across 787,238 tile samples recorded from the vehicle door frames tested at ALT during the 6th Integration

Week. The door featured a Negative Dent of Class B severity and Welding Spatters of Class A severity. All top configurations show small off-diagonal entries for both present defect classes, confirming that false positive rates on clean tiles and false negative rates on both defect types remain within operationally acceptable bounds.

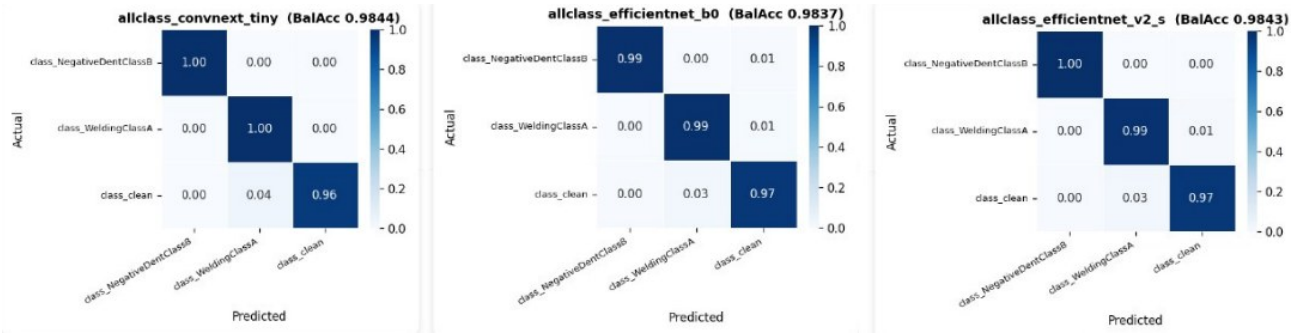


Figure 38: Confusion matrices for the best configurations experimentally derived from 18 candidate models, evaluated on 787,238 tile samples from the ALT vehicle door on the side of the robot that features a Negative Dent (Class B) and Welding Spatters (Class A).

Taken together, Figure 37Figure 37: and Figure 38 demonstrate that the MAGICIAN classifier can process the full incoming frame stream at real-time rates while maintaining the >90% precision and recall in accordance with the requirements of the project KPIs.

2.7 CHALLENGES AND LIMITATIONS

The previous deliverable D3.2 identified four challenge areas for the visual perception component. All four have been substantially addressed during this reporting period. The 3D-printed sensor chassis that transmitted vibrations at high robot speeds has been consolidated, and the bespoke MAGICIAN Main Board PCB replaces the earlier breadboard harness, reducing mechanical coupling between cable slack and the sensor. The illumination challenge (ambient factory lighting threatening to overwhelm the polarimetric signal) has been addressed through sub-millisecond LED pulse strobes synchronised to the 650 µs camera shutter, which reliably suppress ambient light even under the high-intensity overhead sunlight encountered at ALT and UNITN facilities. The focus and distance-sensing challenge has been resolved by a fixed-focus strategy supported by three time-of-flight sensors now operating throughout continuous deployment, confirming that the fixed-focus constraint is met under normal trajectory conditions. Finally, the risk of the classifier learning pen-mark contours rather than defect signatures has been mitigated through the tile-based classification approach combined with scanning samples at multiple partner sites. And the out-of-distribution generalisation results at ALT (Figure 31) confirm that the learned representations transfer to panels and defect instances not seen during training.

Two primary limitations remain for the integration activities:

1. **Training Data Scarcity**

- The primary bottleneck for classifier generalisation performance is the limited size and diversity of the annotated training dataset. Despite the hundreds of kilograms of material shipped to FORTH, IIT and UNITN by TOFAS to provide training data, the data are still ultimately limited, and the coverage of defect types, severities, surface finishes, and illumination conditions remains narrow relative to the full range of automotive production scenarios. Exotic defects such as sealing residues and material deformations are also vastly underrepresented having 1 to 3 exemplars per partner. A systematic data acquisition, annotation and classifier retraining protocol at TOFAS (where hundreds of car bodies per day pass through the production line) would produce a substantially larger and more robust classifier.

- Another issue is the physical degradation of the metal sheets shipped to partners during the early stages of the project. Dust and particulate matter from storage areas has accumulated on the metal sheets prompting careful cleaning of the metals before the recapturing of data with the V2 Camera. However, due to the metals not having protective paint the texture of the metals is in much less pristine state compared to the factory, it may be possible that this introduces some domain shift in the training data.

- A final issue regarding the training data is that metal sheets that are "grinded" using the repair robot effectively are no longer available for being re-recorded with updated versions of the sensor, different exposure settings etc. Similarly repetitive tactile scanning using the plastic scallop also slowly alters the metal surface. The metal sheets available at FORTH are particularly affected by this since they have sustained 6 months of tactile experiments.

2. **Sensor optics and hovering distance**

- The optics are selected to maximise compliance with the project KPIs. The current 16 mm fixed-focus lens is optimised for a 4.5 cm – 6.0 cm robot arm standoff, at which a 300 μm Class C defect subtends approximately 4.3 × 4.3 pixels - 3.2 × 3.2 pixels. Considering the minimum detectable feature for a 3 × 3 Conv2D kernel, at larger standoff distances the field of view increases (relaxing per-pass coverage requirements), but angular resolution decreases and the pixel footprint of the smallest defects shrink below the detectability limit. Sensing accuracy thus (especially for Class C defects) is very sensitive to the ergodic control and the hovering distance of the robot arm.

- Even with the robot hovering perfectly at exactly the intended distance Class C defects fall on the detectability limit. Switching to a higher focal length, for

example 25mm, 35mm, or even 50mm would increase the relative size of the minimum magnitude defect to 11x11 px, 18x18 px and 31x31 px (while at the same time drastically reducing the effective area covered) and mandating using multiple sensing robots for timely execution of the scanning procedure.

3 TACTILE PERCEPTION SYSTEM FOR IMPERFECTIONS DETECTION

3.1 INTRODUCTION

The tactile perception system constitutes the second fundamental component within the MAGICIAN project, complementing the vision system in the detection and classification of surface imperfections. The core challenge addressed by this module is to reproduce, in a robotic and repeatable way, the informative aspects of manual tactile inspection, where human operators rely on contact forces and vibration cues to detect and characterize subtle surface defects that may be difficult to perceive visually.

In D3.2, IIT developed and validated the first tactile acquisition and annotation pipeline, based on handheld exploration of TOFAS-annotated car body frames. That activity enabled the construction of a synchronized tactile dataset including force, acceleration, and motion data, and provided the basis for studying human exploration strategies during surface inspection.

In D3.3, the tactile perception activity progressed from handheld data acquisition toward robotic validation and system integration. The Tactile Perception Module was integrated on a robotic platform and evaluated through controlled surface exploration, where the effects of exploration modality, contact force, and scanning speed were systematically assessed. Human-guided demonstrations collected in the previous phase were used as motion priors for robotic reproduction, while standardized robotic trajectories were introduced as repeatable benchmarks. The activities reported in this deliverable also benefited from the MAGICIAN integration weeks, during which the tactile perception pipeline was further consolidated in view of its integration with the broader sensing and robotic inspection framework. These integration activities supported the transition from standalone tactile acquisition toward a more complete robotic inspection workflow, aligning the tactile module with the overall MAGICIAN objective of combining complementary sensing modalities for robust defect detection and classification.

3.2 STATE OF THE ART

The state-of-the-art review provided in D3.1 and recalled in D3.2 remains valid as the scientific and technological foundation for the tactile perception activities carried out

within WP3. The previously discussed topics, including tactile sensing technologies, multimodal sensor fusion, haptic feature extraction, and data-driven classification, continue to support the development of the MAGICIAN tactile perception module.

In D3.3, the focus moves from the development of the handheld acquisition and annotation pipeline toward the robotic validation of the tactile sensing system. In this context, the most relevant research directions are those related to multimodal tactile inspection, robotic reproduction of human exploratory strategies, force-regulated surface scanning, and the evaluation of learning-based classifiers under different operating conditions. These aspects are directly addressed in the robotic campaign reported in this deliverable, where the Tactile Perception Module is integrated on a robotic platform and evaluated by varying exploration modality, contact force, and scanning speed.

3.3 OBJECTIVES AND REQUIREMENTS

The objectives defined for the tactile perception system remain aligned with those outlined in D3.1 and D3.2, and with the Key Performance Indicators (KPIs) established in D2.1 – "Use Case Definition". These KPIs (Table 2) continue to represent the reference metrics for evaluating the MAGICIAN tactile perception module and its contribution to robotic defect detection on automotive metallic surfaces.

While the general objectives remain unchanged, the work reported in D3.3 represents a significant step forward in their implementation. In D3.2, IIT developed and validated the handheld acquisition and annotation pipeline, enabling the construction of a synchronized tactile dataset based on force, acceleration, and motion data. In D3.3, this framework is extended toward robotic inspection: the Tactile Perception Module is integrated on a robotic platform, tactile data are acquired under controlled and repeatable conditions, and the influence of exploration modality, contact force, and scanning speed is quantitatively assessed.

The tactile system is therefore no longer evaluated only on manually acquired data, but also under robot-executed inspection conditions. This allows IIT and the MAGICIAN consortium to assess how the robotic exploration strategy and the physical interaction with the surface affect defect detection and classification. In this phase, the classification task focuses on defect-type recognition, namely no defect, positive dent, and negative dent, while defect-severity estimation remains a future development.

Scientific and technological objective	KPI ID	KPI definition	After MAGICIAN
(O1) A robotic perception module integrating visual and tactile sensors. The module will be embedded in a robotic	O1-KPI-SR1	Smallest size of defect that can be sensed/detected by the perception module.	≤0.3mm
	O1-KPI-SR2	Detection success rate vs	False positives: ≤120%

sensor module (the SR, hereafter) and will be used for defects analysis and classification. The SR will replicate the skills of human workers through a learning scheme.		humans.	Skipped defects: ≤110%
	O1-KPI-SR3	Car-body scan time compared vs humans on a benchmark set.	≤110%
	O1-KPI-LRN SR1	Misclassification rate with respect to humans.	≤10%
	O1-KPI-LRN SR2	Time to convergence.	Observation time ≤ 15h to achieve KPI LRN-SR1

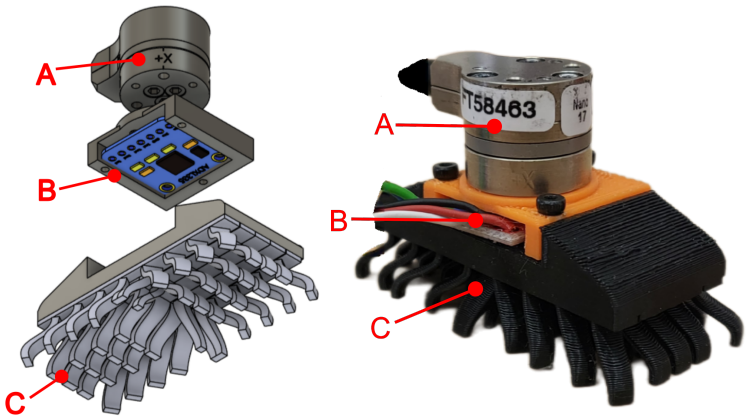
Table 2: KPIs related to the Tactile Perception System for Imperfections Detection.

3.4 TACTILE SENSORS

The tactile sensors employed in this phase remain consistent with those described in D3.1 and D3.2. As shown in Figure 39, the tactile system is based on a compact multimodal sensing module designed to reproduce the informative aspects of human tactile inspection. IIT integrated two complementary sensing modalities: a six-axis force/torque sensor, ATI Nano17, and a tri-axial MEMS accelerometer, Analog Devices ADXL335. The force/torque sensor captures quasi-static interaction forces and moments associated with the global contact state and surface geometry, while the accelerometer records high-frequency vibrations generated by local discontinuities, surface irregularities, and transient contact events.

The sensing hardware is integrated into a custom tactile end-effector, the Tactile Perception Module (TSM), whose exploded and assembled configurations are reported in Figure 39. The TPM includes a compliant contact interface fabricated from thermoplastic polyurethane (TPU) with a Shore A 95 hardness. This material provides a semi-rigid interaction with the inspected metallic surface, ensuring non-destructive contact while preserving sufficient mechanical stiffness to transmit force and acceleration transients to the sensors.

The contact interface has a rectangular footprint of approximately 21 x 50 x 8 mm and includes 44 S-shaped pins, each 8 mm high and with an approximate contact area of 4.5 mm^2. The S-shaped geometry was selected to improve the mechanical transmission of surface-induced vibrations. Unlike a rigid vertical probe, the compliant curved structure behaves as a spring-like contact interface: when the probe slides over a surface defect, the deformation and elastic recovery of the pins generate dynamic signatures that can be captured by the accelerometer. This design allows the tactile module to amplify subtle surface variations that may be difficult to detect through force sensing alone. The integration of force and vibration sensing is therefore central to the MAGICIAN tactile perception strategy. Force measurements provide information about stable contact, normal load, and macroscopic interaction conditions, while acceleration measurements provide sensitivity to fast local events and micro-scale changes in surface profile. The resulting multimodal signal forms the basis for defect detection and classification.



(a) (b)

Figure 39: Overview of the Tactile Perception Module. (a) Exploded view and (b) assembled prototype of the tactile module. The system integrates a six-axis force/torque sensor (A), a tri-axial accelerometer (B), and a compliant TPU-based comb-like contact interface (C) designed to transmit surface-induced vibrations during sliding exploration.

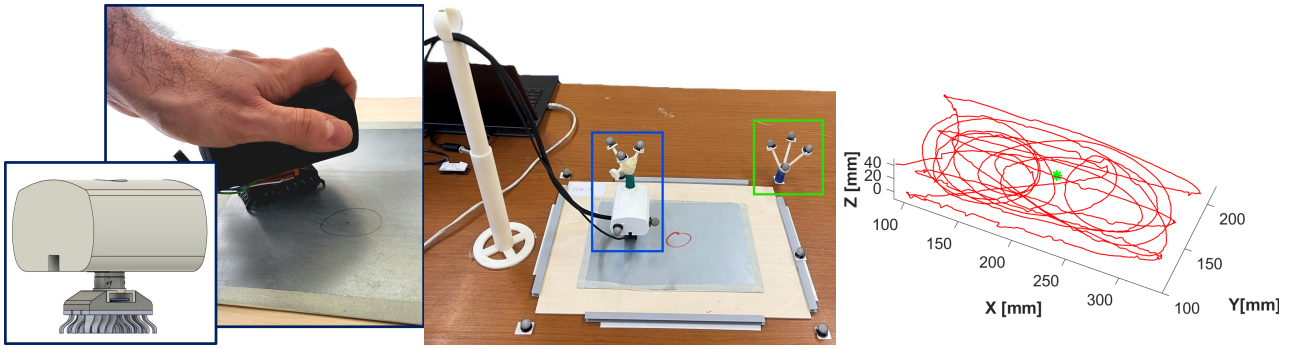
3.5 DATA ACQUISITION AND ANNOTATION

The data acquisition pipeline adopted in D3.3 builds on the framework described in D3.2. In that previous phase, IIT developed and validated a Vicon-based handheld acquisition setup in which human users explored TOFAS-annotated car body frames with the tactile probe, while force, acceleration, and 6-DoF motion data were synchronously recorded. The pipeline included the definition of a common panel reference frame, the calibration of defect locations, the temporal alignment of tactile and motion data, and the automatic annotation of tactile signals based on the spatial overlap between the probe and the known defect regions. The human-guided acquisition setup used as a basis for the robotic campaign is summarized in Figure 40.

Human-guided acquisition

The human-guided acquisition stage represents the link between manual tactile inspection and robotic inspection. As detailed in D3.2 and summarized in Figure 40, human users scanned metallic panels using a handheld tactile probe while a Vicon motion capture system recorded the device's six-degree-of-freedom pose, along with force and acceleration signals. In the robotic validation campaign considered in D3.3, these data were used to extract motion priors and kinematic templates for robotic reproduction.

The human-guided campaign involved 8 subjects, each performing 20 trials, for a total of 160 human exploration trials. Each trial lasted approximately 20 seconds. These demonstrations provide two types of information: the tactile response generated by manual exploration and the kinematic templates used to define user-derived robotic trajectories. They also provide useful priors on scanning velocity and normal contact force, which are used to define the robotic exploration conditions shown in Figure 41.



(a) (b) (c)

Figure 40: Human-Guided Trajectories Acquisition framework for manual exploration. (a-b) The operator scans the metallic surface using a handheld probe integrating the Tactile Perception Module (highlighted in blue). (b) The spatial kinematics are recorded using a Vicon tracking system and passive retroreflective markers (highlighted in green). (c) The acquired data includes the real-time position and orientation of the device as it traverses the panel, providing the kinematic templates for subsequent robotic replication.

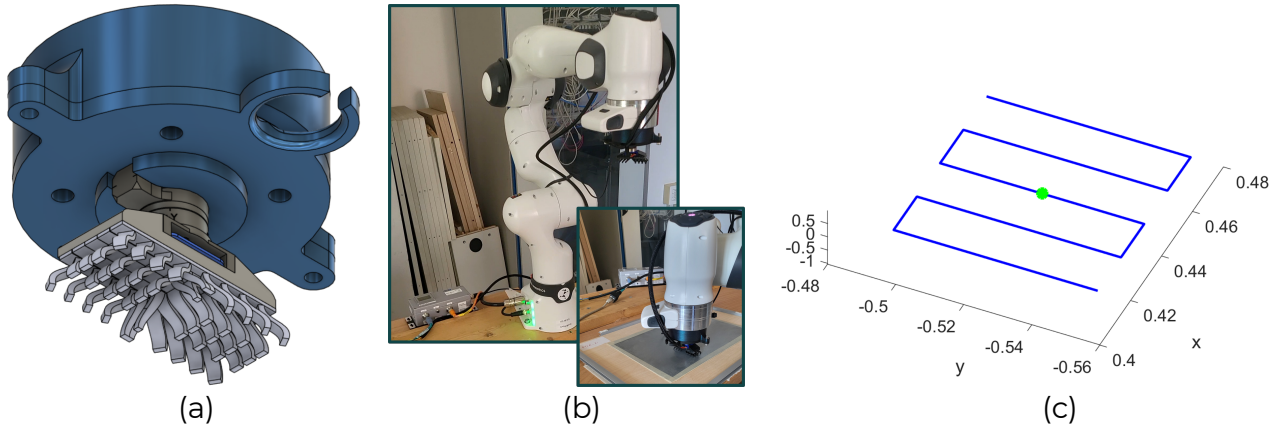
Robotic acquisition

The tactile module was integrated as the end-effector of a Franka Research 3 robotic arm, as shown in Figure 41. This robotic setup enables controlled and repeatable tactile data acquisition while systematically varying the main operational parameters that influence tactile perception.

Three exploration modalities were considered:

- User Trajectory (UT): robotic reproduction of the human-derived position trajectory, while keeping the end-effector orientation fixed to maintain the tactile probe perpendicular to the surface.
- User Orientation Trajectory (UOT): robotic reproduction of the full human-demonstrated pose, including both position and orientation.
- Linear Trajectory (LT): standardized serpentine scanning trajectory analytically defined around the defect location and used as a controlled robotic benchmark.

For each exploration modality, the robot executed tactile scans under different contact force and scanning speed conditions. Contact force was tested at 2 N, 4 N, and 8 N. Scanning speed was modulated with respect to the nominal velocity observed during human trials using three scaling factors: 0.5x, 1.0x, and 1.1x. The full factorial combination of 3 exploration modalities, 3 force levels, and 3 speed levels resulted in 27 operating conditions. Overall, the robotic campaign produced 4320 robotic acquisitions. Figure 41 summarizes both the robotic integration of the TPM and the standardized linear scanning pattern used as a repeatable reference trajectory. Figure 42 and Figure 43 shows data acquisition performed with the integrated platform during the 6th Integration Week in ALT.



(a) (b) (c)

Figure 41: Robotic data acquisition setup for the MAGICIAN tactile perception module. (a) Tactile Perception Module integrated into a custom-designed flange for robotic mounting. (b) TPM mounted as the end-effector of a Franka Research 3 robot during controlled surface exploration. (c) Example of the standardized linear scanning trajectory executed around the nominal defect location, used as a repeatable benchmark for robotic tactile inspection.

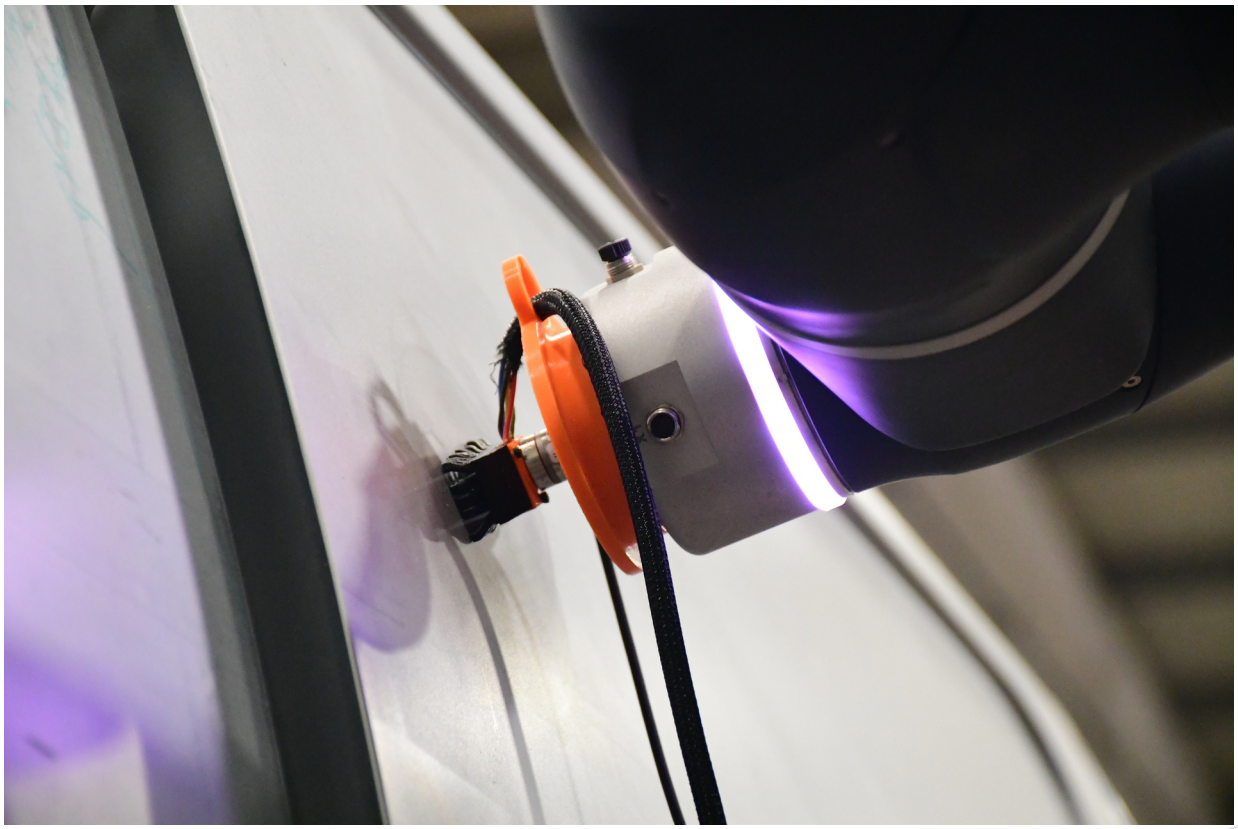


Figure 42: Robotic data acquisition at 6th Integration Week in ALT

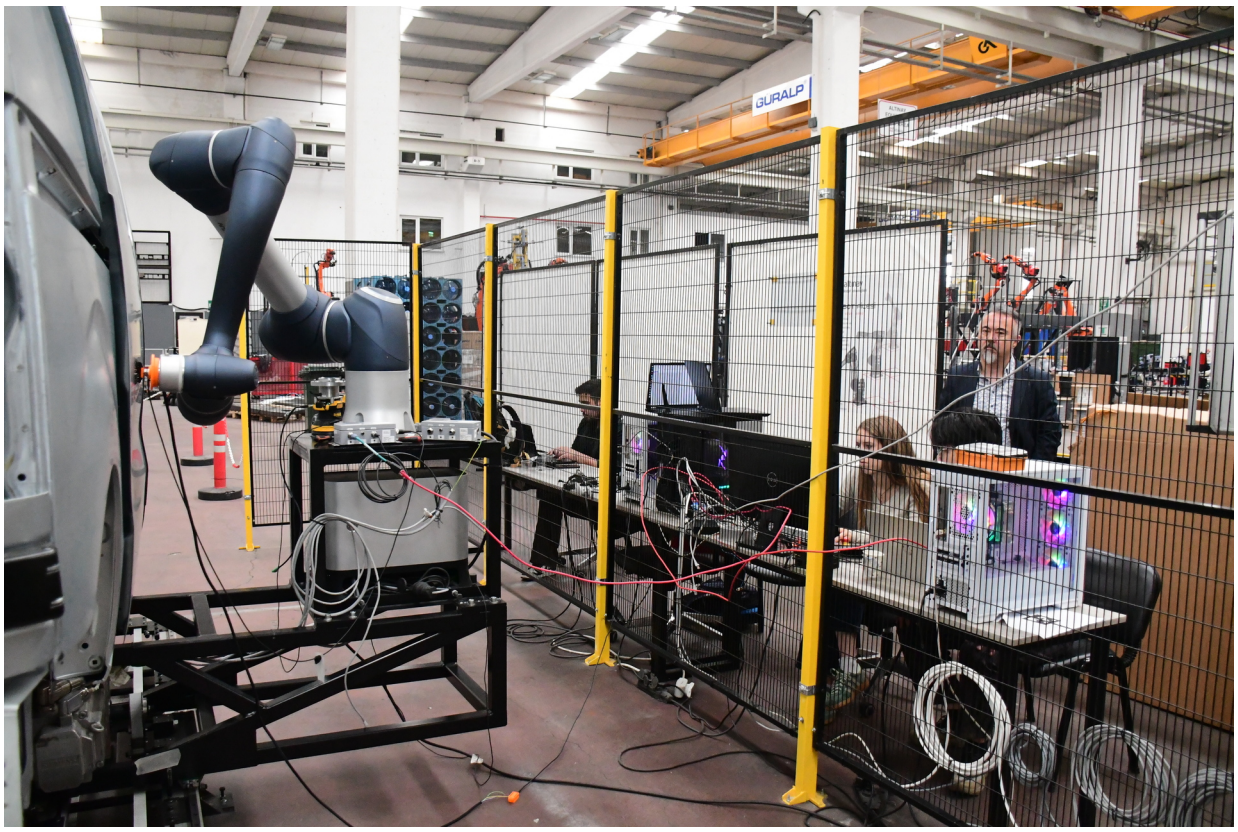


Figure 43: Setup for tactile data acquisition in ALT

Sensor acquisition and synchronization

During robotic exploration, the MAGICIAN tactile system records multimodal time series composed of contact forces and linear accelerations. The ATI Nano17 force/torque sensor is streamed through a Net/FT interface using the Real-Time Data protocol at 7 kHz. The ADXL335 accelerometer is sampled at 4 kHz using a custom data acquisition system based on a Teensy 4.0 microcontroller, with data transmitted to the host computer through high-speed USB serial communication.

The force and acceleration streams are acquired through different hardware pipelines and are therefore synchronized in software. A custom C++ layer performs temporal alignment between the force, acceleration, and robot-pose streams. This ensures that each tactile sample can be associated with the corresponding robot pose and with the instantaneous interaction between the tactile module and the inspected surface.

Automatic annotation

The general concept of spatial annotation was introduced in D3.2, where Vicon trajectories and calibrated defect locations were used to label handheld tactile data. In the updated robotic campaign, the same principle is retained, but the annotation

procedure is implemented directly with respect to the robot reference frame. The nominal location of each defect is defined during a preliminary calibration phase, in which the robot end-effector is manually guided to the defect location and the corresponding Cartesian pose is stored as the ground-truth defect position. Each defect is represented as a localized circular region on the panel surface, with a radius selected according to the characteristic scale of the defect. The tactile probe contact footprint is approximated as a rectangle whose centroid and in-plane orientation are determined from the robot end-effector pose. At each time step, the four projected vertices of the probe footprint are evaluated with respect to the circular defect region. A sample is labelled as defect-related when at least one vertex of the projected footprint falls within the defect region. This procedure produces labelled multimodal time series including robot pose, force, acceleration, and binary defect labels. The resulting dataset provides synchronized and spatially annotated tactile data suitable for supervised classification. Compared with the handheld-only dataset described in D3.2, this updated dataset allows the MAGICIAN tactile perception module to be evaluated under controlled and repeatable robotic inspection conditions.

3.6 METHODOLOGIES EMPLOYED

The robotic tactile inspection pipeline developed in this phase includes trajectory generation, force-regulated execution, multimodal signal processing, feature extraction, classification, and validation across different generalization regimes.

Exploration strategies

The robotic exploration strategies are derived from two complementary sources: human demonstrations and analytical trajectory generation. For user-derived trajectories, the position and orientation of the tactile probe recorded during human-guided acquisition (Figure 40) are encoded using Cartesian pose Dynamic Movement Primitives (DMPs). Separate translational and rotational components are used to reproduce the demonstrated exploratory motion. This formulation enables the generation of smooth robotic references from human demonstrations and the application of temporal scaling to test different scanning speeds.

The standardized Linear Trajectory (LT), illustrated in Figure 41c, is defined as a serpentine scanning path centred on the defect location. The trajectory is designed to cover the target region around the defect using repeated back-and-forth passes. This provides a controlled benchmark that reduces the variability associated with human demonstrations and enables a direct comparison between human-derived and standardized robotic exploration strategies.

Execution controller

All exploration modalities are executed using the same low-level control architecture. The robot tracks the desired Cartesian trajectory using an admittance controller driven by the wrench measured at the tactile module. The direction normal to the panel is

assigned to force control, while the tangential directions are used for trajectory tracking. This allows the robot to maintain approximately constant contact force during sliding exploration while remaining compliant to local surface variations.

The commanded Cartesian twist is mapped to joint velocities through damped least-squares inverse kinematics with null-space regularization. This ensures smooth and stable execution during contact-rich exploration.

Multimodal learning pipeline

The learning pipeline transforms the synchronized tactile time series into a structured dataset for defect classification. The force and acceleration signals are first resampled to a common frequency of 4 kHz. The three force channels and three acceleration channels are then concatenated into a six-dimensional multimodal signal. In addition, a force magnitude and an acceleration magnitude are computed from the corresponding axes to provide two derived scalar features. These magnitude describe the overall force and acceleration independent of direction. The magnitude is calculated as follows:

$$AccMag = \sqrt{\left(A_x^2 + A_y^2 + A_z^2\right)}$$

$$ForceMag = \sqrt{\left(F_x^2 + F_y^2 + F_z^2\right)}$$

The synchronized signal is segmented into fixed-length, non-overlapping windows of 0.1 s. Each window represents a short local interaction between the tactile probe and the surface. The label of each window is assigned based on the fraction of defect-labelled samples within that window. In the multi-class classification setting, each window is assigned to one of three classes: no defect, positive dent, and negative dent. For the binary classification setting, positive and negative dents are both flagged as defect.

Outlier removal is applied independently to each force and acceleration channel by clipping values outside the selected percentile bounds and interpolating the affected portions of the signal.

Each window is then mapped to a feature vector by computing statistical and temporal descriptors for each channel. These include first- and second-order statistics, distribution descriptors, amplitude-related features, temporal dynamics, signal-complexity measures, and zero-crossing rate. The resulting feature vectors provide a compact representation of contact physics and surface-induced tactile signatures.

Classification models

Three ensemble classifiers were evaluated:

- Random Forest (RFC),
- Gradient Boosting (GBC),
- Bagging (BC).

These models were selected because surface defects generate nonlinear correlations between force and vibration signals and because ensemble methods are generally

robust in settings with limited physical samples. Two dummy baselines were also included: a Majority Classifier (MC) and a Random Classifier (RC). These baselines provide lower-bound references for evaluating whether the tactile features contain discriminative information.

Validation protocols

Four validation strategies were used. First, a Stratified Group K-Fold protocol with five folds was adopted to evaluate whether the robot-acquired tactile representation contains informative signals for defect classification while preventing data leakage across groups.

Second, a Leave-One-Group-Out protocol based on the demonstrated scanning style was used. This validation regime, referred to as LOGO-ScanStyle, evaluates whether the tactile representation generalizes to previously unseen demonstrated scanning styles. Third, a Leave-One-Group-Out protocol based on the physical panel was used. This validation regime, referred to as LOGO-Panel, evaluates whether the classifier generalizes to previously unseen surfaces and defect geometries.

Fourth, the dataset was split into training, validation, and test sets. The models were trained on the training set, the best model was selected using the validation set, and the test set was used to confirm that the final model generalizes as expected. This method requires more data then the other methods but is less computationally expensive and provides a straightforward evaluation workflow.

Different metrics were adopted depending on the validation regime. F1-score was used for LOGO-ScanStyle, since each held-out demonstrator group includes all classes. Accuracy was used for LOGO-Panel, because each physical panel contains the no-defect class and only one defect type.

Finally, Linear Mixed Models were used to quantify the effect of the operating parameters. Exploration Modality, Contact Force, and Scanning Speed were treated as fixed factors, while Demonstrator or Panel was used as random intercept depending on the validation protocol. This analysis was used to identify the main drivers of performance and to select a robust operating condition for MAGICIAN tactile inspection.

3.7 RESULTS AND FINDINGS

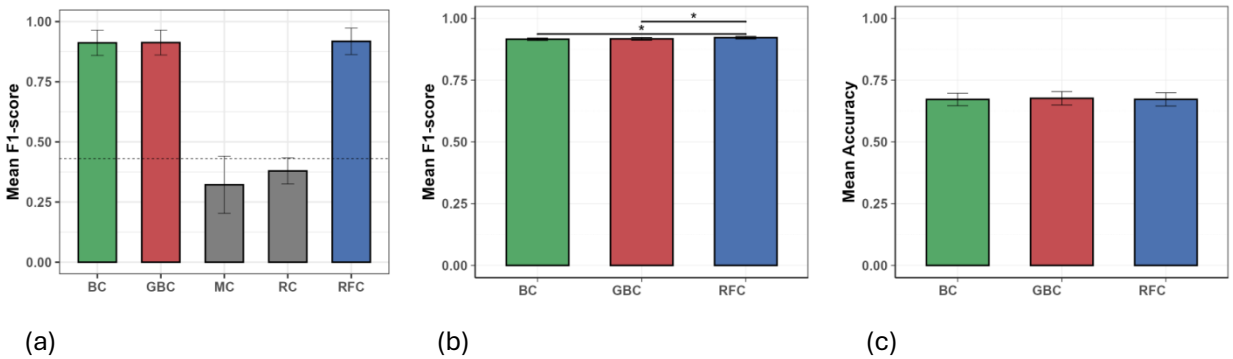
The updated results show a significant advancement with respect to the preliminary handheld-based evaluation reported in D3.2. The new analysis is based on robot-acquired tactile data and provides a systematic evaluation of the influence of exploration modality, contact force, and scanning speed on defect classification.

Tactile perception module validity

The first objective of the analysis was to verify whether the proposed tactile module produces signals that are informative for defect classification. Under the Stratified Group K-Fold protocol, the three ensemble classifiers achieved highly consistent performance.

Bagging obtained a mean accuracy and F1-score of 0.912, Gradient Boosting obtained 0.913, and Random Forest obtained 0.918. In contrast, the Majority Classifier achieved a mean accuracy of 0.484 and F1-score of 0.322, while the Random Classifier achieved a mean accuracy of 0.377 and F1-score of 0.379.

As summarized in Figure 44, these results show that the multimodal tactile signals acquired by the TPM contain discriminative information for classifying surface defects. Compared with the average baseline performance, the ensemble classifiers achieved gains of approximately +0.48 in accuracy and +0.56 in F1-score.



(a) (b) (c)

Figure 44: Performance evaluation and generalization analysis of the tactile classification models. (a) Comparison between ensemble classifiers and baseline models under the Stratified Group K-Fold protocol, showing that Bagging (BC), Gradient Boosting (GBC), and Random Forest (RFC) clearly outperform the Majority Classifier (MC) and Random Classifier (RC). (b) Comparison between ensemble classifiers under the LOGO-ScanStyle validation regime, assessing generalization to previously unseen demonstrated scanning styles. (c) Comparison between ensemble classifiers under the LOGO-Panel validation regime, assessing generalization to previously unseen physical panels.

The comparison among ensemble classifiers was also repeated under the two Leave-One-Group-Out regimes, as shown in Figure 44b-c. Under LOGO-ScanStyle validation, the classifier architecture had a statistically significant effect on F1-score, with Random Forest achieving the highest average performance. However, the average performance of the three ensemble classifiers remained tightly clustered around 0.91. Under LOGO-Panel validation, the classifier comparison was not statistically significant at the standard 0.05 level, suggesting that when generalizing to unseen physical panels, the limiting factor is more closely related to tactile feature quality and panel variability than to the specific classifier architecture.

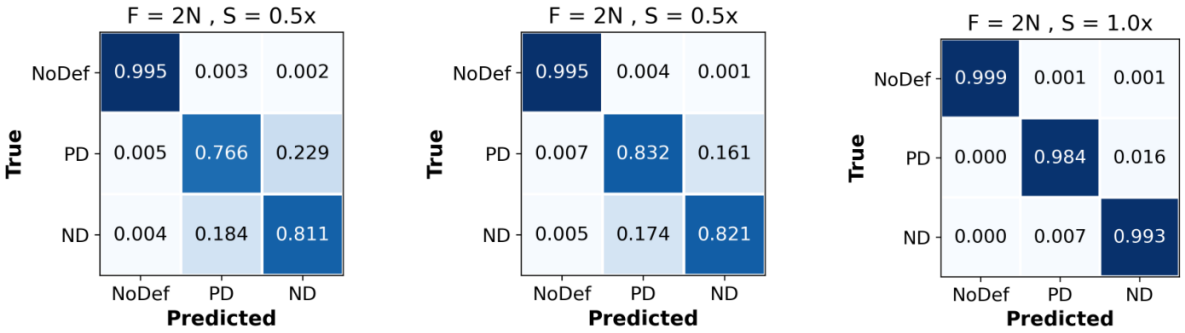
For this reason, the subsequent analysis adopted a model-agnostic perspective, averaging the results across the three ensemble classifiers and focusing on the physical parameters of the exploration process.

Performance across exploration modalities

The results show that all three exploration modalities are effective in detecting the presence of a defect. However, the Linear Trajectory provides stronger discrimination between positive and negative dents. This suggests that a standardized and repeatable motion pattern can generate more consistent tactile signatures for defect-type

classification than user-derived trajectories.

Figure 45 reports the best-performing confusion matrices for the three exploration modalities. All modalities achieve high recall for the no-defect class and good defect detection performance. Nevertheless, user-derived trajectories exhibit more confusion between positive and negative dents, while the standardized linear trajectory provides the clearest separation between the two defect classes.



(a) (b) (c)

Figure 45: Best-performing confusion matrices for each robotic exploration modality. The matrices are obtained under the Stratified Group K-Fold protocol and normalized by row to show class-wise recall for (a) User Trajectories with Orientation (UOT), (b) User Trajectories with fixed orientation (UT), and (c) Linear Trajectories (LT). All modalities show high performance in separating defect and no-defect samples, while the standardized Linear Trajectory provides the strongest discrimination between positive dents (PD) and negative dents (ND).

The descriptive statistics highlight a clear difference between generalization across scanning styles and generalization across physical panels. Under LOGO-ScanStyle, the aggregated F1-score ranged from 0.793 to 0.992, with a median of 0.919 and a mean of 0.918. This indicates that the tactile representation generalizes well across previously unseen demonstrated scanning styles.

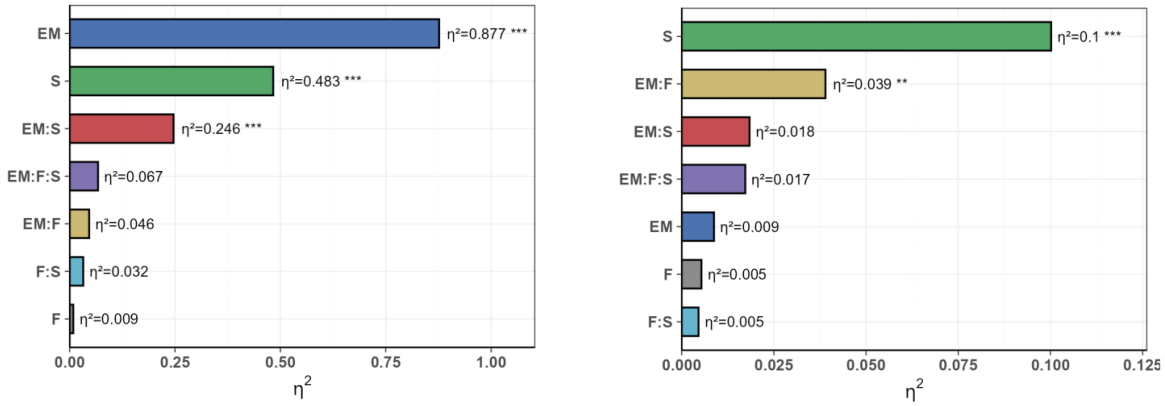
Under LOGO-Panel, the aggregated accuracy ranged from 0.218 to 0.995, with a median of 0.681 and a mean of 0.674. This wider variability indicates that generalization across previously unseen physical panels is more challenging and is strongly affected by panel-specific morphology and defect geometry.

Drivers of performance

The inferential analysis identified different performance drivers for the two validation regimes, as summarized in Figure 46.

Under LOGO-ScanStyle, the fixed operating parameters explained most of the observed variability. The marginal coefficient of determination was 0.815, while the conditional coefficient of determination was 0.887. Exploration Modality emerged as the dominant factor, followed by Scanning Speed. The interaction between Exploration Modality and Scanning Speed was also significant, indicating that the effect of scanning speed depends on the adopted exploration strategy. Contact Force did not show a significant

main effect in this regime. The Linear Trajectory remained stable across speed levels, whereas the user-derived modalities were more sensitive to changes in scanning speed. In particular, reducing the speed to 0.5x improved the performance of the user-derived trajectories, while increasing speed introduced additional variability. This result suggests that standardized motion helps decouple defect-related tactile cues from operator-dependent kinematic signatures.



(a) LOGO-ScanStyle (F1): Effect sizes (η^2)

(b) LOGO-Panel (Accuracy): Effect sizes (η^2)

Figure 46: Statistical analysis of the factors affecting classification performance. (a) Effect sizes under LOGO-ScanStyle validation, where exploration modality and scanning speed dominate performance. (b) Effect sizes under LOGO-Panel validation, where scanning speed becomes the main factor and the interaction between exploration modality and contact force remains relevant.

Under LOGO-Panel, the fixed operating parameters explained a smaller portion of the total variability. The marginal coefficient of determination was 0.119, while the conditional coefficient of determination was 0.365. This confirms that panel-dependent effects represent a major source of variability when the system is evaluated on previously unseen physical surfaces. In this regime, Scanning Speed emerged as the dominant factor. The best performance was obtained at the slowest speed, 0.5x, with an estimated marginal accuracy of 0.747. Performance decreased to 0.648 at 1.0x and to 0.625 at 1.1x. This indicates that slower scanning improves the extraction of discriminative tactile features, most likely because it increases the temporal resolution of contact with subtle surface irregularities.

The interaction between Exploration Modality and Contact Force was also significant. At 2 N, the user-derived modalities outperformed the Linear Trajectory. However, at 4 N and 8 N, these differences disappeared, and the Linear Trajectory reached performance comparable to the best user-derived configurations. This suggests that the standardized linear motion benefits from sufficient mechanical coupling between the tactile probe and the surface. Increasing force from 2 N to 4 N improved LT performance, while increasing force further to 8 N did not provide a clear practical advantage.

Robust operating condition

A final analysis was conducted to identify a single operating condition that remains

reliable across both validation regimes. Each combination of Exploration Modality, Contact Force, and Scanning Speed was treated as a complete operating condition and compared against the best-performing configuration.

As shown in Figure 47, this analysis identified LT x 4 N x 0.5x as the most robust operating condition. This configuration achieved an estimated F1-score of 0.9868 under LOGO-ScanStyle and an estimated accuracy of 0.7735 under LOGO-Panel. The absolute best condition under LOGO-Panel was LT x 8 N x 0.5x, with an estimated accuracy of 0.8066. However, the difference between LT x 8 N x 0.5x and the selected LT x 4 N x 0.5x condition was not statistically significant. Therefore, the additional force did not provide a compelling advantage. Conversely, replacing the selected standardized linear trajectory with user-derived trajectories produced significant performance drops under LOGO-ScanStyle.

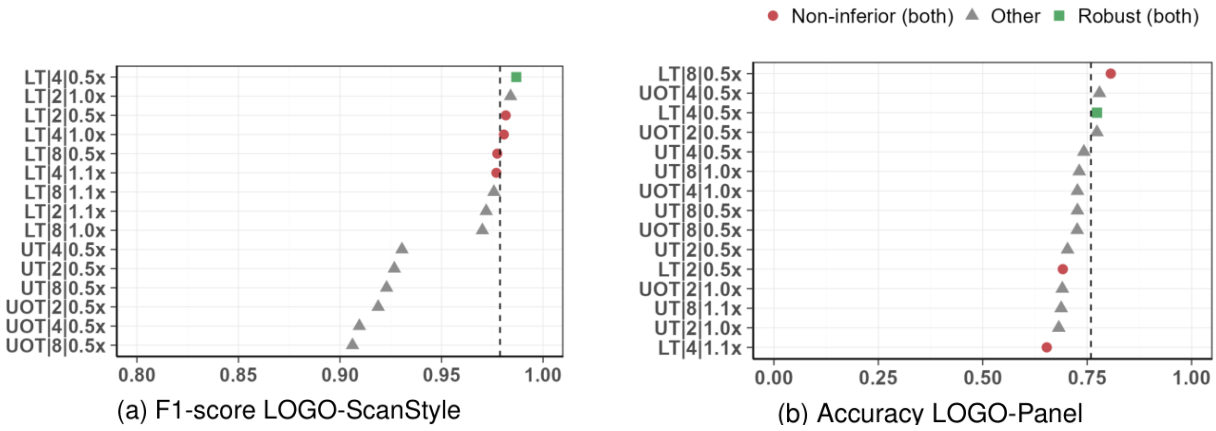


Figure 47: Identification of the most robust operating condition across the two validation regimes. (a) Ranking of the operating configurations under LOGO-ScanStyle validation, using F1-score as performance metric. (b) Ranking of the operating configurations under LOGO-Panel validation, using accuracy as performance metric. The green markers indicate configurations that are robust in both validation regimes, while red markers indicate configurations that are statistically non-inferior in both regimes.

The selected condition should therefore not be interpreted as universally optimal, but as the most robust compromise across the considered validation regimes. A contact force of 4 N provides sufficient mechanical coupling to capture relevant defect morphology while avoiding unnecessary interaction variability. A scanning speed of 0.5x improves the stability and discriminative quality of tactile signatures. The Linear Trajectory provides repeatable motion and reduces sensitivity to user-dependent exploration patterns.

Overall, the updated results demonstrate that the tactile perception module is suitable for robotic defect detection on metallic surfaces. The MAGICIAN system achieves strong classification performance under controlled validation, robust generalization across unseen scanning styles, and meaningful performance on unseen physical panels. At the same time, the results highlight that cross-panel generalization remains the main challenge for future work.

3.8 CHALLENGES AND LIMITATIONS

The robotic validation campaign presented in D3.3 addressed several limitations of the previous handheld-only evaluation reported in D3.2. In particular, IIT validated the Tactile Perception Module on a robotic platform, controlled the main acquisition conditions, and quantitatively assessed the influence of exploration modality, contact force, and scanning speed. The availability of robot-acquired tactile data substantially improved the maturity of the classification results and enabled a more realistic evaluation of the MAGICIAN tactile perception pipeline.

The main remaining challenge concerns generalization across previously unseen physical panels. While the system showed robust generalization across different demonstrated scanning styles, performance was more variable when the held-out group corresponded to a physical panel. This suggests that panel-specific morphology, local surface properties, and defect geometry can significantly affect the tactile signatures acquired during contact. To address this aspect, future work will expand the dataset with additional TOFAS-provided annotated components and investigate panel-specific adaptation strategies, such as in-situ baseline acquisition on defect-free regions before inspection.

A second aspect to be further developed concerns the current scope of the classification task. In this phase, the MAGICIAN tactile module focuses on defect-type recognition, distinguishing between no defect, positive dent, and negative dent. Although the dataset includes defects with different severity levels, severity estimation was not explicitly addressed in the present validation. Future developments will therefore extend the learning pipeline toward severity grading or regression-based estimation of defect magnitude, enabling a more detailed characterization of the detected imperfections.

The operating condition identified in D3.3 and summarized in Figure 47, namely the standardized linear trajectory with 4 N contact force and reduced scanning speed, provides a practical reference configuration for the next integration steps within the MAGICIAN sensing and robotic platform. Future activities will focus on consolidating the tactile perception pipeline across broader sets of components, defect morphologies, and acquisition conditions, while preserving the robustness observed in the robotic validation campaign.

Overall, D3.3 marks a significant advancement from the D3.2 handheld acquisition pipeline toward a robot-integrated tactile inspection framework. The next steps will focus on dataset expansion, improved cross-panel generalization, defect severity estimation, and consolidation of the tactile module within the MAGICIAN use-case scenarios.

4 HUMAN MOTION PERCEPTION

4.1 INTRODUCTION

The SR and CR arms share a workspace with skilled automotive technicians. Our collaborative robots thus need to continuously monitor co-worker location and intent, and to modulate or halt motion when a human enters a proximity zone.

For MAGICIAN these demands: accurate real-time person detection, 3D body-pose estimation and short-horizon motion prediction (50–500 ms) for anticipatory velocity scaling before a collision zone is entered. The perception capabilities are provided by *magician_body_pose_estimation* and *magician_general_visual_perception* repositories.

Learning reworking skills, the hand motions, force profiles, and tool orientations used by an expert panel-beater are captured from human demonstrations for encoding as generalisable motion primitives for the CR arm. This requires multi-modal, time-synchronised observations of: (a) tool 6-DoF pose; (b) full-body kinematics and (c) contact forces repository and the TOFAS capture campaign address these needs.

4.2 OBJECTIVES AND REQUIREMENTS

Collaborative robotics is often seen as the cornerstone of next-generation manufacturing solutions, commonly referred to as "Industry 5.0." However, distinguishing between robotic applications that qualify as "collaborative" and those that do not, is not always straightforward. Rather, it is easier and potentially more informative to identify a spectrum of possible scenarios. At one end of the spectrum, we find stand-alone, classical robotic stations, which are classified as “collaborative” to simplify the physical layout of the production line (collaborative robots do not need to be segregated from humans). At the opposite end, we find truly collaborative applications in which humans and robots engage in a direct and physical collaboration (e.g., a robot can hand over tools to humans or help them move heavy loads). In the MAGICIAN project, we find ourselves in an intermediate scenario. Robots and humans share the same workspace, but they do not directly collaborate. The stations where the sensing robots identify defects and the cleaning robots remove them are shared areas, where humans work alongside the robots to supervise their operation or handle particularly complex tasks that exceed the robots' capabilities. In this setting, our problem is to ensure a safe coexistence without sacrificing productivity. We can identify two situations which can lead to potential problems:

1. A mobile robot is moving along a possible collision course with a human,
2. A robotic manipulator is executing an activity following a path that can collide with some part of the body of the human operator and/or trap them.

Our approach to deal with these potential problems hinges on human-aware motion

planning. A general overview of the approach is illustrated in Figure 48. In this figure we assume the presence of a robot, which must execute a given set of tasks (such as scanning the surface of a car body in search of defects). Appropriate environment sensors detect the presence of humans in the scene and observe their motion. Based on this observation, a system produces a prediction of the human motion for a time horizon of 1.5 to 2 seconds. Based on this prediction and the knowledge of the task that the robot must execute, the human-aware motion planner decides a trajectory that minimises the risk of accidents while guaranteeing a satisfactory level of performance.

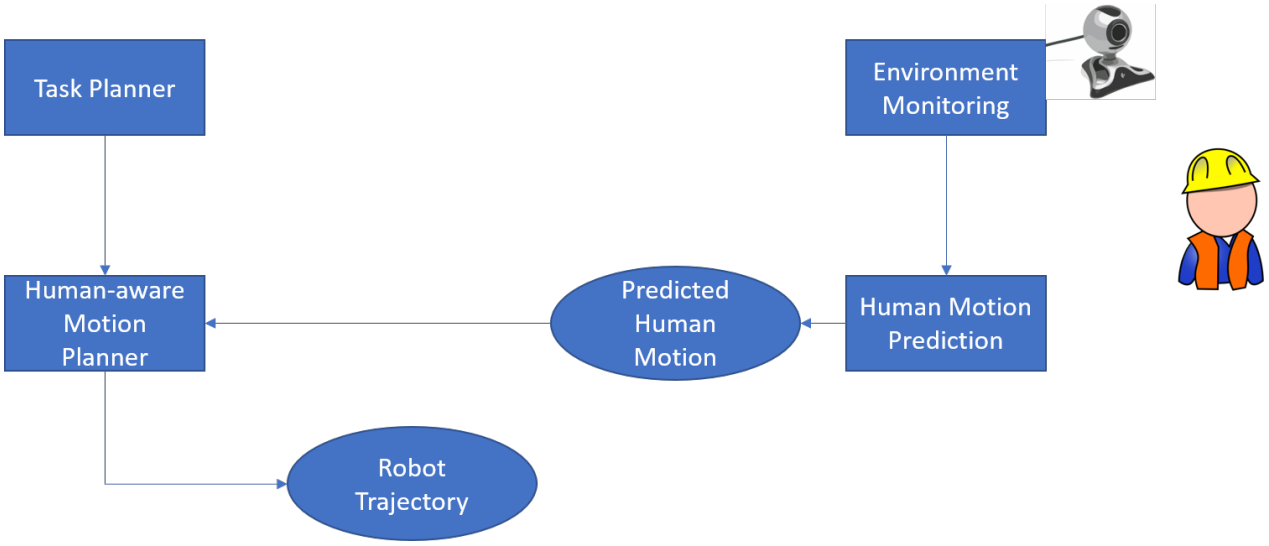


Figure 48: The framework of Human-Aware Motion Planning.

In this section we are focusing on ways to produce an acceptable prediction for the human motion. The problem takes on a different form depending on whether we are dealing with a mobile robot or a manipulator. In the first case, we need a prediction on the position of the human body as a whole (for instance the centroid of the point cloud associated with the human). This topic was explored in previous projects, and we will build upon the results obtained at that time [Ant23]. For the second case, the robot and the human need to work at a close distance. Therefore, we need to consider the position of the different parts of the body since these fine details can prove useful. For instance, for a human with open and stretched arms, it is possible for the robot to use the space between the two arms. This motion prediction will be one of the outcomes of the project and it must meet the following requirements:

1. Accuracy: the acceptable margin of error is in the order of a few centimetres.
2. Time horizon: to be useful for motion planning, the prediction must be reliable for a time horizon of approximately 2 seconds.
3. Efficiency: the system must demonstrate sufficient reactivity, meaning the prediction should be delivered within a few tenths of a second after new data is

collected.

4. Multi-scenario: in cases where uncertainty remains about the person's possible movements, the system can generate multiple scenarios, each associated with a probability level.

We can think of human motion perception as the interplay between two different conceptual modules. The first (elaborated in Section 4.3.1) deals with pattern recognition in the RGB level, with the task of correctly extracting the pose of observed humans regardless of their various optical appearance differences. The second module (Section 4.3.2 of D3.2) deals with pattern recognition in the pose coordinate level (regardless of their RGB appearance), trying to calculate, observe and predict patterns of motion in human joint coordinate trajectories. Both modules form a common mechanism and play an important role for a successful motion prediction framework. A pose estimation module without good accuracy cannot provide accurate data for reliable motion predictions. Similarly, an accurate but slow rate of pose estimation predictions will not provide enough time resolution for detailed understanding of motion, thus resulting in skewed and unrealistic, erratic motion predictions. Furthermore, even with very good pose estimation having a powerful motion prediction technique is essential for a system that can properly model and anticipate the complex and intricate human motions that can be encountered in an industrial environment like the one we target.

4.3 METHODOLOGIES EMPLOYED

To meet the demanding requirements of the human motion predictor outlined above, the MAGICIAN team has thoroughly evaluated the best state-of-the-art solutions that are assessed in Sections 4.3.1 and Section **Error! Reference source not found.**. For pose detection, we are developing a novel U-NET based architecture that regressed 2D pose, depth and normal vectors in real-time to facilitate the motion prediction task while also providing depth perception to the MAGICIAN Cobot. The 2D pose can also be augmented using a MocapNET that performs inverse kinematics regression to provide higher level data. For motion prediction, after careful assessment, we narrowed our options to two alternatives: applying deep learning approaches or using classic clustering techniques informed by our understanding of the specific process.

4.3.1 TECHNIQUES FOR POSE DETECTION

We will begin by describing the pose detection sub-problem, which in general consists of the task of receiving an RGB image featuring persons, identifying them and the joints of their skeleton and providing this data as high-level output for use by other modules. Tackling the task is very challenging since the appearance of humans in an image widely ranges when they are recorded as 2D projections of red, green and blue light intensities. The human body is very flexible with many configurations, human appearance is very varied, parts of the scene may be occluded by obstacles, cameras suffer from lens deformations, thermodynamic noise, motion blur, vibrations and other potential

artifacts, observations might have multiple explanations and differences in lighting given the low dynamic range of typical camera CCDs pose significant challenges that need to be systematically overcome. Deep learning approaches have recently managed to tackle the very high dimensional space of RGB images successfully performing pose estimation and the next sections will briefly describe the various involved techniques adopted for use in the context of the MAGICIAN project.

4.3.1.1 DATA ACQUISITION AND ANNOTATION

Data acquisition for human pose estimation presents significant challenges in the European Union. Collecting images of individuals in manufacturing environments requires careful compliance with GDPR and the EU AI Act; all footage collected at TOFAS premises was carried out with the written consent of the recorded individuals and with care not to focus on their face or other identifying characteristics, but instead overlaying a 3D rendered skeleton (Figure 51, Figure 52, Figure 55, Figure 57).

At the ALT integration site, an Intel RealSense camera (Figure 50) was deployed in the worker cell. Critically, only the RGB stream is used with the depth stream produced by the structured-light sensor being used as a backup to cross reference the Depth information from RGB so that the method remains applicable to sites equipped only with commodity RGB cameras. Complementary to the defect annotation tool described in Section 2.4.2, a human pose annotation counterpart was developed and deployed to label ground-truth body configurations in the collected footage. The full D-PoSE 3D pose estimation stack runs even on an a relatively resource constrained laptop based on the RTX 1000 Ada Laptop 6GB VRAM GPU (Figure 51), ensuring ample performance on the RTX 5090 32GB of the Pilot PC (Figure 59).

A dedicated recording campaign was conducted at the TOFAS factory to capture expert technician defect rework demonstrations under realistic production conditions. The goal was to build a multimodal learning-by-demonstration dataset that pairs full-body motion, tool trajectory, and measured applied force while performing defect repairs with the goal of repairing the defects sensed by the sensing robot using similar motion-primitives such as the real human experts. Four heterogeneous cameras were deployed around the TOFAS working cell: a GoPro Hero 13 (4K, overhead wide, height 3.0 m, HFOV 94°), a GoPro Hero 4 (2.7K, side wide, height 3.0 m, HFOV 94°), a Nikon D7000 (1080p30, side long-ish, height 1.8 m, HFOV 52°), and a Nikon D7500 (1080p60, side long-ish, height 1.8 m, HFOV 52°). Camera positions and fields of view were planned in advance to use a custom top-down FOV calculator, ensuring overlapping coverage of the full ~4.6 m × 1.9 m TOFAS van body footprint from multiple angles (Figure 56, Figure 57) and further fine-tuned after experiments in TOFAS to provide the best possible recording results.

Frame-accurate synchronisation across all four cameras was achieved via a QR timecode system. The same laptop running the *magician_grabber* software (Section 2) recording the ATI force sensor with microsecond precision timestamps emitted a full screen QR code also in sync with the laptop monitor refresh rate of 60 Hz, with each code encoding

the current Unix timestamp. All four cameras each of which has its own internal clock and recording capabilities pointed toward this screen so that the QR code appears in every video frame. After retrieving the video streams from the SD-Cards of the cameras post-hoc decoding of the QR payload aligns streams to within a single frame regardless of differences in frame rate or recording start time. Two ArUco fiducial markers where affixed to the scene: an 18 cm marker mounted on the car body for vehicle pose and conveyor-position tracking, and a 6 cm x 6 cm x 6 cm cube of markers attached to the repair tool, providing continuous 6DoF tool trajectory estimates throughout the rework sequence. The ATI six-axis force/torque sensor recorded Fx, Fy, Fz and Tx, Ty, Tz concurrently, with timestamps aligned to the camera timeline via the same QR timecode. Figure 49 shows an example ATI recording from one rework session:

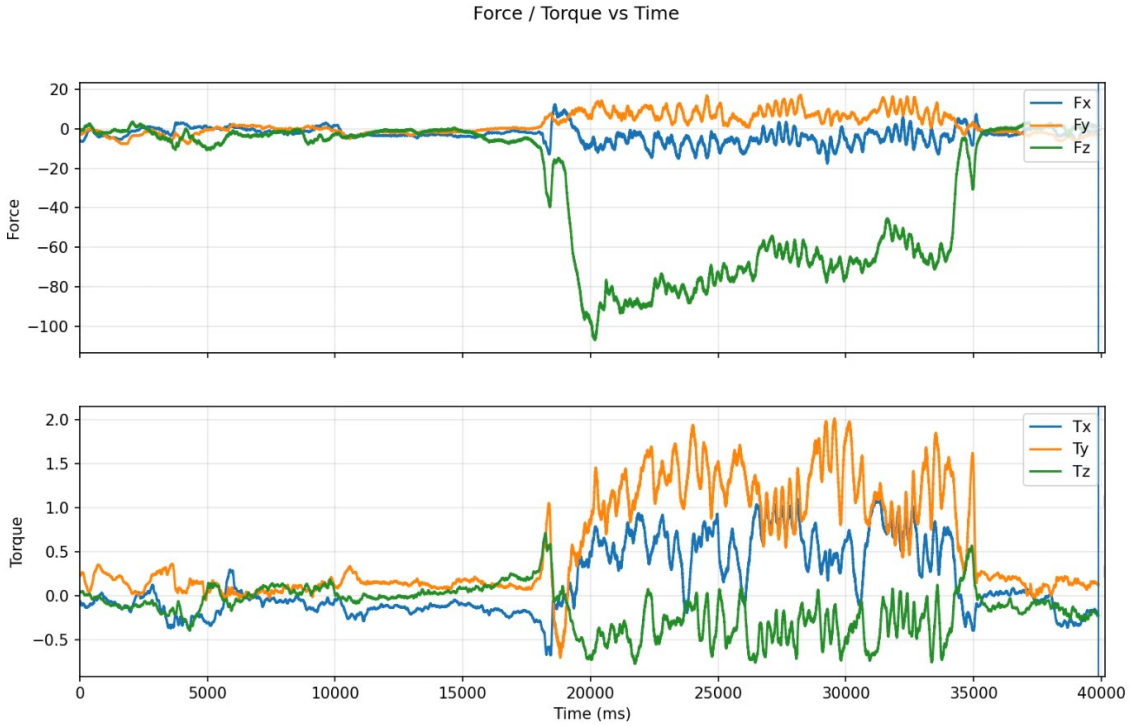


Figure 49: The forces recorded using the ATI Force/Torque sensor during the fourth demonstration of the first expert operator at TOFAS. The resting phase (0–15 s, near-zero forces) transitions sharply to the active rework phase as the technician engages the panel.

D-PoSE (4.3.1.2) was then applied offline to each of the four synchronized camera streams to extract per-frame 3D body pose sequences. Figure 55 shows representative tracking results across the camera views: up to four people are tracked simultaneously in the wide overhead GoPro Hero 13 stream, and the side-view Nikon cameras resolve fine-grained arm and hand configuration during tool contact. The resulting dataset provides, for each frame, a multi-view SMPL-X pose estimate, the 6DoF tool trajectory from the ArUco marker, and the simultaneous force/torque signal from the ATI sensor, a fully synchronised multimodal record of expert rework behaviour at TOFAS production

cadence.

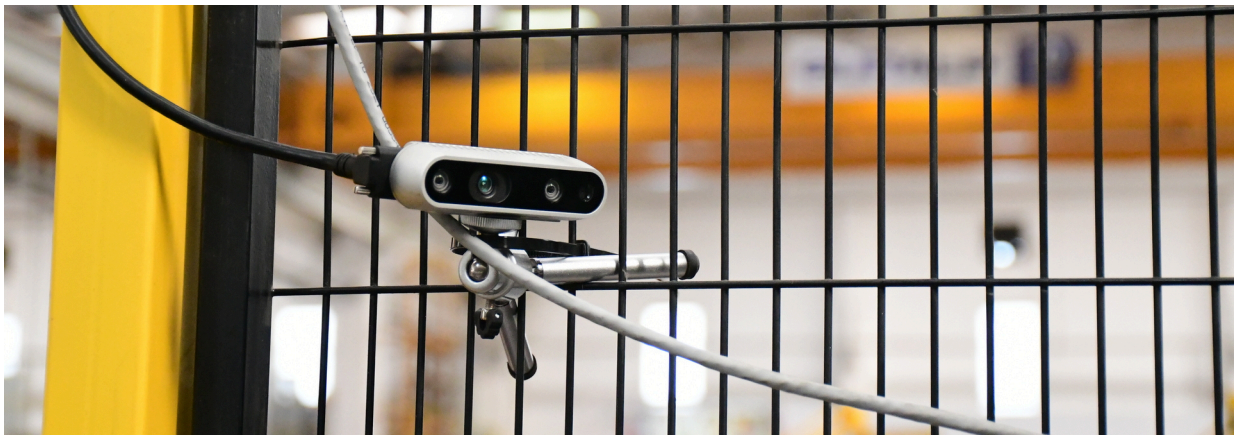


Figure 50: Deployed Intel Real-sense camera on the worker cell in ALT premises. It is worth noting that we use only the RGB stream of the camera, however the additional depth output that it features can be contrasted to the 3D skeleton derived and used as an additional safety check and guarantee for the robustness of the 3D pose estimation fidelity.

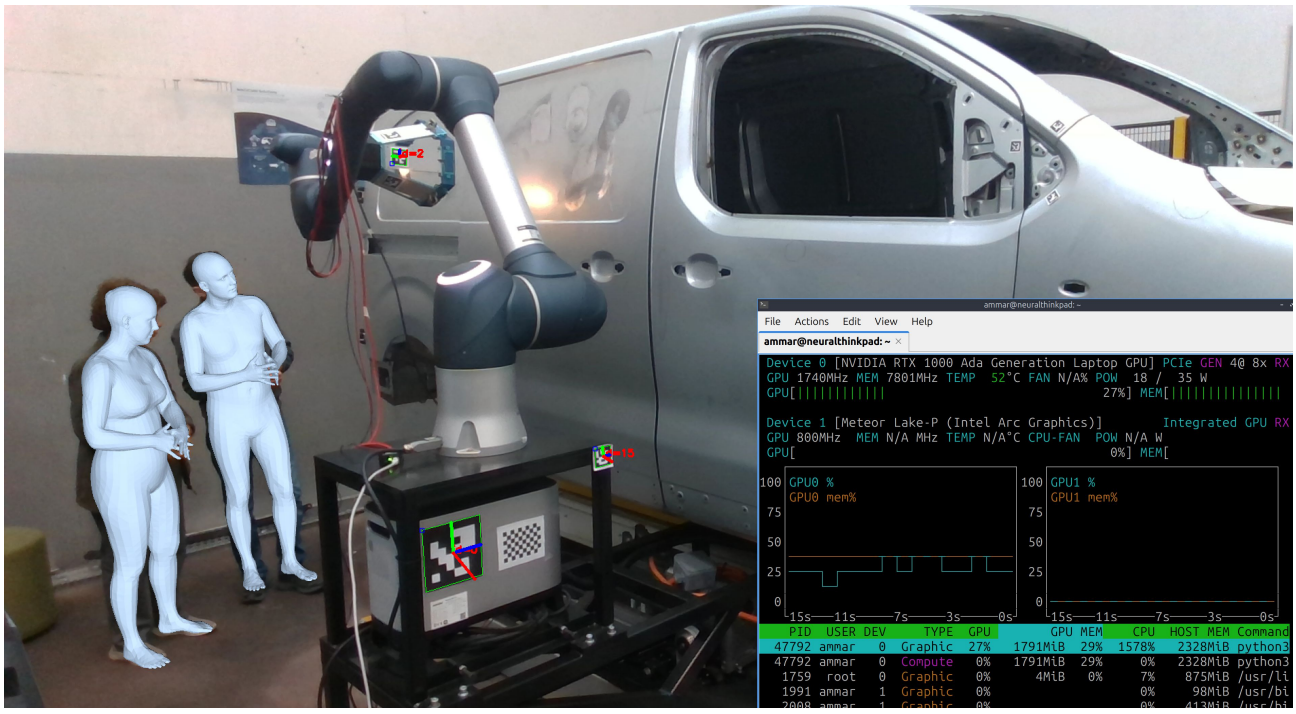


Figure 51: The 3D Pose estimation stack running on an RTX 1000 Ada Laptop GPU and tracking multiple persons while also performing ArUco marker localization and only utilizing a fraction of the available GPU resources

4.3.1.2 HUMAN POSE DETECTION METHOD

Real-time performance and high 3D accuracy are essential in a production-line setting. FORTH developed D-PoSE (Depth as an Intermediate Representation for 3D Human Pose and Shape Estimation), a one-stage method that estimates the full 3D body model from a single RGB image. D-PoSE was accepted at the IEEE International Symposium on

Virtual Reality (ISVR) 2026.



Figure 52: Stress testing D-PoSE in ALT premises when tackling the whole factory floor (left) and tracking a person interacting with the robot arm highlighting the accuracy of the method.

D-PoSE processes a single 224 × 224 RGB image through a HRNet-W48 convolutional backbone — chosen over Vision Transformer (ViT) backbones specifically to keep the parameter count and inference cost low enough for interactive robotic applications. The backbone generates spatially precise feature maps at four resolutions: $F_1 \in \mathbb{R}^{384\times7\times7}$, $F_2 \in \mathbb{R}^{192\times14\times14}$, $F_3 \in \mathbb{R}^{96\times28\times28}$, and $F_4 \in \mathbb{R}^{48\times56\times56}$. Two parallel decoders process these features through U-Net-like up sampling and refinement modules: a depth decoder producing a relative human depth map, and a part-segmentation decoder producing 23 SMPL-X body-part labels. A soft-attention mechanism combines the up sampled CNN features with the part-segmentation maps, concentrating the regressor's attention on body-relevant regions and suppressing background activations. The attention-weighted features are concatenated with the bounding-box coordinates (following the CLIFF supervision strategy) and the estimated depth map, then fed to an MLP regressor with 22 parallel heads that directly output the SMPL-X pose parameters θ, shape parameters β, and facial expression parameters ψ.

D-PoSE is trained entirely on synthetic data using two datasets: BEDLAM (subsampled at 6 fps, providing ground-truth SMPL-X parameters and depth maps) and AGORA (providing SMPL-X supervision and 3D human loss, but not depth). Training runs for 200K iterations with batch size 64 using the Adam optimiser (learning rate 1×10^{-5}, gradient clipping at 1.5) on a single NVIDIA RTX-A100 GPU in approximately three days.

	Training Datasets	Method	EMDB [20]			3DPW [47]		
			MVE	MPJPE	PA-MPJPE	MVE	MPJPE	PA-MPJPE
HRNet	SD	PARE	-	-	-	97.9	82.0	50.9
	SD	CLIFF	-	-	-	87.6	73.9	46.4
	BL	BEDLAM-HMR	-	-	-	93.1	79.0	46.4
	BL	BEDLAM-CLIFF	113.2	97.1	61.3	85.0	72.0	46.6
	BL	**D-PoSE (Ours)**	**99.0**	**85.5**	**53.2**	**81.4**	**68.9**	**43.6**
ViT	BL	HMR2.0	106.6	90.7	51.3	88.4	72.2	45.1
	BL	TokenHMR	106.2	89.6	**49.8**	85.7	71.6	44.0
HR Net	BL	**D-PoSE (Ours)**	**99.0**	**85.5**	53.2	**81.4**	**68.9**	**43.6**

Figure 53: Comparison of HPS errors on the EMDB and 3DPW datasets. SD denotes standard realistic datasets, while BL denotes training exclusively with synthetic datasets (BEDLAM and AGORA).

In Figure 53 we compare our method to the current state of the art methods. D-PoSE contains 81.2 M trainable parameters — 83.8% fewer than the current state-of-the-art TokenHMR (681 M, ViT backbone). It requires 64.70 GFLOPs per forward pass compared to 254.31 GFLOPs for TokenHMR and achieves 265 ms CPU inference time versus 441 ms for TokenHMR (Figure 54). This lightweight design makes D-PoSE practical for concurrent deployment alongside the visual defect classifier.

Method	GFLOPs	# Params	Infer. time
TokenHMR	254.31	681.0 M	441 ms
D-PoSE (ours)	**64.70**	**81.2 M**	**265 ms**

Figure 54: Comparing complexity of TokenHMR and D-PoSE with respect to GFLOPs, number of parameters and inference time (CPU).

In Figure 54 we present the floating-point operations, the number of model parameters and the inference time of D-PoSE. The current D-PoSE implementation operates on individual frames, recovering 3D pose and shape from a single RGB image without explicitly modelling temporal continuity. In parallel with the deployment of D-PoSE within MAGICIAN, FORTH has investigated a temporally consistent extension of the method for monocular video-based human pose and shape estimation [Shehaj26]. This work augments the depth-guided D-PoSE representation with recurrent and transformer-based temporal modules that process short sequences of consecutive frames. The temporal components are designed to exploit inter-frame motion cues while preserving the advantages of the original single-image spatial representation.

This direction is particularly relevant for MAGICIAN because the perception stack is expected not only to estimate accurate human pose in isolated frames, but also to provide stable motion trajectories for human-aware planning and learning from demonstration. Frame-wise jitter can propagate to downstream motion prediction, tool-use analysis, and motion-primitive extraction. The temporal D-PoSE extension addresses this limitation by adding velocity- and acceleration-based supervision, as well as an online shape-memory mechanism for reducing frame-to-frame fluctuations in

predicted body shape. Preliminary experiments on standard video-based benchmarks indicate that temporal modelling substantially reduces acceleration error while retaining competitive frame-wise reconstruction accuracy. These results support the use of temporally aware pose estimation as a future upgrade path for the MAGICIAN human motion perception module, especially in scenarios where smooth and reliable human trajectories are required for safe robot behaviour and faithful skill transfer.

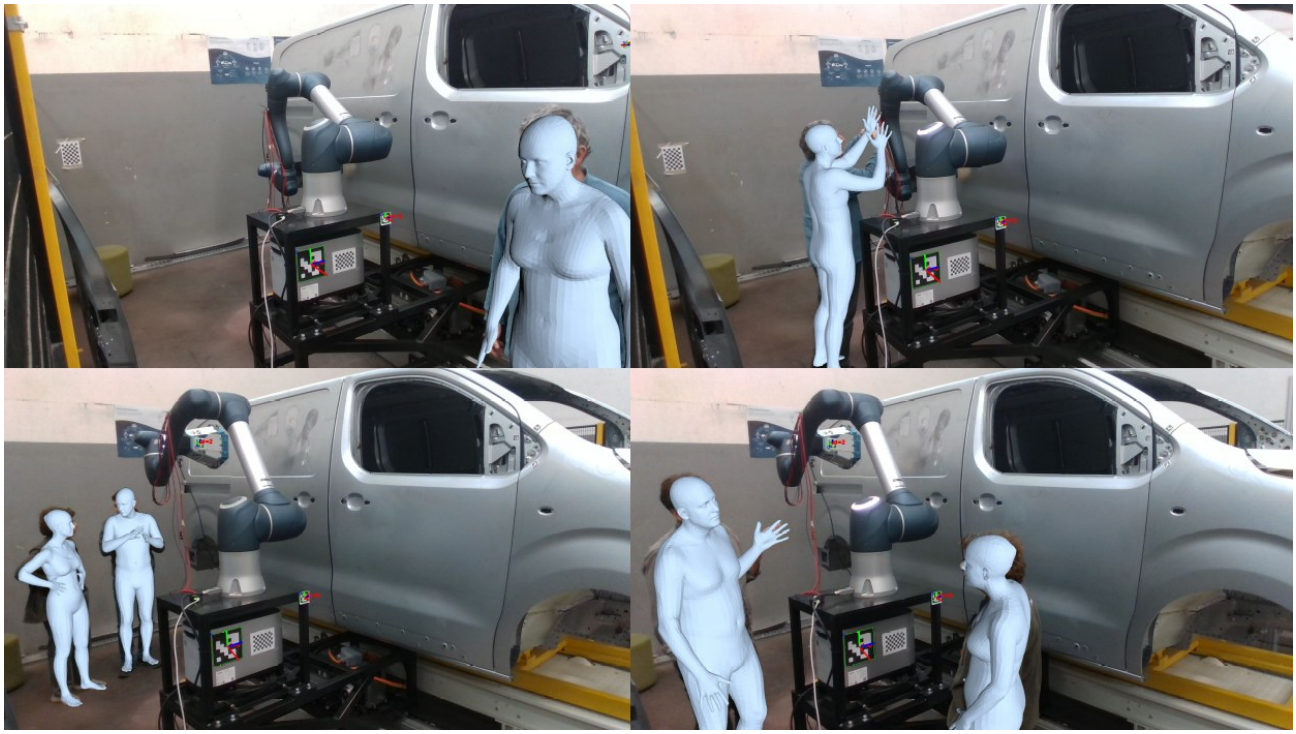


Figure 55: Interactions close to the robot tracked by the human pose estimation module of Magician during the 6th Integration Week in ALT.

Our qualitative results provide evidence on the effectiveness of our method across a diverse set of challenging scenarios. Table consolidated results from four key datasets, illustrating the versatility and robustness of our approach across a variety of environments and challenges. Figure 55 showcases our 3D estimation capabilities and high-fidelity output. It is worth noting that the same node that performs real-time 3D pose estimation also has the capacity to decode ArUco marker positions and is thus able to immediately resolve the distance between the robot base and robot endpoint to all visible persons in the scene. This using purely visual means with immediate calculation intended as a backup not even relying on the robot state. However, a very extensive motion prediction planning stack based on the robot trajectory as well as human motion profile is also present. The same approach has been used to record the car rework, where we used four different cameras with different specifications and lens configurations (Figure 56).

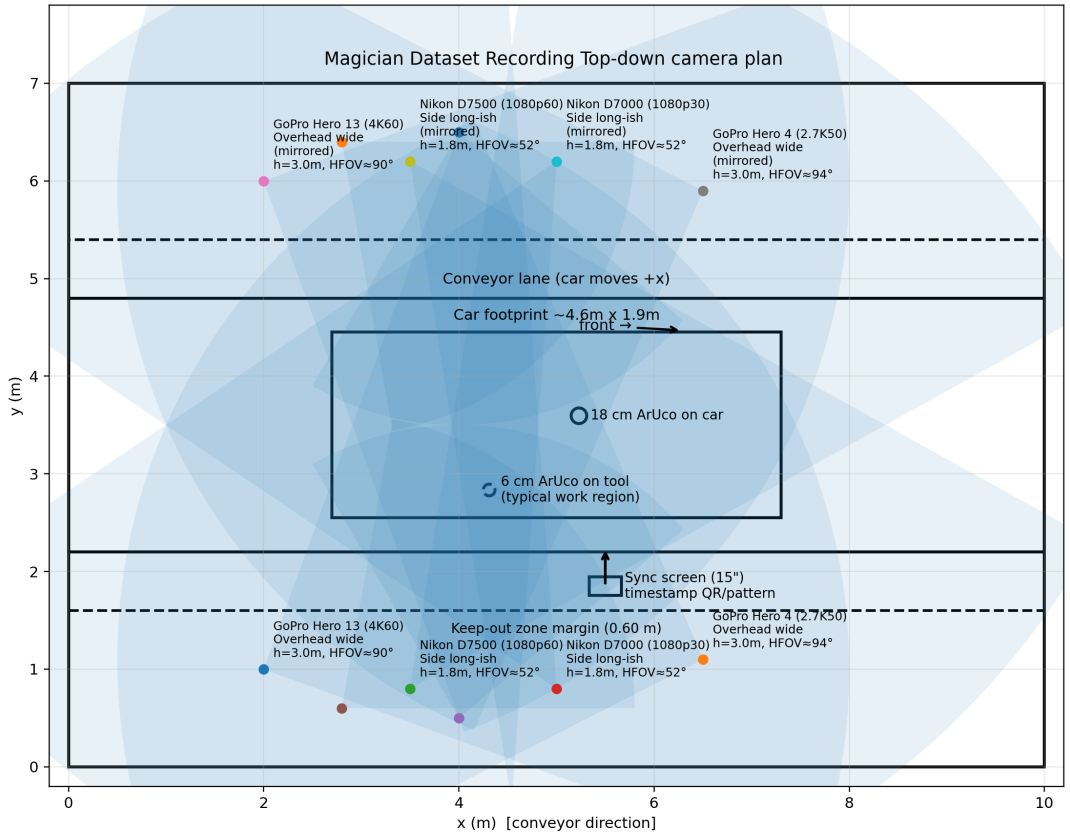


Figure 56: To record the car rework, we used four different cameras with different specifications and lens configurations. Two of them had zoom lens (NIKON based) to help focus on details and two wide-angle lenses (GoPro) to simultaneously capture a wider area. To facilitate accurately tracking the various ArUco markers we created a camera position calculator tool to make an informed decision about camera placement.

4.3.1.3 SENSING HUMANS AND THEIR CONTEXT

D-Pose provides state of the art accuracy and understanding of the 3D human form and pose using a low-cost RGB camera even in very challenging scenes with multiple visible humans. It is also excellent with respect to its computational efficiency allowing it to be executed in parallel with the defect detection node on a single GPU. By its design, however, and despite on the addition of ArUco marker decoding that simply and effectively reduces this gap it is not built to provide any information on the spatial context outside of the human domain. To address this aspect of the observed scene we developed a Y shaped Multi Attribute Prediction neural network that is called Y-MAP-Net: Learning from Foundation Models for Real-Time, Multi-Task Scene Perception [Qammaz2026]. This work was accepted and presented to a large audience as an oral presentation in the IEEE International Conference on Robotics and Automation (ICRA 2026) (Figure 63).

Y-MAP-Net adopts a Y-shaped topology with a single shared encoder branching into two dedicated decoder paths (Figure 60). The encoder progressively downscales input RGB

through convolutional blocks with Batch Normalisation and Leaky ReLU activations, arriving at a compact bridge representation. The pictorial decoder path produces 44 output heatmap/image channels: channels 1–17 encode COCO-17 2D joint heatmaps; channels 18–29 encode Part Affinity Fields (PAFs) for joint association; channel 30 encodes per-pixel relative depth; channels 31–33 encode surface normal in camera space; and channels 34–44 encode per-pixel semantic segmentation across 11 broad categories. The caption decoder path produces 8 GloVe token embeddings forming a multi-label scene description. Both paths share all encoder weights, so the total model (347 M parameters) uses only 48.94% of the combined weight of its four teacher models. Y-MAP-Net adopts a multi-teacher, single-student knowledge distillation approach: task-specific foundation models provide automatic supervision on the COCO 17 training set (118K images). The depth teacher is Depth Anything V2 (ViT-L, 335M parameters). The segmentation teacher is Detectron2. The text segmentation teacher is DPText. For resource-constrained or embedded deployments the model was also quantized as a TFLite and JAX FP16 (~210 MB) or ONNX FP16 (~209 MB) variant. In select experiments, Figure 57 shows Y-MAP-Net executed in ALT premises and Figure 62 demonstrates the method applied to monocular footage from a security camera at the TOFAS factory, extracting 2D joints, person and vehicle segmentation, and scene context without site-specific fine-tuning. The 2D joint heatmaps and PAFs serve as an independent verification signal for D-PoSE poses and as a fallback detector when D-PoSE's bounding-box input is unreliable (Figure 58, Figure 59 and Figure 61). Work is ongoing to expose per-instance person masks from the PAF association output so that Y-MAP-Net can eventually replace the YOLO-based person detector used as D-PoSE's Stage 1.



Figure 57: Successfully tracking workers performing defect repairs on TOFAS premises. 3D pose estimation is conducted using D-Pose, ArUco marker 3D transformations are overlayed on the visualization and the QR time-sync codes emitted by the laptop screen synchronize all streams with each other and also with respect to the ATI F/T sensor whose data is recorded using the same laptop.

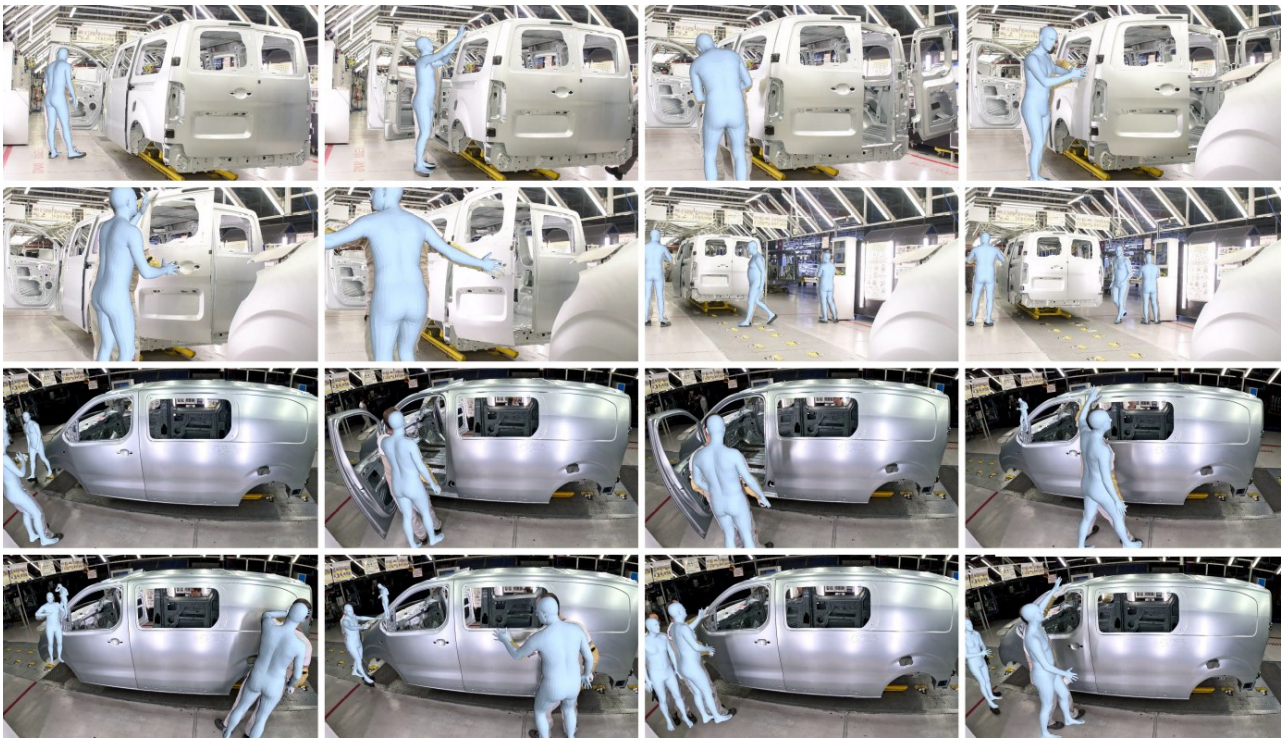


Figure 58: Successfully tracking people performing defect sensing on TOFAS premises.

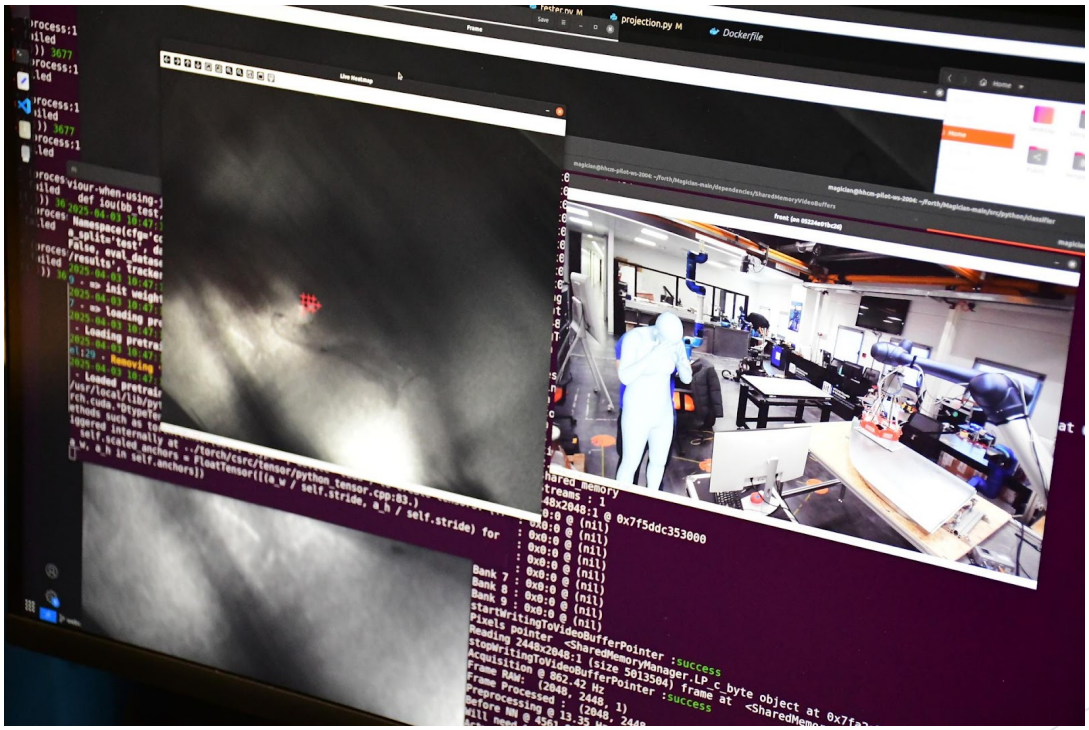


Figure 59: Successfully running D-Pose while also performing defect detection from the same computer system host during the 1st integration meeting in IIT.

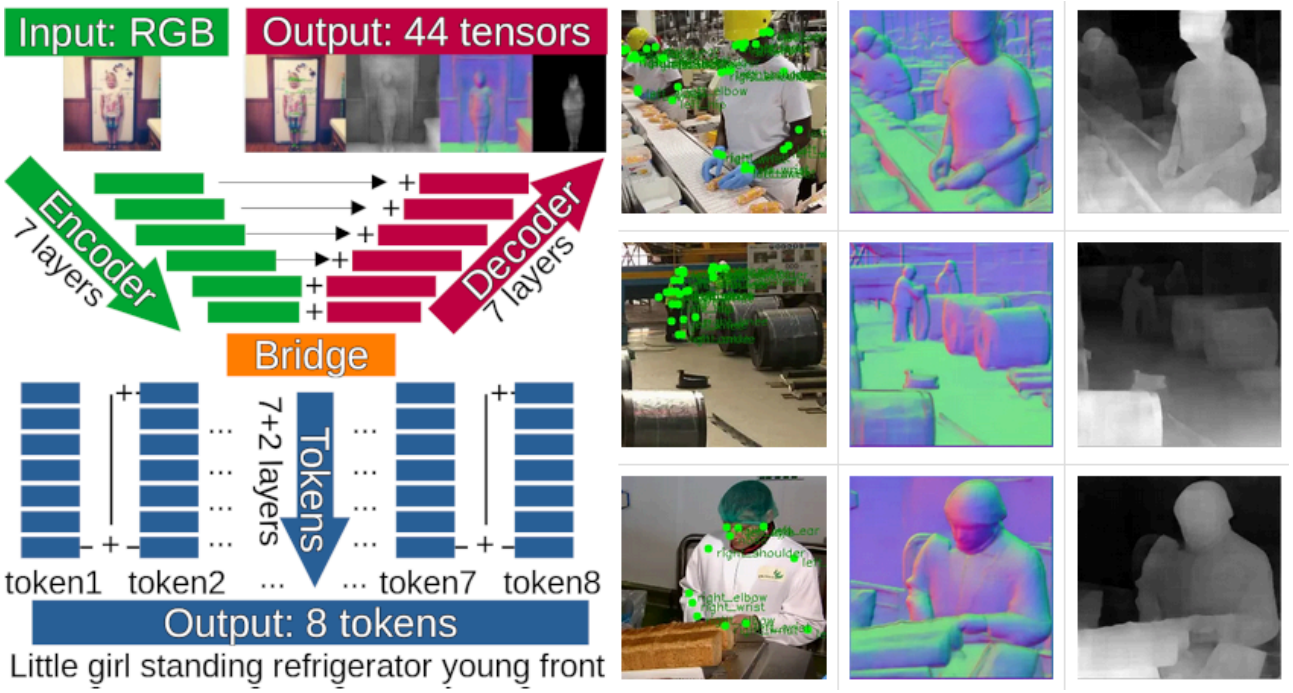


Figure 60: Left: The Y-MAP-Net (https://rasevents.org/presentation?id=177998) network architecture. Right: The network extracts 2D pose, normal vectors and depth maps for the whole scene from monocular RGB something useful in the context of industrial applications.

Accurate reconstruction of hand shape and pose from monocular RGB is critical for capturing the fine-grained tool manipulation skills that expert panel-beaters employ during surface inspection and repair. FORTH has developed and published work on this problem in close connection with the D-PoSE and Y-MAP-Net efforts. Our recent work [Karvounas26] proposes a lightweight texture module that embeds per-pixel observations into UV texture space, enabling a novel dense alignment loss between predicted and observed hand appearance. The key observation is that even high-performing 3D hand models produce reconstructions that do not perfectly align with the observed image appearance a discrepancy that texture-guided supervision can directly correct. The module is self-contained and pluggable into existing monocular reconstruction pipelines; in the paper it is attached to HaMeR, a high-performing transformer architecture for 3D hand pose estimation, producing measurable improvements in both accuracy and visual realism without requiring ground-truth texture annotations or multi-view input. In the MAGICIAN context, this work is directly relevant to the skill-transfer objective: fine-grained hand pose during tool use (pen marking, sanding, hammering) is a key input to the motion-primitive learning pipeline.

The perception of human motion and scene context also opens the way to higher-level interpretation of procedural activities. In MAGICIAN, this is relevant not only for safety monitoring, but also for analysing whether an inspection or rework action follows the intended procedure. A recent FORTH-led review on vision-based mistake analysis in

procedural activities [Bacharidis25] surveys methods for detecting and predicting execution errors in structured tasks, including incorrect action ordering, inappropriate technique, timing errors, and deviations from expected task progress.

This perspective is useful for MAGICIAN because expert repair demonstrations are not arbitrary motion sequences: they have procedural structure, tool-use constraints, and permissible variations. Distinguishing acceptable stylistic variation from true mistakes is therefore essential when demonstrations are used for learning robotic behaviour. The review highlights open challenges such as intra-class variability, viewpoint changes, error propagation, and the need to combine visual activity understanding with task-level structure. These issues match the MAGICIAN setting, where the system must interpret human rework actions from visual, pose, tool, and force/torque signals. Consequently, mistake-analysis methods provide a conceptual and methodological basis for future extensions of the MAGICIAN perception stack, including procedural validation of demonstrations, detection of anomalous repair executions, and richer feedback during human-robot collaborative workflows.



Figure 61: Successfully using Y-MAP-Net [Qammaz2026] to extract 2D pose, depth maps, normal vectors and segmentation maps in the observed scene.

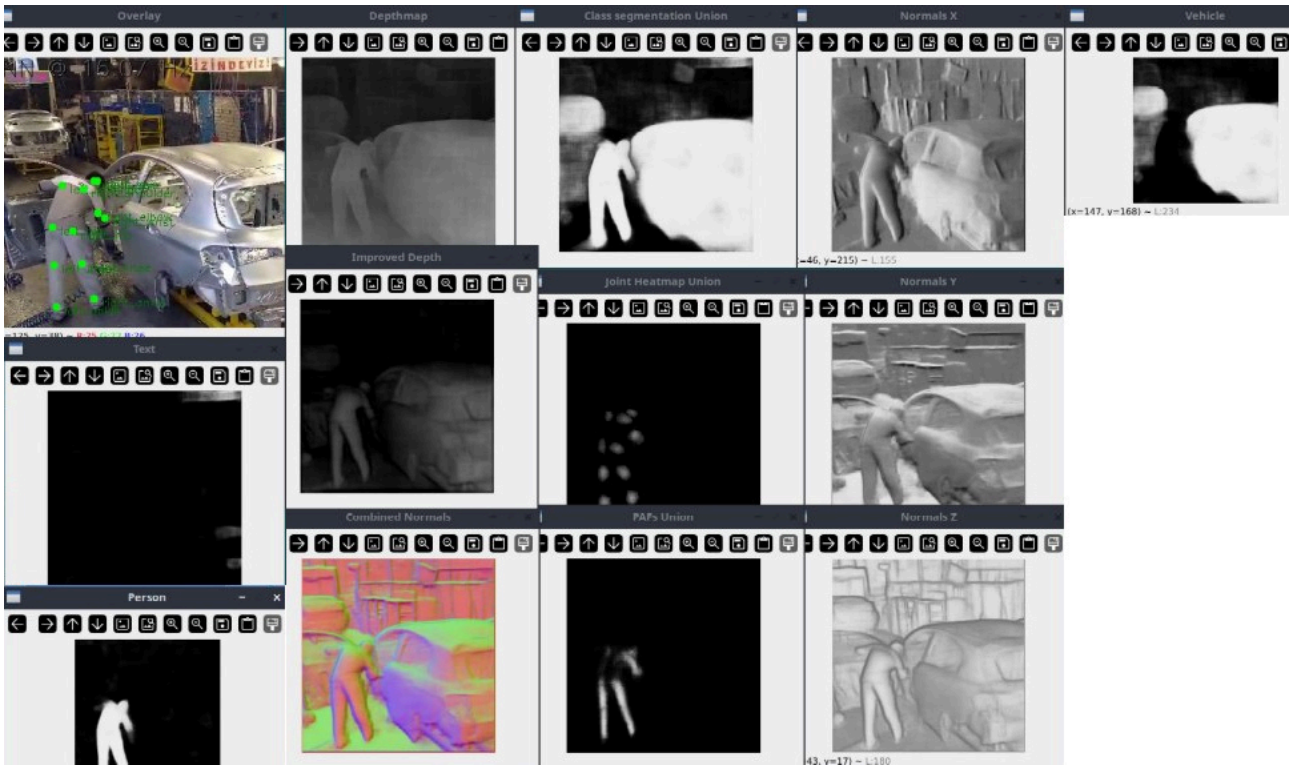


Figure 62: Employing Y MAP Net on monocular videos from a security camera located in the TOFAS factory and extracting 2D joints, person and vehicle segmentation masks, Depth map, Normal maps and Part Affinity Fields using one-pass evaluation in real-time (19Hz @ RTX 4070 GPU).

4.3.1.4 RESULTS AND FINDINGS

We successfully track human-related datasets and the surrounding environment both from the factory lines of the TOFAS plant as well as the ALT integrator premises. By using real-time segmented depth streams directly extracted from RGB, along with a minimum safety distance policy, we can establish a baseline human-aware motion planning module. Using the same ROS nodes and developed tools we can both support real-time closed-loop human motion prediction for the purposes of avoiding accidents on the space shared between people working in the factory lines where the MAGICIAN robots are deployed, as well as recording data from experts performing repair operations to allow learning the motion primitives of their unique repair execution.

D-PoSE is evaluated on the 3DPW and EMDB benchmark datasets using Mean Per Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), and Mean Vertex Error (MVE), all in millimetres. On 3DPW D-PoSE achieves PA-MPJPE = 43.6 mm, MPJPE = 68.9 mm, MVE = 81.4 mm outperforming the state-of-the-art TokenHMR (ViT backbone: PA-MPJPE = 44.0, MPJPE = 71.6, MVE = 85.7) on all three metrics despite using 83.8% fewer parameters (Figure 51). On EMDB D-PoSE achieves PA-MPJPE = 53.2 mm, MPJPE = 85.5 mm, MVE = 99.0 mm, reducing MPJPE by 4.1 mm and MVE by 7.2 mm compared to TokenHMR.

Y-MAP-Net is evaluated on the COCO 17 validation set using the Heatmap Distance Metric (HDM) and MSE. Joint detection achieves HDM = 0.98, PAFs HDM = 0.97, surface normal HDM = 0.64, text segmentation HDM = 0.98, class segmentation HDM = 0.97. Depth is the hardest output with MSE = 58.49, consistent with the inherent ambiguity of monocular depth in in-the-wild images. Despite training on only 118K samples, Y-MAP-Net matches its individual teacher models on their respective tasks while running all tasks simultaneously. By using real-time segmented depth streams directly from Y-MAP-Net we obtain per-pixel workspace geometry in addition to the skeletal information provided by D-PoSE, both from the same low-cost RGB camera and without any dedicated depth hardware. Figure 55 shows the system successfully resolving interactions between technicians and the SR arm during the 6th ALT Integration Week.

Similar to the case of the defect detection stack, the human and environment perception stack is also deep learning based. Despite the extensive training datasets additional annotated data will once again further enhance results. Additionally, depending on the available computational resources, by adding more layers and increasing the capacity of the neural networks they can yield even cleaner output, based on our needs following experimental evaluation of the solution.



Figure 63: Y-MAP-Net oral presentation during the IEEE International Conference on Robotics and Automation (ICRA) 2026

5 CONCLUSIONS

This deliverable presented the third and most advanced reporting stage of the perception systems developed within the MAGICIAN project. Building upon the foundations established in D3.1 and the technological advancements reported in D3.2, the work described in this document demonstrates the substantial maturation of the project's sensing technologies and their progression towards deployment in representative industrial environments.

Significant progress has been achieved across all perception-related activities. The visual perception module has evolved into a robust defect detection framework combining dedicated sensing hardware, advanced image processing techniques, and machine-learning-based classification methods.

The tactile perception module has further strengthened the inspection capabilities of the platform by providing complementary information through force and vibration sensing, enabling the detection of defects that may be difficult to identify through vision alone. Together, these technologies establish the foundations of a multi-modal perception framework capable of supporting reliable industrial inspection tasks.

In parallel, substantial advances have been made in human-centred perception. The human motion perception and prediction modules provide the platform with awareness of operators and their activities, supporting the project's objective of enabling safe and effective human-robot collaboration.

A key achievement of the reporting period has been the integration of the individual perception components into a unified software and hardware ecosystem. The successful deployment of the complete sensing stack during project integration activities demonstrated not only the functionality of the individual modules but also their ability to operate jointly under conditions representative of the target industrial use case. This transition from isolated laboratory prototypes to integrated system validation represents an important milestone for the MAGICIAN project.

The results presented in this deliverable confirm that the perception technologies developed within WP3 have reached a level of maturity suitable for supporting the final integration, demonstration, and validation activities of the project. The next phase of work will focus on further improving robustness and scalability, strengthening multi-modal fusion strategies, and supporting the integration of perception outputs with the planning, control, and repair functionalities of the complete MAGICIAN platform.

6 REFERENCES

[Ant23] Antonucci, A., Bevilacqua, P., Leonardi, S., Paolopoli, L., & Fontanelli, D. (2023). Humans as path-finders for mobile robots using teach-by-showing navigation. Autonomous Robots, 47(8), 1255-1273.

[Bacharidis25] K. Bacharidis and A. A. Argyros, "Vision-based mistake analysis in procedural activities: A review of advances and challenges," arXiv preprint arXiv:2510.19292, 2025.

[Karvounas26] G. Karvounas, N. Kyriazis, I. Oikonomidis, G. Pavlakos, and A. A. Argyros, "Enhancing monocular 3D hand reconstruction with learned texture priors," in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, Arizona, USA, 2026. Also available as arXiv preprint arXiv:2508.09629.

[Shehaj26] A. Shehaj, N. Vasilikopoulos, I. Oikonomidis, and A. A. Argyros, "Temporally consistent 3D human pose and shape estimation from video," under submission to the International Conference on Virtual Reality (ICVR), 2026.

[Yang24] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything v2," arXiv preprint arXiv:2406.09414, 2024.

[Tian2024] S. Tian, M. Zheng, and X. Liang, "Transfusion: A practical and effective transformer-based diffusion model for 3D human motion prediction," IEEE Robotics and Automation Letters, vol. 9, no. 7, pp. 6232–6239, 2024.

[Qammaz2026] A. Qammaz, N. Vasilikopoulos, I. Oikonomidis and A.A. Argyros, "Y-MAP-Net: Learning from Foundation Models for Real-Time, Multi-Task Scene Perception", In IEEE International Conference on Robotics and Automation (ICRA 2026), June 2026.