

CORVINUS UNIVERSITY OF BUDAPEST

**COMPUTER VISION-BASED FRAMEWORK
FOR PRODUCT IDENTIFICATION IN
DYNAMIC ENVIRONMENTS**

PH.D. DISSERTATION

Supervisor: Dr. Andrea Kő

Supervisor: Dr. Tamás Ruppert

Zeleny Klaudia Éva

Budapest, 2026

Zeleny Klaudia Éva

Computer Vision-Based Framework for Product
Identification in Dynamic Environments

Doctoral School of Economics and Business Informatics

Supervisor: Dr. Andrea Kő

Supervisor: Dr. Tamás Ruppert

Corvinus University of Budapest

Doctoral School of Economics and Business Informatics

**Computer Vision-Based Framework for
Product Identification in Dynamic
Environments**

Ph.D. Dissertation

Zeleny Klaudia Éva

Budapest, 2026

CONTENTS

1	Introduction	8
1.1	Background and motivation	8
1.2	Visual identification in warehouse environments	11
1.3	Thematic positioning of the dissertation	12
1.4	Structure of the Dissertation	13
2	Related Work	15
2.1	Search strategy and logic	15
2.2	Structured overview of the selected literature	15
2.2.1	Overview of relevant review papers	15
2.2.2	Overview of relevant research papers	21
2.3	Research gap	29
3	Research problem, objectives and questions	32
4	Methodology	38
5	Theoretical Framework	41
5.1	HMLV intralogistics context	41
5.2	Visual detection, instance segmentation, and structured identification	42
5.3	Data Matrix	44
5.4	Binary and continuous module similarity	48
5.5	Evidence aggregation across observations	49
6	Model and System Design	50
6.1	System-level architecture	52
6.2	Dataset design and data acquisition	55
6.2.1	Acquisition setup	55
6.2.2	Macro-scene concept	55
6.2.3	Controlled environmental variability	56
6.2.4	Annotation protocol	58
6.2.5	Episode annotation and construction	59
6.2.6	Train-validation-test separation	59
6.2.7	Dataset statistics	60
6.3	Data Matrix detection model	61
6.3.1	Model selection	61
6.3.2	Training configuration	63
6.3.3	Data augmentation	63

6.3.4	Detection output	64
6.4	Primary decoding	67
6.4.1	Mask-derived quadrilateral and crop extraction	67
6.4.2	Primary Data Matrix decoding and reference validation	67
6.5	Recovery branch and similarity-based identification	69
6.5.1	Frame-level evidence and dominance rule	75
6.6	Temporal integration and episode-level decision	77
6.7	Operational interpretation of episode-level identification outcomes	80
7	Results	82
7.1	Evaluation scope and reporting conventions	82
7.1.1	Evaluation measures	85
7.2	Object detection	90
7.3	Primary decoding	93
7.4	Recovery branch	94
7.4.1	Sensitivity to an absolute support floor	102
7.5	Frame-level pipeline	103
7.5.1	Performance under controlled variability	105
7.5.2	Effect of lighting	107
7.5.3	Effect of blur	108
7.5.4	Effect of object condition	108
7.5.5	Effect of viewpoint angle	109
7.5.6	Effect of relative symbol area	109
7.5.7	Absolute symbol size and the decoding operating range	110
7.5.8	Frames containing more than one symbol	112
7.5.9	Effect of localization on identification	113
7.5.10	Summary of frame-level findings	114
7.6	Sequence-level pipeline	114
7.6.1	Evaluation scope and overall outcome	115
7.6.2	Outcomes under controlled lighting variability	116
7.6.3	Evidence basis of the sequence decisions	117
7.6.4	Contribution of temporal integration	118
7.6.5	Latency to the first correct observation	120
7.6.6	How much of an episode the decision requires	121
7.6.7	Cost of frame sampling	121
7.6.8	Measured implementation cost and throughput scope	123
7.6.9	Effect of confirmed symbol visibility	125
7.6.10	Summary of sequence-level findings	125
7.7	Ablation of the recovery stages	126

7.7.1	Ablation of the similarity measure	128
7.7.2	Ablation of the accumulation rule	129
8	Discussion	130
8.1	Discussion of RQ1: system-level design principles	130
8.2	Discussion of RQ2: implementation and empirical evaluation under controlled variability	132
8.3	Discussion of RQ3: operational implications for inventory-related decision support	134
8.4	Positioning the results against prior work	136
9	Conclusion	140
9.1	New scientific contributions	143
A	Literature search protocol	144
A.1	Conceptual dimensions and keyword formulation	144
A.2	Query construction	145
A.3	Query results and interpretation	146
A.4	Selection and screening process	148
B	Supplementary qualitative and distribution figures	150

1. INTRODUCTION

This dissertation investigates the role of visual perception in supporting product identification and the generation of inventory-related evidence in indoor warehouse environments. The research focuses on high-mix low-volume (HMLV) intralogistics systems, where dynamic product placement and operational variability limit the applicability of conventional identification workflows.

The proposed approach examines video recorded by a mobile two-dimensional camera, representing the type of visual observations that could later be acquired from material-handling platforms such as forklifts. Rather than assuming that each frame provides a complete and reliable identification, the study investigates how repeated observations can be processed and combined into product-identification evidence. The resulting pipeline is examined as a conceptual system architecture under controlled, HMLV-inspired environmental variability, rather than as a finalized warehouse deployment or an implemented inventory-management solution.

1.1 Background and motivation

Logistics systems are characterized by complex interactions between physical processes and information flows. Effective coordination of these processes requires timely and reliable operational data, particularly in environments where product variety, demand variability, and spatial dynamics challenge traditional planning assumptions. Data-driven and perception-based approaches have therefore become increasingly relevant in supporting decision-making and operational transparency in contemporary logistics systems.

A central objective of logistics research is to improve the efficiency and effectiveness of logistics processes by enhancing the quality, availability, and usability of information. Data-driven methods and decision-support tools are increasingly used to assist logistics managers and practitioners in monitoring system performance, identifying bottlenecks, and adapting operations to changing conditions. The growing availability of data, combined with advances in data-driven and machine learning-based methods, has further expanded the range of analytical approaches applicable to logistics problems.

Within this broader context, intralogistics has undergone significant transformation in recent years, driven by increasing automation, digitalization, and the growing complexity of supply chains. Warehouses are no longer static storage facilities but dynamic operational environments that must continuously adapt to changing demand, product variety, and process requirements. As a result, internal logistics systems are expected to provide accurate, timely information that supports operational transparency, traceability, and data-driven decision-making across a wide range of warehouse processes (Tadjine et al., 2025; Winkelhaus et al., 2021).

One important aspect of improving intralogistics performance is the reliable identification and tracking of products and handling units. In warehouse operations, Warehouse Management Systems (WMS) play a central role in coordinating storage, picking, and internal transportation activities. These systems typically rely on explicit machine-readable identifiers, such as one-dimensional barcodes, QR codes, or Data Matrix codes, to link physical objects to digital records throughout the warehouse lifecycle. While WMS platforms integrate data from multiple sources, the accuracy and timeliness of this information often depend on manual scanning, human data entry, or controlled acquisition processes. Machine-readable identifiers remain widely used in warehouse operations because they provide an explicit connection between physical objects and digital records and can be integrated with existing information systems (Winkelhaus et al., 2021; Losada del Olmo et al., 2025).

Product location and product identification constitute related but distinct operational problems. When the recorded storage location is assumed to be correct, a WMS or picking-support system can direct an operator to that location. Once the location has been reached, the product may still have to be identified or verified. A different problem arises when the recorded location is missing, insufficiently precise, or inconsistent with the physical warehouse state. In this case, navigation to the recorded location does not by itself resolve the discrepancy, and additional searching or inventory verification may be required (Winkelhaus et al., 2021; Vaccari et al., 2024).

In the industrial case motivating this dissertation, incomplete or unreliable information about the physical location of products could result in manual product searches. This observation is treated as case-specific motivation rather than as evidence that manual searching is the only or generally dominant practice in HMLV warehouses. The literature reviewed for this dissertation does not provide a sufficiently comparable quantitative basis for estimating the prevalence or operational cost of such searches across HMLV facilities. The motivation is therefore based on the observed industrial problem, while the broader technological context is established through the literature.

The effectiveness of such identification-based approaches is closely tied to the structural characteristics of the warehouse environment. In highly structured, large-volume warehouse systems, where product types are limited and storage layouts are stable, predefined processes and standardized identification workflows can be applied successfully. In contrast, HMLV warehouses are characterized by a large number of distinct product types, low repetition rates, frequent product changes, and limited process standardization (Gan et al., 2023; Simeth et al., 2025). In the HMLV warehouse environment motivating this dissertation, these general characteristics are accompanied by non-fixed or temporary storage locations, non-uniform packaging, and changing product and identifier orientations. These conditions complicate the application of automation and visual-identification strategies that assume stable layouts, standardized physical appearances, consistent orientations, or

predefined object models.

HMLV operating conditions occur across multiple industrial contexts rather than being limited to a single sector. Manufacturing and assembly applications are prominently represented in the HMLV literature (Gan et al., 2023; Simeth et al., 2025). From the intralogistics perspective adopted in this dissertation, relevant application contexts include automotive manufacturing, custom assembly, maintenance logistics, and spare-part distribution. In these contexts, the identification problem may involve many distinct parts or products with low individual repetition rates and heterogeneous physical characteristics. High product variability, dynamic space utilization, and evolving operational constraints introduce substantial complexity into both physical handling and information management processes. Recent studies in HMLV manufacturing and logistics contexts highlight that variability and low repetition can challenge the robustness of automation and perception systems, particularly when visual information must be interpreted under changing environmental conditions (Simeth et al., 2025).

Beyond traditional transactional and planning-related data, contemporary intralogistics systems increasingly rely on heterogeneous data sources to achieve operational visibility and responsiveness. These include sensor readings, machine logs, positional data, and visual information captured within warehouse environments. While structured data supports higher-level planning and optimization, many operational decisions require awareness of the physical state of the warehouse in real time. In this context, visual data represents a particularly rich but complex source of information, as it directly reflects the spatial arrangement, condition, and identity of physical objects. As warehouses become more dynamic and less structured, the role of visual information becomes increasingly central to data-driven intralogistics, motivating growing research interest in computer vision-based identification approaches (Tadjine et al., 2025; Losada del Olmo et al., 2025).

A range of operational and technological approaches may be used to support warehouse product location and identification. WMS-based location management and paper-based or digital picking instructions support navigation to recorded storage locations. Pick-by-light, pick-by-voice, and pick-by-vision technologies assist operators during picking activities. Barcodes and Data Matrix codes provide explicit visual identifiers, while RFID and indoor positioning technologies can support contactless identification, tracking, or location estimation. Robot-, UAV-, and camera-based systems provide further possibilities for inventory inspection and warehouse monitoring (Winkelhaus et al., 2021; Vaccari et al., 2024; Tadjine et al., 2025).

These technologies address different operational requirements and are not directly interchangeable. WMS-based navigation assumes that recorded locations are sufficiently accurate. Conventional visual-code scanning requires an explicit acquisition event and adequate visibility of the identifier. RFID and indoor positioning systems require compatible tags, readers, and supporting infrastructure, and their performance depends on the physical

environment and the selected technology. Picking-support systems primarily assist the execution of known tasks and do not necessarily resolve discrepancies between the recorded and physical warehouse states (Losada del Olmo et al., 2025; Vaccari et al., 2024).

Computer vision is therefore not positioned in this dissertation as a universally preferable alternative to WMS, barcode, RFID, or indoor positioning technologies. Its relevance follows from the conditions of the investigated problem: the products already carry machine-readable visual identifiers, multiple observations can be collected from video, and the resulting evidence may complement transaction-based records without requiring an additional identification technology on each observed item. This rationale motivates the dissertation’s focus on computer vision-based Data Matrix identification under HMLV-inspired environmental variability.

1.2 Visual identification in warehouse environments

The increasing flexibility of warehouse operations in HMLV environments introduces substantial challenges for perception and identification systems. While identifier-based tracking remains a core component of intralogistics workflows, the practical effectiveness of such systems depends not only on the presence of identifiers but also on their reliable detection and interpretation under real operating conditions. In contrast to structured warehouse settings, HMLV environments exhibit a high degree of visual and operational variability that directly affects the performance of identification technologies.

In HMLV warehouses, variable product placement and mobile image acquisition cause identifiers to appear under changing scale, viewpoint, illumination, and visibility. Occlusion, perspective distortion, reflections, low contrast, and motion blur therefore affect both localization and decoding. These acquisition conditions motivate a system design that can use uncertain frame-level observations without treating each observation as an independent final decision (Almeida et al., 2021; Simeth et al., 2025).

Object detection localizes candidate regions, after which a dedicated Data Matrix decoder attempts to extract the encoded identifier. Detection errors may propagate to decoding, however successful detection does not necessarily result in successful decoding. Conversely, a decoder may return a syntactically plausible payload that cannot be accepted as a valid identifier within the applicable reference set. Detection, decoding, and identification must therefore be treated as related but distinct stages of the complete system.

Prior research on Data Matrix recognition in cluttered indoor environments has combined deep learning-based symbol detection with a dedicated conventional decoder, demonstrating that localization and payload extraction constitute separate components of the complete recognition process (Almeida et al., 2021). This architectural separation is important because uncertainty generated during detection, cropping, or geometric normalization may affect the information subsequently available to the decoder.

Within the proposed pipeline, identification refers to the decision that an available payload provides admissible evidence for a known item-level identifier. A directly decoded or recovery-generated candidate is therefore not automatically treated as an accepted identification. It must first be evaluated against the reference set and the applicable decision rules. This distinction separates perception outputs from the higher-level decision that determines whether the available evidence is sufficient for an accepted, ambiguous, or unresolved outcome.

Video acquisition also provides multiple observations of the same physical identifier. Instead of treating every frame as an independent final decision, the proposed architecture interprets frame-level results as uncertain evidence that may be accumulated across a tracked symbol episode. Direct decoding results, recovery candidates, and their supporting observations can therefore contribute to an episode-level decision. Frame-level evaluation characterizes the behaviour of the individual perception and recovery stages, whereas episode-level evaluation assesses whether evidence distributed across multiple observations can support a reliable identification outcome.

Accordingly, the dissertation treats visual identification as a system-level problem in which uncertainty may propagate from perception to decoding and decision, and in which identification reliability must be evaluated at both frame and episode levels.

1.3 Thematic positioning of the dissertation

This dissertation is positioned within the broader research area of data-driven intralogistics, with a particular emphasis on the use of visual information in warehouse environments. The work is situated at the intersection of logistics, artificial intelligence, and computer vision, reflecting the increasing role of data-driven and perception-based approaches in supporting operational visibility and decision-making in complex logistics systems.

The thematic focus is limited to indoor warehouse settings, where visual data captured by cameras represents an important source of information about the physical state of the environment. Within this context, the dissertation concentrates on visual identification tasks related to machine-readable product codes, which play a central role in linking physical objects to digital logistics systems. The investigated task is identifier-based product identification rather than open-world recognition based solely on product appearance.

By narrowing its thematic scope to visual identification in warehouse environments, the dissertation addresses a specific aspect of intralogistics where physical variability, environmental conditions, and data interpretation intersect. This positioning provides the conceptual foundation for the subsequent chapters, which progressively refine the research problem and research objectives of the dissertation.

The empirical investigation is conducted in a controlled, HMLV-inspired environment

designed to reproduce selected sources of visual variability. Although the proposed architecture provides a conceptual basis for using validated visual observations as inventory-related evidence, the implemented system does not estimate warehouse coordinates, update WMS records, or validate operational performance in an industrial deployment. Product identification, geometric localization, and inventory-state updating are therefore treated as distinct system functions.

This dissertation contributes to the system-level understanding of visual product identification in dynamic warehouse environments by proposing an integrated perception-driven identification framework, a structured evaluation design under controlled environmental variability, and a conceptual interpretation of operational implications for intralogistics systems. The contribution is consequently positioned at the levels of system architecture, methodological evaluation design, and applied operational interpretation rather than as a new standalone detection or decoding algorithm.

Throughout this dissertation, product identification refers to the identification of an individual physical item through its item-level Data Matrix identifier, rather than to product-category or SKU classification.

1.4 Structure of the Dissertation

The dissertation is organized into nine chapters, progressing from problem formulation and theoretical foundations toward system design, experimental validation, and practical implications.

Section 2 reviews the **related literature**. It first introduces the search strategy and selection logic, followed by a structured overview of review papers and primary research articles. The chapter synthesizes findings across intralogistics, computer vision, deep learning-based object detection, Data Matrix detection, dedicated decoding, and visual identification systems. It concludes with a summary of identified research gaps that inform the methodological design and contributions of this dissertation.

Section 3 defines the **research problem** in detail, focusing on open and challenging issues related to visual identification approaches in HMLV warehouse environments. It formulates the research objectives and research questions, translating the identified problem into clearly defined goals and outlining the scope and expected theoretical and practical contributions of the study.

Section 4 presents the research **methodology** and the overall research framework. It describes the design-oriented approach of the study and introduces the structured research workflow, including dataset design, model development, dedicated Data Matrix decoding, similarity-based recovery, temporal decision logic, and evaluation strategy.

Section 5 introduces the **theoretical framework** required to understand the study. It covers the characteristics of HMLV intralogistics systems, fundamental concepts of computer

vision and deep learning relevant to instance segmentation, and the structural and decoding properties of Data Matrix codes as visual identifiers.

Section 6 presents the **model and system design**. Subsection 6.1 describes the system-level architecture of the proposed detection, dedicated decoding, recovery, and temporal-reasoning pipeline. Subsection 6.2 details the dataset design and data acquisition principles, including macro-scene-based train–validation–test separation and controlled environmental variables. The chapter further explains the Data Matrix detection model, decoding and geometric preprocessing steps (cropping, warping, perspective correction), similarity-based validation logic, and the end-to-end pipeline integration. It concludes with the experimental setup and evaluation protocol.

Section 8 **discusses** the results in relation to the research questions and existing literature. It interprets the observed performance patterns, evaluates practical applicability in HMLV intralogistics systems, and reflects on methodological limitations and environmental constraints.

Finally, Section 9 **concludes** the dissertation by summarizing the main contributions, including the structured HMLV dataset design, the integrated detection, Data Matrix decoding, recovery, and temporal-reasoning pipeline, and the empirical validation results. It also outlines directions for future research, such as movement-data integration, domain adaptation, multi-camera extensions, and real-time deployment scenarios.

2. RELATED WORK

2.1 Search strategy and logic

The literature search aimed to systematically map research at the intersection of computer vision, deep learning, and intralogistics, with particular emphasis on visual code detection and Data Matrix recognition in HMLV environments. Scopus was selected as the primary database for its coverage of peer-reviewed journals and conference proceedings and to support reproducibility of the search process.

The search logic was structured around four conceptual dimensions, each denoted by a shorthand used throughout this chapter: **CV** (computer vision and deep learning methods, e.g. “deep learning”, “computer vision”, “object detection”), **QR** (two-dimensional visual code recognition, e.g. “QR code”, “datamatrix”, “data matrix”, “barcode”), **Domain** (intralogistics and warehouse localization tasks, e.g. “intralogistics”, “inventory management”, “smart warehouse”, “product localization”), and **HMLV** (high-mix low-volume settings, e.g. “high-mix low-volume”, “HMLV”). An additional *review* keyword was used to separate survey-oriented queries from research-oriented ones, since the two were not mutually exclusive and overlaps were resolved during manual screening.

The four dimensions were combined using Boolean operators, restricted to English-language journal articles, conference papers, and review papers published from 2019 onward. Combining all four dimensions returned zero publications, so the search was systematically relaxed to three-dimension and then two-dimension combinations across fourteen queries in total. Publications were then screened in two stages: a title-abstract-keyword screening against a fixed set of relevance criteria (problem definition, publication type, industrial domain, HMLV relevance, environmental or visual factors, 2D-code focus, and overall relevance), followed by a full-text review assessing methodological depth, technological contribution, and applicability to HMLV warehouse environments. This process yielded a final corpus of 34 review papers and 28 research papers, which forms the empirical basis of this chapter and of the structured identification of research gaps.

The full query log, the result counts across all fourteen dimension combinations, and the detailed screening statistics are reported in Appendix A.

2.2 Structured overview of the selected literature

2.2.1 Overview of relevant review papers

This chapter synthesizes recent systematic reviews and surveys on vision-based perception, localization, and object detection in robotics, autonomous systems, and industrial environments, with a focus on methodological patterns, dataset characteristics, environ-

ment realism, and research gaps, particularly as they pertain to HMLV intralogistics and related domains. The objective of this overview is to position the existing literature within a coherent thematic structure and to identify common trends and persistent limitations relevant to real-world industrial deployment.

To structure the discussion and enable a coherent comparison of the reviewed literature, the identified review papers are grouped according to their primary research focus. Table 1 summarizes the thematic grouping and the corresponding publications.

Group A (Object-centric visual perception) encompasses surveys focused on object-level visual perception, including detection, segmentation, recognition, and emerging open-vocabulary and unsupervised approaches. While these methods achieve strong performance on standard benchmarks, their applicability to industrial and HMLV environments is constrained by dataset bias, incomplete annotations, and limited robustness to occlusion and environmental variability.

Group B (Localization, mapping, and ego-motion) addresses spatial understanding through visual localization, mapping, and motion estimation, with SLAM and odometry forming the methodological backbone. Despite increasing robustness through sensor fusion, most approaches are evaluated in structured or idealized settings, limiting their transferability to dynamic, cluttered HMLV intralogistics environments.

Table 1: Thematic grouping of the reviewed survey papers according to their primary research focus

Group	Thematic focus	Publications
A	Object-centric visual perception	Du et al. (2021), Ilioudi et al. (2022), Wudhikarn et al. (2022), Zhang et al. (2021), Diwan et al. (2023), Abosinnee et al. (2025), Awan et al. (2025), Nani et al. (2025), Yu et al. (2025), Papadopoulos et al. (2025), Alam et al. (2021), Varghese and Sambath (2023), Yadav and Kumar (2023), Zilpe et al. (2023), Krishna et al. (2024), Issaoui et al. (2024), Rekavandi et al. (2025)
B	Localization, mapping, and ego-motion	Morar et al. (2020), Agostinho et al. (2022), Dias et al. (2023), Couturier and Akhloufi (2024), Rodríguez-Martínez et al. (2025), Prasad et al. (2025), Yu (2023), Nwobodo et al. (2023)
C	Integrated perception pipelines and industrial systems	Al-hamadani and Al-Darraj (2025), Faqeer and Khajavi (2025), Albayrak Ünal et al. (2023), Heruatmadja et al. (2024), de Moura Junior et al. (2024)

Group C (Integrated perception pipelines and industrial systems) focuses on system-level vision pipelines and industrial applications, including warehouse environments, manufacturing systems, and digital twin-based integration. Although these surveys emphasize deployment context and operational monitoring, they often lack quantitative evaluation of perception robustness under real-world HMLV constraints such as high object diversity and frequent scene reconfiguration.

Methodological patterns

Across the surveyed review literature, vision-based localization, detection, and scene understanding are predominantly driven by deep convolutional architectures, often enhanced with attention mechanisms, transformer components, or multi-modal fusion. At the same time, classical geometric and probabilistic formulations remain foundational in SLAM and visual odometry pipelines.

In indoor localization and navigation, (Morar et al., 2020; Agostinho et al., 2022) describe a broad methodological spectrum ranging from feature-based and direct SLAM systems to hybrid frameworks that integrate learning-based depth or pose estimation into otherwise geometric pipelines. Classical approaches rely on handcrafted descriptors (e.g., SIFT, SURF, ORB, FAST) combined with geometric verification and probabilistic filtering techniques (such as RANSAC-based outlier rejection, bundle adjustment, and Kalman or particle filtering) for pose refinement and uncertainty modeling. In contrast, deep learning-based methods employ convolutional backbones pretrained on large-scale benchmarks (e.g., ImageNet, COCO) for feature extraction, detection, and semantic augmentation (Alam et al., 2021; Diwan et al., 2023; Zilpe et al., 2023; Krishna et al., 2024; Nani et al., 2025). While fully end-to-end CNN-RNN odometry systems have been proposed, reviews consistently report that hybrid designs embedding learned modules within geometric optimization pipelines currently demonstrate more stable and robust performance.

Object detection research consistently distinguishes between the two-stage and one-stage paradigms. Two-stage detectors (R-CNN family, Mask R-CNN, FPN-based models) remain reference architectures for high-precision localization and instance segmentation in cluttered scenes (Varghese and Sambath, 2023; Zilpe et al., 2023; Krishna et al., 2024). Their strength lies in accuracy, though at higher computational cost. One-stage detectors (YOLO family, SSD, RetinaNet, EfficientDet) prioritize real-time performance, making them suitable for embedded or mobile platforms, albeit sometimes at reduced performance for small or densely packed objects (Diwan et al., 2023; Rekavandi et al., 2025). In warehouse and inventory-related applications, both families are used, reflecting a trade-off between runtime constraints and localization precision. Although several studies report accuracies exceeding 97%, these results are frequently obtained under controlled acquisition conditions (Wudhikarn et al., 2022; Heruatmadja et al., 2024; de Moura Junior et al., 2024).

Robotic grasping research structures perception into three interconnected subtasks: object localization, 6D pose estimation, and grasp estimation (Du et al., 2021). While localization overlaps methodologically with general detection frameworks, 6D pose estimation introduces correspondence-based, template-based, and regression or voting-based strategies for recovering full spatial orientation. Grasp estimation extends this pipeline using planar or full 6DoF grasp networks operating on depth maps or point clouds. Many contemporary approaches rely on 3D representations and point cloud encoders (e.g., PointNet-style architectures); some bypass explicit pose estimation and directly regress grasp configurations from partial observations.

In autonomous driving, SLAM and visual odometry remain predominantly geometry-driven. According to Agostinho et al. (2022), learning-based components improve robustness—particularly for depth and pose estimation—but fully end-to-end approaches have not consistently surpassed classical pipelines in complex, dynamic scenarios. Multi-modal fusion (camera–LiDAR–IMU) enhances robustness but increases system complexity and computational demands. Core algorithmic elements such as bundle adjustment, pose-graph optimization, and sliding-window estimation remain central across implementations.

Across domains, preprocessing and post-processing steps are rarely emphasized despite their practical relevance. Typical preprocessing includes denoising, contrast normalization, histogram equalization, geometric transformations (rotation, scaling, cropping), and task-specific operations such as background subtraction or depth-based separation (Morar et al., 2020; Rekavandi et al., 2025). Post-processing commonly involves non-maximum suppression, mask refinement, clustering, and voting schemes (Dias et al., 2023; Papadopoulos et al., 2025). These components significantly influence end-to-end performance but are often treated as implementation details rather than systematically evaluated design choices.

Evaluation protocols vary across domains but are internally consistent within subfields. Object detection relies on AP/mAP at predefined IoU thresholds, frequently complemented by recall measures and runtime indicators (Diwan et al., 2023; Krishna et al., 2024). SLAM performance is measured using trajectory-based metrics such as ATE and RPE on benchmarks including KITTI and EuRoC (Agostinho et al., 2022). Small-object and unsupervised localization research adapts detection-style metrics, though several reviews note that inherited supervised protocols may inadequately capture the challenges of small-object perception or open-world discovery (Rekavandi et al., 2025; Papadopoulos et al., 2025). As a result, cross-paper comparability remains limited due to heterogeneous backbones, post-processing schemes, and training regimes.

Dataset characteristics and limitations

The reviewed surveys reveal a recurring dependency on a small number of widely used

benchmarks supplemented by domain-specific or proprietary datasets.

In general detection tasks, ImageNet and MS COCO dominate as pretraining and benchmarking corpora (Alam et al., 2021; Krishna et al., 2024). While these datasets provide scale and category diversity, they underrepresent extreme lighting, small objects, and domain-specific industrial artifacts.

Indoor SLAM evaluation primarily relies on TUM RGB-D, EuRoC, KITTI, and related benchmarks (Morar et al., 2020). These datasets represent offices, apartments, and structured driving environments with limited semantic diversity. Similarly, robotic grasping benchmarks are few in number and largely instance-level, limiting generalization to unseen industrial objects (Du et al., 2021).

In barcode recognition, dataset fragmentation is particularly pronounced. As reported by Wudhikarn et al. (2022), most barcode datasets are private or synthetic, with only a small subset publicly accessible. Public datasets are often limited in scale and captured under favorable acquisition conditions. This constrains reproducibility and cross-study comparability. Moreover, deep learning-based barcode systems are typically evaluated on clean, well-lit images that differ substantially from harsh warehouse environments.

Warehouse and inventory management research exhibits similar constraints. Studies frequently rely on self-collected datasets with limited documentation regarding environmental diversity, scale, or acquisition protocols (de Moura Junior et al., 2024). Inventory-focused AI reviews report no standardized shared datasets across contributions (Albayrak Ünal et al., 2023). Digital twin-computer vision integrations are predominantly validated in laboratory-scale scenarios, with minimal evidence of large-scale industrial deployment (Faqeer and Khajavi, 2025).

Small object detection research, particularly in maritime and remote sensing domains, also suffers from benchmark scarcity and heterogeneity (Rekavandi et al., 2025). Unsupervised localization inherits detection benchmarks but adapts them in ways that may penalize valid object discoveries due to incomplete ground truth annotations (Papadopoulos et al., 2025).

Across domains, annotation limitations are recurrent. Dense, exhaustive labeling is rare; small or ambiguous objects are often omitted. In barcode datasets, distorted or partially occluded codes lack standardized ground truth. In unsupervised settings, pseudo-label generation introduces additional uncertainty.

Environmental representativeness remains uneven. Autonomous driving datasets provide structured urban diversity but limited extreme conditions. Indoor SLAM relies heavily on office-like environments. Robotic grasping focuses on tabletop setups. Barcode and warehouse datasets are frequently collected under favorable lighting and structured layouts, insufficiently modeling occlusion, damage, clutter, or dynamic interactions typical of operational warehouses.

Environment realism and deployment gaps

A consistent theme across reviews is the discrepancy between benchmark evaluation and real-world deployment.

Indoor localization methods are often tested in small, static environments, whereas industrial settings are large, cluttered, and continuously changing (Morar et al., 2020). Marker-based systems achieve high precision but require dense artificial infrastructure.

Autonomous driving benchmarks capture structured challenges but do not fully reflect dynamic, adverse-weather, or feature-sparse conditions (Agostinho et al., 2022). Multi-modal systems improve robustness but are evaluated on limited sequences.

Robotic grasping research largely relies on controlled laboratory setups with planar surfaces and fixed viewpoints (Du et al., 2021). High-mix industrial environments, characterized by deformable, reflective, or occluded objects, remain insufficiently modeled.

Barcode recognition and warehouse vision systems are typically validated on clean images captured with smartphone or industrial cameras under controlled lighting (Wudhikarn et al., 2022; de Moura Junior et al., 2024). Pilot deployments dominate; systematic large-scale industrial validation is rare (Faqeer and Khajavi, 2025).

Sensor placement assumptions further reduce realism. Static cameras, predefined trajectories, and standardized sensor mounting dominate experimental setups, while real industrial vehicles operate in narrow aisles with varying occlusion patterns and sensor configurations.

Collectively, these patterns reveal a substantial gap between laboratory-scale validation and operational warehouse conditions.

Missing or underexplored research directions

Several cross-domain gaps emerge, particularly relevant for HMLV intralogistics.

First, there is a lack of representative HMLV warehouse datasets capturing mixed SKUs, dense shelving, partial occlusion, dynamic interactions, and variable lighting. Existing warehouse datasets are small, proprietary, or task-specific (Wudhikarn et al., 2022; de Moura Junior et al., 2024).

Second, while multi-modal fusion is well-studied in autonomous driving, analogous integration of vision with warehouse-relevant sensing modalities (e.g., RFID, weight sensors) is rarely systematically evaluated (Agostinho et al., 2022; Albayrak Ünal et al., 2023).

Third, open-vocabulary and unsupervised localization research has not been explicitly extended to intralogistics domains (Yu et al., 2025; Papadopoulos et al., 2025). This omission is significant in HMLV settings, where exhaustive manual labeling is impractical.

Fourth, small-object detection paradigms developed for maritime or remote sensing contexts have not been systematically adapted to warehouse-scale small artifacts such as labels, compact Data Matrix codes, or damaged packaging under mixed lighting and occlusion (Rekavandi et al., 2025).

Fifth, digital twin–computer vision integration remains largely conceptual or laboratory-based, with limited quantitative validation in large-scale operational logistics systems (Faqeer and Khajavi, 2025).

Overall, the reviewed literature demonstrates substantial methodological progress in object detection, localization, pose estimation, and scene understanding. However, each survey focuses on a specific subdomain—such as SLAM, robotic grasping, small-object detection, barcode recognition, or digital twin integration—without addressing their intersection in high-mix, real-world warehouse environments.

No single review simultaneously covers:

1. Data Matrix or visual code recognition under industrial constraints,
2. Deep learning-based detection tailored to small and partially occluded objects,
3. Warehouse-specific environmental variability, and
4. Integration into operational intralogistics workflows.

The approximately thirty reviewed survey papers collectively touch upon these components, yet only partially and from distinct disciplinary perspectives. Their fragmented coverage highlights the absence of an integrated synthesis tailored to HMLV intralogistics and visual product identification.

For this reason, a dedicated systematic literature review is required. The following section presents the search strategy, inclusion logic, and structured analysis conducted to bridge these domains and establish a consolidated knowledge base directly aligned with the research objectives of this dissertation.

2.2.2 Overview of relevant research papers

This subsection presents the outcome of the systematic search and screening process applied to peer reviewed research articles directly related to the intersection of computer vision, visual code recognition, and high mix low volume warehouse environments.

As introduced in the search strategy section, the literature exploration was structured along four conceptual dimensions:

1. Computer Vision as the enabling technology,
2. Data Matrix or QR code recognition as the technical task,

3. Intralogistics or warehouse processes as the application domain,
4. High mix low volume environments as the operational context.

The combination of all four dimensions proved to be highly restrictive. In fact, only three partial combinations of these dimensions yielded research papers with tangible overlap to the dissertation topic.

Specifically, the following query combinations produced relevant results:

- Computer Vision + QR or Data Matrix + domain context resulted in 24 papers.
- Computer Vision + HMLV resulted in 39 paper.
- Computer Vision + HMLV + domain resulted in 1 papers.

Other combinations of the four dimensions did not produce directly relevant research outputs.

After applying the predefined inclusion and exclusion criteria, and conducting a structured selection and screening process including title and abstract, 18 research articles remained that demonstrated substantial methodological or application level relevance to the present study.

In addition, during the review and survey oriented search phase, several papers initially classified among review results were identified as original research contributions. These papers were methodologically aligned with the core research questions and were therefore reclassified and included in the research paper corpus.

As a result, the final set consists of 28 research articles forming the analytical basis of this subsection. Table 2 summarizes the search outcomes and filtering process.

Query	Query Combination	Initial hits	After screening
6	CV + QR + Domain	24	12
8	CV + HMLV + Domain	1	0
12	CV + HMLV	39	6
Total after screening			18
Additional resarch papers from review search			10
Final research corpus			28

Table 2: Search Results and Selection Process for Research Papers

The number of publications simultaneously addressing computer vision, visual code recognition, and high mix low volume intralogistics environments remains limited. Most identified studies intersect only two or three of the four conceptual dimensions, typically

focusing either on visual code detection in controlled industrial settings, or on broader warehouse automation without explicit emphasis on HMLV variability.

This distribution reflects the interdisciplinary and still evolving nature of the field. Rather than forming a consolidated research stream, the relevant contributions are dispersed across industrial inspection, warehouse automation, robotics, and applied computer vision domains.

The following sections synthesize these 28 research contributions by identifying recurring methodological patterns, architectural choices, dataset characteristics, and evaluation strategies. Particular attention is given to aspects relevant for robust Data Matrix detection and validation under dynamic and heterogeneous HMLV warehouse conditions.

Across the reviewed work, several consistent architectural patterns emerge in the application of deep learning for industrial and logistics-related visual perception tasks, as summarized in Table 3. Object detection constitutes the dominant paradigm, particularly for codes, markers, components, and small industrial features, where spatial localization is a prerequisite for downstream decoding, inspection, or control.

Convolutional neural networks, and especially modern one-stage detectors from the YOLO family, represent the most frequently adopted model class. YOLO variants are employed across diverse applications, including barcode and QR detection for UAV-based warehouse stocktaking, QR decoding under challenging lighting conditions, HMLV assembly hole localization, perspective-aware QR detection, depalletizing support, and tag-less stock counting with drones (Kalinov et al., 2020; Almeida et al., 2021; Edula et al., 2023; Fonseca et al., 2024; Simeth et al., 2025; Töper et al., 2025; Antoniou et al., 2025; Corpuz et al., 2025). Their prevalence reflects a consistent prioritization of real-time inference and deployment feasibility in operational environments.

Two-stage detectors, most notably Faster R-CNN, remain relevant in scenarios where small-object accuracy or precise localization is emphasized over latency. Representative examples include Data Matrix landmark detection for AGV localization and burr detection in metallic components (Almeida et al., 2020; Kyosuke et al., 2022; Almeida et al., 2021). In settings where pixel-level delineation is required, instance segmentation architectures such as Mask R-CNN are adopted, for example in wrinkle-like defect segmentation or tree crown extraction from UAV LiDAR data (Manyar et al., 2023; Jarahizadeh and Salehi, 2026).

Complementary architectural choices further illustrate adaptation to deployment constraints. Lightweight backbones, including MobileNet and separable convolution-based variants, are selected in embedded or resource-constrained scenarios such as automated optical inspection for PCB marks or barcode restoration on edge devices (Chang et al., 2023; Muallim et al., 2025). Backbone architectures such as ResNet and DenseNet are frequently integrated within detection or segmentation frameworks to enhance feature ex-

traction capacity while balancing computational cost.

Table 3: Deep Learning Model Families and Representative Papers in the Reviewed Corpus

Family	Models	Count	Publications
One-stage detectors	YOLO variants	12	Kalinov et al. (2020); Almeida et al. (2020); Edula et al. (2023); Daasan and Ishak (2023); Fonseca et al. (2024); Vivek et al. (2024); Simeth et al. (2025); Corpuz et al. (2025); Muallim et al. (2025); Töper et al. (2025); Antoniou et al. (2025); Jarahizadeh and Salehi (2026)
Two-stage detectors	Faster R-CNN	6	Almeida et al. (2020, 2021); Kyosuke et al. (2022); Matuszczyk and Weichert (2023); Simeth et al. (2025); Jarahizadeh and Salehi (2026)
Instance segmentation	Mask R-CNN	4	Manyar et al. (2023); Xie et al. (2023); Simeth et al. (2025); Jarahizadeh and Salehi (2026)
Segmentation networks	U-Net, DeepLab, PSPNet	5	Yoon et al. (2023); Kalinov et al. (2020); Muallim et al. (2025); Töper et al. (2025); Jarahizadeh and Salehi (2026)
Backbone networks	ResNet	8	Almeida et al. (2020, 2021); Kyosuke et al. (2022); Edula et al. (2023); Bauer and Daub (2023); Manyar et al. (2023); Matuszczyk and Weichert (2023); Yoon et al. (2023)
Lightweight backbones	MobileNet	2	Almeida et al. (2021); Chang et al. (2023)
Dense connectivity models	DenseNet	2	Almeida et al. (2021); Manyar et al. (2023)
3D point-based models	PointNet	2	Xie et al. (2023); Jarahizadeh and Salehi (2026)
Sequence models	RNN	3	Kyosuke et al. (2022); Xie et al. (2023); Muallim et al. (2025)
Unspecified CNN-based	CNN (generic)	2	Asare-Manu et al. (2023); Vijayaraj et al. (2025)
Unspecified	Other or non-CNN-based approaches	6	Deniša et al. (2023); Chen and Huang (2024); Al-hamadani et al. (2025); Bengueddoudj et al. (2025); Gunawan et al. (2023); Pohudina et al. (2024)

Typical processing pipelines across the reviewed studies exhibit a broadly consistent multi-stage structure. In code and marker-related tasks, a deep learning-based localization module is first applied to identify candidate regions of interest. These modules most frequently rely on YOLO or Faster R-CNN detectors, although U-Net-like architectures are also used where segmentation-based region extraction is preferred (Almeida et al., 2020, 2021; Edula et al., 2023; Matuszczyk and Weichert, 2023; Corpuz et al., 2025; Muallim et al., 2025). Non-maximum suppression or similar post-processing steps are routinely employed to refine overlapping detections.

Following localization, a second stage performs decoding, restoration, or domain-specific interpretation. In code recognition tasks, classical decoding libraries such as libdmtx, ZBar, Dynamsoft Barcode Reader, or Pyzbar are commonly retained. Deep learning components are typically positioned upstream of these decoders, for example by cropping detected regions, performing geometric rectification, or restoring partially damaged symbols before invoking a deterministic decoder (Almeida et al., 2021; Matuszczyk and Weichert, 2023; Corpuz et al., 2025; Muallim et al., 2025).

In industrial inspection settings beyond code reading, similar structural patterns are observed. Domain-specific pre-processing techniques—such as polar coordinate transformations for cylindrical burr detection, physics-based wrinkle simulation, or multiband rasterization of LiDAR point clouds—precede detection or segmentation modules (Kyosuke et al., 2022; Manyar et al., 2023; Jarahizadeh and Salehi, 2026). Downstream stages may include rule-based decision logic, quality classification, or integration with motion control systems.

Across these applications, fully end-to-end neural architectures that directly map raw images to decoded symbols or operational decisions remain uncommon. Instead, hybrid pipelines dominate, in which learned localization is coupled with classical decoding, geometric correction, or task-specific control logic. This separation between detection and symbolic interpretation reflects both industrial reliability requirements and the continued maturity of traditional decoding algorithms.

Evaluation strategies across the reviewed literature display a high degree of methodological consistency. For object detection tasks, performance is predominantly assessed using mean Average Precision (mAP) computed at predefined Intersection over Union (IoU) thresholds. Many studies adopt COCO-style evaluation protocols, reporting $AP@[0.50:0.95]$ and, in some cases, scale-specific metrics for small, medium, and large objects (Almeida et al., 2020, 2021; Edula et al., 2023; Matuszczyk and Weichert, 2023; Simeth et al., 2025; Antoniou et al., 2025; Töper et al., 2025; Corpuz et al., 2025; Jarahizadeh and Salehi, 2026). This reflects the transfer of benchmark evaluation practices from general computer vision research into applied industrial contexts.

Complementary metrics such as precision, recall, F1-score, and IoU distributions are fre-

quently reported, particularly in applications where spatial localization accuracy directly affects downstream processes. Examples include hole localization in HMLV assembly, burr region detection in metal parts, and tree crown delineation from UAV LiDAR data (Simeth et al., 2025; Jarahizadeh and Salehi, 2026). These metrics provide additional insight into detection reliability beyond aggregate AP values.

In application-driven studies, task-level or operational metrics are also introduced to capture system-level performance. For code recognition tasks, decoding success rate and symbol restoration accuracy are commonly reported. In warehouse and UAV-based scenarios, inventory recognition percentage, flight time, stock-count accuracy, or navigation performance are included as complementary indicators of practical feasibility (Kalinov et al., 2020; Edula et al., 2023; Manyar et al., 2023; Bauer and Daub, 2023; Chen and Huang, 2024; Muallim et al., 2025; Antoniou et al., 2025). Such metrics highlight the applied orientation of many contributions, where perception quality is evaluated in relation to operational constraints.

Dataset characteristics vary substantially across the reviewed studies, yet several recurring patterns are observable. Many industrial vision datasets are relatively small and task-specific, often comprising only a few hundred images. This reflects the practical difficulty, cost, and operational constraints associated with data collection in production environments. Examples include Data Matrix datasets containing 156–224 frames per subset, burr inspection datasets with a few hundred samples per domain, SLS Data Matrix datasets with 204 images, and QR datasets ranging from 477 to 7,308 images depending on acquisition conditions (Almeida et al., 2020, 2021; Kyosuke et al., 2022; Matuszczyk and Weichert, 2023; Edula et al., 2023).

To compensate for limited real-world data, several works incorporate synthetic data generation or simulation-based augmentation. Synthetic wrinkle datasets with 10,000 generated samples, simulated pallet scenes, and multiscene warehouse simulations are used to increase variability and reduce annotation cost (Manyar et al., 2023; Töper et al., 2025). While such strategies expand dataset scale, they introduce considerations regarding domain transfer between simulated and real environments.

Public availability of datasets and code is inconsistent. Some contributions explicitly release benchmarks, detection models, or simulation tools (Almeida et al., 2021; Edula et al., 2023; Deniša et al., 2023; Manyar et al., 2023). In other cases, datasets remain proprietary, partially shared, or internally restricted due to industrial confidentiality (Chang et al., 2023; Muallim et al., 2025). Several studies report the absence of prior standardized datasets for their specific task, necessitating custom data acquisition and annotation (Kyosuke et al., 2022; Matuszczyk and Weichert, 2023).

Annotation procedures are predominantly manual and supervised, typically relying on bounding boxes or polygonal masks defined by domain experts. In tasks such as burr de-

tection, powder-versus-part segmentation, or tree crown extraction, annotation cost is explicitly acknowledged as a limiting factor (Kyosuke et al., 2022; Asare-Manu et al., 2023; Jarahizadeh and Salehi, 2026). To mitigate this burden, some works explore synthetic label generation, cross-domain annotation transfer, or human-in-the-loop refinement strategies (Deniša et al., 2023; Manyar et al., 2023).

Environmental representativeness varies considerably across the reviewed studies. Several datasets are collected in operational industrial or warehouse settings, including low-light warehouses for UAV-based barcode scanning, SMT production lines, friction stir welding environments, and additive manufacturing chambers (Kalinov et al., 2020; Chang et al., 2023; Bauer and Daub, 2023; Asare-Manu et al., 2023). These works emphasize real-world constraints such as illumination variability, limited access, and production-driven timing.

In contrast, other contributions rely on laboratory approximations or simulation environments, including mock warehouse setups, controlled assembly benches, and simulated industrial scenes in Gazebo or custom frameworks (Deniša et al., 2023; Antoniou et al., 2025). Synthetic environments are frequently employed for bin-picking, depalletizing, and localization studies, often combined with domain randomization techniques to approximate variability (Töper et al., 2025; Pohudina et al., 2024). While such approaches enable scalable experimentation, they also introduce challenges related to transferability between simulated and operational environments.

Evaluation settings are frequently static or semi-controlled, even in studies targeting dynamic warehouse or UAV scenarios. Several barcode and QR detection systems are assessed using static images or offline video processing, with limited reporting of closed-loop flight experiments or continuous deployment conditions (Kalinov et al., 2020; Chen and Huang, 2024). Similarly, depalletizing and bin-picking models are often evaluated on predefined scene configurations where occlusion and motion variability are partially represented (Töper et al., 2025; Deniša et al., 2023).

Assumptions regarding scene structure and sensor placement are also common. Many approaches rely on planar markers, fixed camera viewpoints, constrained trajectories, or optimized sensor positioning, such as overhead conveyor cameras, floor-mounted QR codes for localization, or static LiDAR scans (Yoon et al., 2023; Matuszczyk and Weichert, 2023; Muallim et al., 2025). In several cases, occlusion, clutter, and viewpoint variation are either limited or treated as secondary factors (Bengueddoudj et al., 2025; Jarahizadeh and Salehi, 2026).

Dynamic environmental factors—including illumination changes, motion blur, and noise—are explicitly considered in some works, particularly in UAV-based warehouse scanning and HMLV assembly inspection (Kalinov et al., 2020; Ecula et al., 2023; Simeth et al., 2025). However, long-term deployment studies under continuously changing prod-

uct mixes and layout configurations are less frequently reported. Similarly, explicit mechanisms for online adaptation or continual learning are rarely integrated into warehouse-focused perception systems, with the exception of selected industrial monitoring tasks such as friction stir welding (Bauer and Daub, 2023).

With respect to HMLV intralogistics specifically, only a limited subset of the reviewed works directly address environments characterized by high product variability, heterogeneous stock-keeping units, and dynamically changing layouts. While YOLO-based hole detection for HMLV assembly and drone-based stock counting are reported, these applications typically focus on constrained product families or controlled test environments (Simeth et al., 2025; Antoniou et al., 2025). Broader variability in packaging types, stacking configurations, and ad hoc storage conditions is less frequently explored within a unified experimental framework.

Continual and adaptive learning mechanisms are also sparsely represented in warehouse and intralogistics perception systems. Although industrial monitoring studies such as friction stir welding explicitly address concept drift and continuous calibration (Bauer and Daub, 2023), comparable adaptation strategies are not commonly integrated into visual code recognition or warehouse perception pipelines. Most reviewed approaches rely on static training regimes with evaluation performed on fixed datasets.

Integration between perception and higher-level planning or human interaction is demonstrated in selected contributions, including UAV path optimization driven by visual detection, AGV navigation guided by markers, and human-in-the-loop segmentation interfaces (Kalinov et al., 2020; Xie et al., 2023; Deniša et al., 2023). However, comprehensive end-to-end systems combining robust perception, dynamic task allocation, and operational-scale HMLV deployment remain relatively rare within the surveyed literature.

Reporting practices further influence interpretability and reproducibility. In several publications, details regarding dataset availability, preprocessing configurations, postprocessing logic, or deployment hardware are only partially specified (Chang et al., 2023; Fonseca et al., 2024; Simeth et al., 2025; Antoniou et al., 2025; Töper et al., 2025). While such omissions are common in applied industrial research, they limit cross-domain comparison and systematic evaluation.

In summary, the reviewed literature demonstrates substantial progress in applying deep learning-based detection and segmentation to industrial and warehouse perception tasks. Recurring methodological patterns include the dominance of YOLO-like detectors, COCO-style evaluation protocols, hybrid detection–decoding pipelines, and frequent reliance on synthetic data to supplement limited real-world datasets. At the same time, recurring constraints are observable, including restricted environmental variability, predominantly static evaluation settings, and limited integration of adaptive learning or full-scale HMLV deployment scenarios.

These synthesized observations provide the analytical basis for formulating the specific research gaps and contributions addressed in the following subsection.

2.3 Research gap

Within the reviewed code-recognition literature, hybrid architectures are the dominant pattern: learned localization is followed by dedicated decoding, geometric correction, restoration, or rule-based validation. Classical geometric readers, learned recovery modules, and one-stage or two-stage detectors therefore provide relevant architectural context for the present system. Fully end-to-end neural systems that map a raw industrial image directly to a validated identifier or operational decision are less common in the reviewed corpus. Transformer-based perception architectures are valid future alternatives, but they were not represented as directly comparable Data Matrix identification baselines in the selected literature set.

The systematic review indicates that, although deep learning-based object detection for industrial perception has advanced considerably, the specific intersection of visual code recognition and HMLV intralogistics remains only partially explored. Existing works typically address code detection under controlled industrial conditions, navigation in structured warehouse environments, or inspection tasks with constrained variability. Fully integrated approaches that explicitly target dynamic, heterogeneous HMLV warehouse settings are comparatively limited.

Table 4 summarizes the main limitations identified across the reviewed literature, the requirements they imply, and the specific components of the proposed framework (Chapter 6) that address them.

These observations suggest that the primary limitation of existing approaches is not the absence of effective detection or decoding methods, but the lack of an integrated framework that combines these components with validation, recovery, and temporal reasoning mechanisms; in particular, the interaction between environmental variability, decoding uncertainty, and system-level identification reliability remains insufficiently characterized in the literature.

Several recurring patterns highlight concrete research opportunities. First, industrial code recognition systems predominantly rely on hybrid pipelines, combining deep learning-based localization with classical decoding libraries and domain-specific post-processing. While such architectures demonstrate robustness, they often lack explicit validation mechanisms that address decoding ambiguities under occlusion, distortion, or visual degradation.

Limitation in existing literature	Implied requirement	Addressed by
Detection, decoding, and perception treated as separate problems, mostly evaluated under controlled or semi-structured conditions	A unified architecture linking and jointly evaluating perception stages under variable acquisition conditions	End-to-end pipeline integrating detection, geometric normalization, decoding, and validation
No explicit mechanism for handling uncertainty from occlusion, distortion, or incomplete observations	A dedicated pathway to recover identification when primary decoding degrades or fails	Recovery stage: geometric reconstruction and similarity-based matching
Evaluation predominantly performed at the frame level; identification reliability across multiple observations not considered	Reasoning over temporally extended observations rather than single frames	Sequence-level reasoning: track association, evidence accumulation, and the dominance decision rule
Lack of an integrated framework combining validation, recovery, and temporal reasoning; temporal evidence accumulation not systematically addressed	A system-level architecture in which validation, recovery, and temporal aggregation operate jointly rather than in isolation	Full proposed pipeline, explicitly separating frame-level perception from sequence-level decision-making
Datasets rarely reflect HMLV variability and heterogeneity; standardized benchmarks for such conditions are largely absent	A dataset design with systematically controlled, HMLV-representative variability to support stratified evaluation	Macro-scene-based dataset design with controlled illumination, viewpoint, occlusion, and blur variability

Table 4: Mapping from limitations identified in the reviewed literature to the requirements they imply and the framework components that address them.

Second, evaluation protocols are frequently conducted in static or semi-controlled environments. Even when dynamic applications such as UAV-based scanning or warehouse stocktaking are targeted, long-term deployment under continuously changing product mixes, layouts, and viewpoints is rarely documented. Consequently, the robustness of visual code recognition systems under realistic HMLV variability remains insufficiently

characterized.

Third, dataset scale and heterogeneity are commonly constrained. Many industrial datasets are relatively small and task-specific, while unified benchmarks representing variability across stock keeping units (SKUs) in dynamic warehouse contexts are uncommon. This limits systematic evaluation of perception performance across diverse packaging types, stacking patterns, lighting conditions, and sensor viewpoints typical of HMLV intralogistics.

These observations collectively motivate the research direction of the present dissertation. Rather than focusing solely on incremental improvements in detection accuracy, the study addresses the broader challenge of designing and evaluating a perception framework tailored to dynamic HMLV warehouse environments. The precise formulation of the research problem and its operational objectives are detailed in the following section.

3. RESEARCH PROBLEM, OBJECTIVES AND QUESTIONS

The preceding related work analysis has systematically reviewed existing research at the intersection of computer vision, visual code recognition, and intralogistics, with particular attention to their applicability in HMLV environments. This analysis has shown that, while substantial progress has been achieved in vision-based perception and code recognition under controlled or structured conditions, the integration of these methods into realistic, dynamic warehouse operations remains limited. In particular, the literature reveals a fragmentation between methodological advances in visual detection and decoding, and the operational requirements of HMLV intralogistics, where variability, uncertainty, and real-time constraints dominate everyday processes.

Building on the operational context introduced in Section 1, this section defines the research problem as a gap between the physical state of HMLV warehouses and the information available to identification and inventory systems. Dynamic placement and incomplete location updates may require operators to search manually for products during picking, replenishment, or internal transport. The problem is therefore system-level rather than an isolated limitation of a single detection or decoding algorithm.

From a lean intralogistics perspective, time spent manually locating items does not create direct customer value. In the industrial case motivating this dissertation, such search activity was reported to recur multiple times during a shift, and manual loading and unloading processes were observed to introduce additional variability into item placement, further reducing the reliability of recorded inventory positions. These observations are treated as case-specific motivation rather than as a general claim about the prevalence of search-related waste across HMLV warehouses.

WMS platforms are designed to provide a digital representation of inventory states and product locations. Because location updates typically rely on manual scanning or operator input, this information may be delayed, omitted, or performed under time pressure, so the digital system state can diverge from the actual warehouse state. In the industrial case motivating this dissertation, such discrepancies were observed to increase search times and reduce trust in system outputs. This discrepancy between recorded and physical inventory states limits the effectiveness of data-driven decision support and constrains the degree of achievable automation.

A key characteristic of HMLV warehouses is that products are already labeled with unique machine-readable identifiers, such as Data Matrix codes, which are primarily used for transactional purposes. These identifiers are present in the physical environment but are not continuously exploited as sources of real-time information. Existing workflows typically treat code scanning as a discrete event rather than as part of a continuous perception process. As a result, opportunities to leverage visual information for dynamic product

localization and real-time data updates remain largely unexploited.

In this dissertation, a Data Matrix symbol is treated as carrying a serialized item-level identifier attached directly to a single physical inventory item. The intended objects are individually stored and handled products or components, with spare parts representing a typical HMLV use case. The encoded value distinguishes one physical item instance from other instances, rather than identifying only a product category or stock keeping unit (SKU). The scope therefore excludes identifiers assigned to aggregate handling units, such as pallets, totes, bins, or containers holding multiple items. The identifier also does not determine the item's storage location; associating an identified item with a warehouse position would require an additional spatial-inference and warehouse-management integration layer outside the implemented framework.

A representative scenario illustrates how these elements connect. Consider a spare part stored as an individually identified item and labeled with an item-level Data Matrix code. If its recorded WMS location is missing, imprecise, or inconsistent with its physical position, an operator may need to search for it manually across the affected storage area. A mobile camera carried or mounted during such a search, or during routine internal transport, observes the part across multiple frames as it comes into and out of view under varying distance, angle, and illumination. The proposed system aggregates this evidence across observations and returns one of three outcomes for the item: an accepted identifier, an ambiguous outcome, or an unresolved outcome. An accepted identifier confirms that the specific item was observed and could inform downstream inventory reasoning; it does not, by itself, estimate the item's spatial coordinates or trigger an automatic update to the WMS record. This scenario is used throughout the dissertation as a conceptual illustration of the research problem rather than as a documented case study.

From an intralogistics perspective, this creates a clear unmet need for a system that can support operators during product search while simultaneously contributing to the accuracy of warehouse data. Such a system should be able to identify products visually based on their two-dimensional codes, even under non-ideal conditions, and should allow newly observed information to contribute validated evidence to the digital system state. Importantly, this capability must function in unstructured and visually complex environments, without relying on fixed camera placements or controlled acquisition setups.

The technical challenge is to transform variable visual observations into validated Data Matrix identities. Detection and decoding errors may otherwise propagate into incorrect inventory information. The proposed framework therefore distinguishes detection performance from identification reliability and includes system-level validation and error-control mechanisms.

Based on the related work analysis conducted in this dissertation, existing research does not provide an integrated solution to this problem. Prior studies typically address visual-

code detection, dedicated code decoding, or warehouse perception in isolation, often under controlled or highly structured conditions. No prior work systematically investigates how deep learning-based visual detection, code decoding, and similarity-based validation can be combined into a single pipeline tailored to HMLV intralogistics environments, nor how such a pipeline can support both product search and real-time data consistency.

Accordingly, the research problem addressed in this dissertation can be formulated as follows. There is a lack of robust, empirically validated methods for visual product identification based on Data Matrix codes that operate reliably in dynamic HMLV warehouse environments and that can support product localization and inventory-related decision support. Addressing this problem requires the development of a vision-based identification pipeline that is resilient to environmental variability, aligned with operational constraints, and evaluated using realistic datasets that reflect the actual conditions encountered in HMLV intralogistics.

In response to the challenges highlighted above, the research objectives of this dissertation are defined to translate the identified operational problem into concrete, empirically investigable goals. These objectives focus on the development and evaluation of a vision-based identification pipeline that is technically robust, operationally meaningful, and explicitly tailored to the constraints of HMLV intralogistics environments.

The primary objective of this dissertation is to develop and evaluate a computer vision-based product identification pipeline that enables reliable Data Matrix recognition and supports product localization to inform inventory-related decision support in high-mix low-volume warehouse environments.

In order to achieve the primary objective and to structure the investigation of the identified research problem, the dissertation is guided by three research questions (RQ). Each question addresses a distinct but interconnected dimension of the proposed solution: system-level design, empirical validation, and operational integration.

RQ1: What system-level design enables robust visual product identification with integrated prior knowledge?

The first research question addresses the architectural dimension of the problem. Given that reliable visual product identification in HMLV environments requires the coordinated functioning of detection, decoding, and validation mechanisms, this question seeks to identify the structural principles that govern an effective system design. Particular emphasis is placed on the integration of prior knowledge in the form of a predefined Data Matrix reference list, which constrains and validates decoding outputs in order to mitigate identification errors. The investigation of this question is conceptual and design-oriented: it involves the definition of a modular pipeline architecture, the specification of data flows between components, and the formalization of validation logic that prevents uncontrolled error propagation. The outcome of RQ1 is a coherent system-level framework that aligns

technical robustness with operational requirements.

RQ2: How can the proposed system be implemented and empirically validated under controlled environmental variability?

The second research question focuses on implementation and empirical evaluation. Building on the architectural framework established in RQ1, it examines how the proposed pipeline can be realized using deep learning–based instance segmentation, dedicated Data Matrix decoding, reference validation, failure-activated recovery, and episode-level evidence aggregation, and how its performance can be assessed systematically. To ensure methodological rigor, the evaluation is conducted on a structured dataset that captures controlled environmental variability, including variations in illumination, viewing angle, distance, and the presence of negative samples. The investigation involves both component-level and end-to-end performance analysis, with particular attention to robustness across macro-scene-level data splits. The outcome of RQ2 is an empirically grounded assessment of the system’s reliability under realistic acquisition conditions.

RQ3: How can temporally validated visual product-identification outcomes be interpreted as evidence for inventory-related decision support in dynamic warehouse operations?

The third research question extends the technical investigation toward operational interpretation. While RQ1 and RQ2 address architectural design and empirical robustness, RQ3 examines how accepted, ambiguous, and unresolved episode-level identification outcomes may be interpreted as structured evidence for inventory-related decision support. The focus is conceptual and system-oriented rather than implementation-driven: it defines the information that the proposed system can provide to a downstream process and the boundary between such evidence and an actual inventory update. RQ3 therefore does not develop or evaluate a warehouse-management integration, spatial-localization method, or operational update policy. Instead, it establishes an analytical perspective on how uncertainty-aware visual identification may inform later inventory-related reasoning in HMLV-inspired intralogistics environments.

To achieve the primary objective and to provide a structured pathway for answering the research questions, the dissertation formulates a set of detailed operational research objectives. These objectives translate the conceptual aims of the study into concrete development, implementation, and evaluation steps. While more specific than the research questions, they remain aligned with the system-level perspective introduced above.

The research objectives are defined as follows:

1. **O1.** To conceptualize a modular system architecture for vision-based Data Matrix identification that integrates detection, decoding, and validation components.

2. **O2.** To incorporate structured prior knowledge in the form of a predefined Data Matrix reference list into the identification process through a similarity-based validation mechanism.
3. **O3.** To define architectural principles that ensure robustness against environmental variability and limit error propagation between pipeline components.
4. **O4.** To construct a structured dataset that captures controlled environmental variability, including variations in illumination, viewing angle, distance, and the presence of negative samples.
5. **O5.** To implement the proposed detection–decoding–validation pipeline using deep learning–based instance segmentation, dedicated Data Matrix decoding, reference validation, recovery, and episode-level evidence aggregation.
6. **O6.** To empirically evaluate both component-level and end-to-end system performance under macro-scene-level data separation.
7. **O7.** To formulate how accepted, ambiguous, and unresolved episode-level identification outcomes may be interpreted as uncertainty-aware evidence for inventory-related decision support.
8. **O8.** To assess conceptually the integration requirements and boundaries between the proposed identification outputs and warehouse-management processes.

These objectives are directly linked to the three research questions and collectively structure the empirical and conceptual contributions of the dissertation (Table 5). Objectives O1–O3 primarily address **RQ1**, as they focus on system-level architectural design and the integration of prior knowledge within a robust identification framework. Objectives O4–O6 correspond to **RQ2**, as they concern the implementation of the proposed system and its empirical validation under controlled environmental variability. Objectives O7–O8 relate to **RQ3**, as they define the conceptual interpretation of episode-level identification outcomes as evidence for inventory-related decision support and the boundary to warehouse-management integration. Together, these objectives ensure conceptual coherence between system design, empirical validation, and operational interpretation.

The dissertation makes three bounded contributions. First, it formulates a system architecture that separates perception, identification evidence, and episode-level decision-making while integrating reference validation and failure-activated recovery. Second, it provides a methodological evaluation design that decomposes component-level and end-to-end outcomes under recorded acquisition variability and macro-scene-level separation. Third, it offers an applied interpretation of the resulting decision states as potential inputs to inventory-related decision support in HMLV-inspired environments. These contributions

concern architectural integration, evaluation design, and operational interpretation; they do not claim a new recognition algorithm, a deployed warehouse-management system, or state-of-the-art detector performance.

Table 5: Mapping between research questions and operational objectives

Research Question	Formulation	Associated Objectives
RQ1	What system-level design enables robust visual product identification with integrated prior knowledge?	O1, O2, O3
RQ2	How can the proposed system be implemented and empirically validated under controlled environmental variability?	O4, O5, O6
RQ3	How can temporally validated visual product-identification outcomes be interpreted as evidence for inventory-related decision support in dynamic warehouse operations?	O7, O8

The three research questions are addressed through distinct but interconnected analytical stages within the dissertation.

RQ1 is answered through the conceptualization and formalization of the proposed system-level architecture. Its resolution relies on the theoretical grounding of HMLV intralogistics constraints and computer vision principles, as well as on the explicit definition of a modular pipeline that integrates detection, decoding, and similarity-based validation. The architectural framework and its underlying design principles provide the structural answer to how robust visual product identification with integrated prior knowledge can be realized.

RQ2 is addressed through the implementation and systematic empirical evaluation of the proposed pipeline. By constructing a structured dataset that captures controlled environmental variability and by assessing both component-level and end-to-end system performance, the dissertation provides quantitative evidence regarding robustness and reliability. The empirical findings establish the extent to which the architectural principles defined under RQ1 translate into measurable identification performance under realistic acquisition conditions.

RQ3 is examined through the conceptual interpretation described above, applied to episode-level outcomes as decision-support evidence.

Together, these analytical stages ensure that each research question is not only conceptually formulated but systematically examined through architectural design, empirical validation, and operational interpretation.

Table 6 summarizes how the research questions and operational objectives are addressed across the chapters of the dissertation.

Table 6: Mapping of research questions, objectives, and dissertation chapters

RQ	Associated Objectives	Chapter(s) Addressing Them
RQ1	O1–O3: System-level architecture design; integration of prior knowledge via similarity-based validation; definition of robustness and error-mitigation principles.	Chapter 5 (Theoretical framework); Chapter 6.1 (System-level architecture)
RQ2	O4–O6: Structured dataset construction; implementation of the detection–decoding–validation pipeline; empirical component-level and end-to-end evaluation.	Chapter 6 (Model and System Design)
RQ3	O7–O8: Conceptual interpretation of episode-level outcomes as inventory-related decision evidence; analysis of integration requirements and boundaries.	Chapter 8 (Discussion); Chapter 9 (Conclusions)

4. METHODOLOGY

The methodological framework of this research is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology (Chapman et al., 1999), which provides a structured and systematic approach for data-driven projects across different domains. CRISP-DM covers the complete lifecycle of a data science project and defines a set of well-established phases, tasks, and deliverables that guide the development, evaluation, and application of machine learning models.

Although the research involves the construction of a novel computer vision–based identification pipeline and a purpose-built dataset, the study is positioned primarily as an empirical, data-driven investigation. Therefore, the CRISP-DM methodology is adopted as the main research framework. The work shares certain characteristics with design science research, particularly in the creation and evaluation of methodological artefacts; however, design science is not used as the primary methodological paradigm.

CRISP-DM was originally developed for industrial data mining projects, where the primary objective is to support business decision-making and deployment. In this disser-

tation, the framework is adapted to the context of scientific research, where the focus is on methodological rigor, reproducibility, and the generation of new knowledge. The six phases of CRISP-DM are retained, but their interpretation and outputs are aligned with research objectives rather than operational business goals. The phases are not strictly linear; instead, iterative feedback between them is possible, reflecting the exploratory and experimental nature of applied machine learning research.

The research process is aligned with the six CRISP-DM phases as follows.

Table 7: Alignment of the CRISP-DM methodology with the dissertation structure

CRISP-DM phase	Study-specific decisions and outputs	Corresponding dissertation sections
1. Business Understanding	Research boundary, research questions, and design principles: modularity, controlled variability, reference validation, and frame–episode separation	1; 2; 3; 5
2. Data Understanding	Pre-specified acquisition factors: illumination, viewpoint, blur, apparent size, partial visibility, multiple symbols, and negative frames	6.2
3. Data Preparation	Frame and episode manifests, polygon annotations and metadata, and macro-scene-level code-disjoint separation	6.2
4. Modeling	YOLOv8 instance segmentation, geometric normalization, dedicated decoding, reference validation, recovery, soft-Hamming ranking, and evidence aggregation	6.1; 6.6; 6.7
5. Experimental Design and Evaluation	Held-out test protocol, ground-truth target, reporting units, component ablations, and condition-specific assessment	7.1
6. Deployment	Documentation and bounded operational interpretation; no WMS integration, inventory update, or production deployment	8

In this dissertation, the CRISP-DM phases are used as a traceability structure for concrete research decisions rather than as a generic description of a data-mining lifecycle. Table 7 maps each phase to the decisions, artefacts, and evaluation boundaries of the study.

Business Understanding defined the research boundary as visual Data Matrix identification under variable mobile-camera acquisition, rather than warehouse-management deployment or inventory-state updating. The systematic review, research questions, and objectives established modularity, controlled variability, reference validation, and the separation of frame-level perception from episode-level identification as design principles.

Data Understanding identified the acquisition factors to be represented in the data: illumination, viewpoint, blur, apparent symbol size, partial visibility, multiple visible symbols, and negative frames (Section 6.2).

Data Preparation converted the recordings into annotated frame and episode manifests. It defined polygon-based object annotations, the *normal*, *too small*, and *occluded* object-condition categories, the independent blur flag, and macro-scene-level train–validation–test separation. The split prevents the same physical code layout and identifier set from appearing across partitions.

Modeling implemented a modular rather than end-to-end system: YOLOv8 instance segmentation, mask-based geometric normalization, dedicated Data Matrix decoding, reference validation, failure-activated recovery, soft-Hamming candidate ranking, and frame- and episode-level evidence aggregation. The model-design phase therefore included both learned perception and explicit decision logic.

Experimental Design and Evaluation fixed the held-out test split, the ground-truth target, and the reporting units before interpreting the results. Detection, decoding, frame-level identification, and episode-level identification were evaluated separately, with component-level ablations and condition-specific reporting used to locate losses within the pipeline.

Deployment is interpreted as research dissemination and operational interpretation rather than system rollout. The study documents the dataset, pipeline, decision rules, results, and limitations, and it explains how accepted, ambiguous, and unresolved episode outcomes could serve as bounded evidence for later inventory-related decision support. No WMS integration, inventory update, or production deployment was implemented.

5. THEORETICAL FRAMEWORK

5.1 HMLV intralogistics context

Warehouses constitute a central operational node within modern supply chains, serving as interfaces between production, distribution, and customer fulfillment. From a systems perspective, warehouses can be characterized along multiple dimensions, including throughput volume, product variety, automation level, and information integration. Recent literature emphasizes that contemporary warehouses increasingly deviate from traditional high-volume, standardized configurations, driven by mass customization, shorter product life cycles, and volatile demand patterns (Tubis and Rohman, 2023; Zhen and Li, 2022). Within this broader landscape, intralogistics refers specifically to the organization, control, and optimization of internal material and information flows inside a facility. Intralogistics encompasses inbound handling, internal transportation, storage, order picking, and outbound preparation, and acts as a connective layer between physical operations and digital control systems (Winkler and Zinsmeister, 2019; Płaczek and Osieczko-Potoczna, 2024). Unlike external logistics, which focuses on inter-facility transportation and distribution, intralogistics operates under tighter spatial constraints and stronger coupling between physical layout, human activity, and information systems (Szeszák et al., 2025). HMLV warehouse environments represent a particularly demanding subclass of intralogistics systems. HMLV settings are characterized by a large number of distinct SKUs, low throughput per item, and limited repetition of handling operations. While the HMLV concept is frequently discussed in manufacturing, recent studies highlight that its implications extend directly to warehousing and internal logistics, especially in sectors such as spare parts logistics, customized assembly, and e-commerce fulfillment (Gan et al., 2023).

From an intralogistics standpoint, HMLV environments fundamentally alter the nature of operational decision making. Storage strategies tend to be less standardized, item locations may change frequently, and picking paths are harder to optimize due to low predictability and limited historical repetition. Classical warehouse optimization approaches, which assume stable demand patterns and homogeneous handling units, therefore exhibit reduced effectiveness in HMLV contexts (Azadeh et al., 2019).

Recent research further indicates that automation strategies must be carefully adapted to the mix–volume profile of the warehouse. Highly automated systems such as shuttle-based storage or fully robotic picking solutions are typically designed for high-throughput, low-variability environments. In contrast, HMLV warehouses benefit more from flexible automation concepts, including mobile robots, autonomous guided vehicles, and human–machine collaboration, which can accommodate variability without excessive reconfiguration costs (Ellithy et al., 2024).

A recurring challenge identified in HMLV intralogistics is limited real-time visibility of

individual items. The high number of unique identifiers and dynamic placement of products complicate inventory tracking, picking verification, and process monitoring. Several studies emphasize that insufficient visibility leads to increased error rates, longer picking times, and higher reliance on manual verification steps (Lee et al., 2018; Kumar et al., 2022).

Digital technologies associated with Industry 4.0 have been proposed as enablers to address these challenges. IoT-based warehouse management systems, sensor networks, and cyber-physical integration aim to improve transparency and responsiveness within intralogistics processes. However, existing implementations often assume structured environments or rely on fixed identification infrastructure such as scanning stations or RFID gates (Zhen and Li, 2022). These assumptions are frequently violated in HMLV settings, where products may be handled by mobile agents and observed from varying viewpoints.

Recent intralogistics research increasingly recognizes the need for perception-based solutions that can operate under such variability. Instead of enforcing rigid structural order, perception-driven approaches seek to infer state information directly from the physical environment. Visual identification technologies, in particular, have been identified as promising tools for enabling flexible, context-aware intralogistics operations without requiring extensive infrastructure modifications (Tubis and Rohman, 2023).

In HMLV warehouses, the role of identification extends beyond mere labeling. Unique identifiers serve as anchors that connect physical objects to digital records, enabling traceability, localization, and process control. Errors or delays in identification therefore propagate through multiple intralogistics functions. As highlighted by Gan et al. (Gan et al., 2023), reliable product identification and traceability are foundational requirements for managing high product variety and low production or handling volumes.

In summary, HMLV intralogistics environments are defined by high variability, limited repetition, and strong coupling between physical handling and information systems. These characteristics challenge traditional intralogistics solutions while amplifying the importance of flexible automation and robust identification mechanisms. Establishing this context provides a necessary theoretical foundation for introducing perception-based identification approaches in subsequent sections, where computer vision and learning-based methods are examined as enabling technologies for reliable operation under HMLV conditions.

5.2 Visual detection, instance segmentation, and structured identification

This section retains only the concepts needed to interpret the proposed hybrid identification pipeline. It distinguishes object detection from instance-level boundary estimation, introduces the representation used for reference comparison, and explains why evidence

from repeated observations requires an explicit acceptance decision. General introductions to artificial intelligence, elementary neural-network operations, and text recognition are not part of the final theoretical argument.

Computer vision transforms visual measurements into structured representations that support perception and decision making. In the proposed hybrid pipeline, learned perception supplies the location and visible boundary of each Data Matrix symbol. Geometric normalization, dedicated decoding, reference comparison, and temporal decision making remain explicit stages rather than an end-to-end learned predictor.

The distinction retained for this dissertation is between object detection and instance segmentation: the former returns a class label and an approximate rectangular location, whereas the latter estimates a separate pixel mask for each object instance.

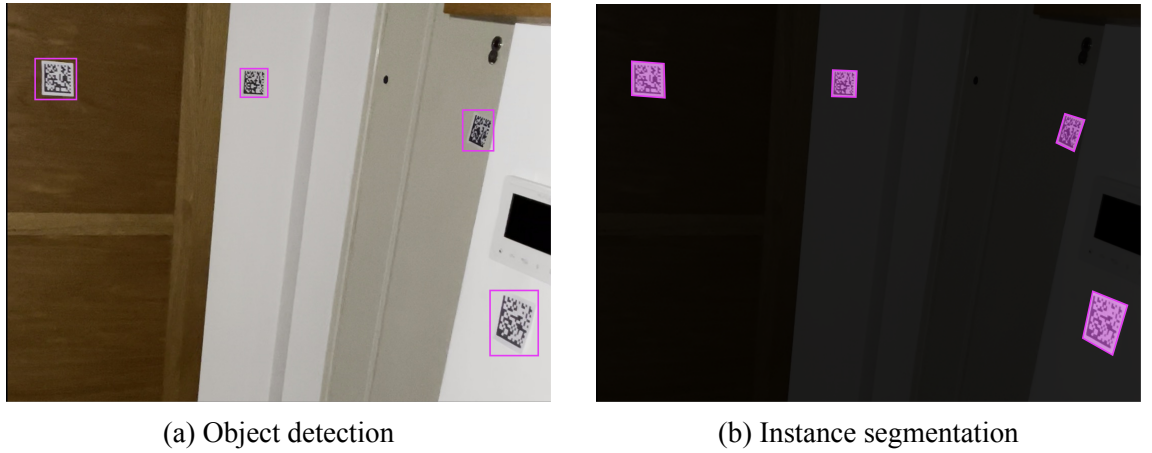


Figure 1: Conceptual distinction between object detection, which localizes an object with a bounding box, and instance segmentation, which estimates an object-specific pixel mask. Source: own elaboration.

Object detection answers whether and approximately where an instance is present, conventionally reporting a rectangular bounding box and a class label. A bounding box is adequate when a downstream task needs only coarse localization. Figures 1a and 1b contrast this representation with instance segmentation, which assigns pixels to each separate instance and thereby retains a closer approximation of the visible boundary. The combined comparison in Figure 1 illustrates why the mask preserves more precise object geometry for the subsequent rectification step. The design rationale for selecting a segmentation model, however, is given with the implemented system in Chapter 6, not inferred from this conceptual distinction alone.

For the present study, the relevant learning-theoretic point is limited to this: convolutional feature hierarchies can localize visual instances across changing scale and appearance, and an instance-specific mask preserves spatial boundary information. The selected detector is treated as the perception component of a hybrid pipeline; the rationale for selecting its segmentation variant and the associated training configuration are given in Chapter 6.

5.3 Data Matrix

Two-dimensional (2D) barcodes have largely replaced traditional one-dimensional (1D) barcodes in many industrial and logistics applications due to their significantly higher data capacity and built-in error correction mechanisms (Karrach and Pivarčiová, 2021). While linear barcodes encode information along a single dimension, matrix codes such as Data Matrix and QR codes store information in both horizontal and vertical directions, allowing substantially more data to be represented within a compact physical area.

Among 2D codes, the Data Matrix symbology is widely used in industrial environments where space is limited and reliable machine-readable identification is required. A typical Data Matrix symbol can encode up to 3,116 numeric or 2,335 alphanumeric characters (approximately 1,556 bytes) depending on the symbol size (Karrach and Pivarčiová, 2021). Compared to QR codes, Data Matrix symbols generally achieve higher data density for small payloads and therefore require less physical space, which makes them particularly suitable for marking small objects such as electronic components, medical devices, or industrial parts (Hu et al., 2025).

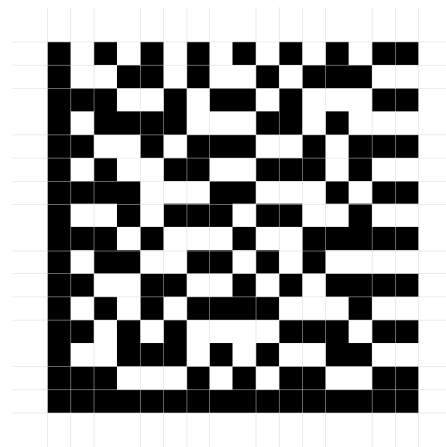


Figure 2: Schematic example of a 16×16 Data Matrix symbol. The solid L-shaped finder pattern fixes orientation, while the alternating borders support estimation of the module grid. Source: own elaboration.

A Data Matrix symbol consists of a square or rectangular grid of black and white cells, commonly referred to as modules. Each module represents a binary value (0 or 1), depending on whether the cell is light or dark. The symbol structure contains several characteristic elements that support reliable detection and decoding. Two adjacent borders form a solid “L”-shaped finder pattern, which defines the orientation and location of the symbol in the image. The other two borders contain alternating dark and light modules forming the so-called timing pattern, which helps determine the size of the grid and the spacing between modules (Hu et al., 2025). Within these borders lies the data region where the encoded information and error-correction codewords are stored.

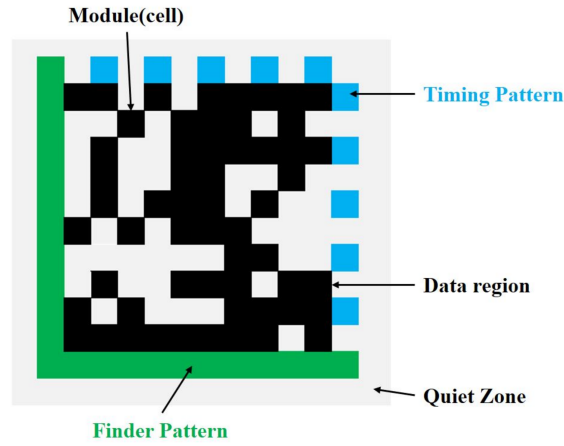


Figure 3: Structural elements of a Data Matrix symbol: the solid finder pattern, alternating timing pattern, data region, individual modules, and surrounding quiet zone. Adapted from Liao et al. (2022).

Figure 2 shows the border organization of a schematic 16×16 symbol, while Figure 3 identifies the functional regions that support orientation, grid estimation, and decoding.

Data Matrix symbols exist in several predefined sizes. Square symbols range from 10×10 to 144×144 modules, while rectangular formats range from 8×18 to 16×48 modules. The specific symbol size is determined automatically based on the amount of encoded data, with the smallest possible symbol selected for efficiency (Hu et al., 2025). Larger symbols may contain additional alignment structures that subdivide the data region into smaller blocks, which improves decoding robustness for high-capacity symbols.

Modern Data Matrix implementations follow the ECC 200 standard (ISO/IEC 16022), which incorporates Reed–Solomon error correction codes. This error correction mechanism enables reliable decoding even when parts of the symbol are damaged or obscured. In practical scenarios, correctly decoding the encoded message is often possible even if approximately 20–30% of the symbol is unreadable (Karrach and Pivarčiová, 2021; Hu et al., 2025). The ECC 200 standard also supports several encoding schemes (such as ASCII, C40, Text, and Base256), allowing efficient data compaction depending on the type of information being stored.

From a computer vision perspective, several structural properties make Data Matrix particularly suitable for automated visual identification. The high-contrast binary pattern of modules simplifies segmentation in camera images, while the distinctive L-shaped finder pattern enables robust localization even under rotation or moderate perspective distortion. Once the symbol is detected, projective transformation techniques can be applied to normalize the grid structure, allowing reliable sampling of the module values. These characteristics make Data Matrix symbols well suited for machine vision systems operating under non-ideal imaging conditions (Karrach and Pivarčiová, 2021).

Due to their compact size, high data density, and strong error correction capabilities, Data Matrix codes are widely used in industrial traceability systems. They are commonly applied in direct part marking (DPM), where codes are laser-etched or printed directly onto components in sectors such as automotive, aerospace, and electronics manufacturing. In logistics and warehouse environments, Data Matrix labels allow automated identification of items, containers, or pallets, enabling reliable tracking and integration with inventory management systems (Matuszczyk and Weichert, 2023). These properties make Data Matrix a particularly suitable visual identifier for computer vision-based product recognition systems in industrial and intralogistics environments.

The decoding of Data Matrix symbols follows a well-defined sequence of geometric and logical processing steps. In classical machine vision systems, decoding typically begins with the detection of candidate regions in the image, often using edge-based or contrast-based segmentation techniques. Once a potential symbol region is identified, the distinctive structural elements of the symbol are analyzed. In particular, the L-shaped finder pattern allows the decoder to determine the orientation of the symbol, while the timing pattern provides information about the grid spacing and module alignment.

After the symbol boundaries and orientation are established, a projective transformation is applied to compensate for perspective distortion and normalize the grid structure. The corrected symbol image is then sampled at the module level, converting the visual pattern of dark and light cells into a binary sequence. This binary representation is subsequently interpreted according to the Data Matrix encoding scheme, and the message is reconstructed using Reed–Solomon error correction codes. If the number of corrupted or unreadable modules remains within the error correction capability of the symbol, the original encoded message can still be recovered reliably.

Data Matrix decoding operates on the symbol's binary module pattern rather than on visually recognizable characters. Consequently, successful decoding produces an exact reconstruction of the encoded data sequence, while decoding failures typically occur when critical structural elements (such as the finder pattern) cannot be reliably detected or when too many modules are unreadable. In such cases, the decoder does not usually produce an incorrect character substitution but instead reports that the symbol cannot be decoded.

In practical industrial environments, however, image degradation factors such as motion blur, partial occlusion, strong perspective distortion, or poor illumination can prevent successful decoding even when parts of the symbol remain visible. In warehouse scenarios, the set of possible identifiers is typically known in advance, as Data Matrix symbols encode product identifiers that belong to a predefined inventory database. Under these conditions, the recognition problem can be interpreted as a constrained identification task rather than a fully unconstrained decoding problem. Therefore, similarity-based validation or correction strategies may be applied after an initial recognition attempt in order to

match partially decoded or noisy outputs against the set of valid product identifiers. Such approaches can improve the robustness of computer vision–based identification pipelines in environments where classical barcode decoding alone may fail.

Despite the robustness of the ECC 200 standard, reliable decoding of Data Matrix symbols can still fail in real-world imaging conditions. In practical machine vision applications several factors may prevent the decoder from successfully reconstructing the encoded bit-stream. One common failure mode occurs when the characteristic structural elements of the symbol cannot be reliably detected. For example, severe blur, low contrast, or partial occlusion may prevent the accurate identification of the L-shaped finder pattern or the timing pattern, which are essential for determining symbol orientation and grid geometry.

Another frequent source of decoding failure is geometric distortion. Strong perspective effects, extreme viewing angles, or motion-induced blur may distort the module grid to the extent that reliable sampling of the individual modules becomes impossible. Similarly, insufficient image resolution or excessive distance from the camera may reduce the apparent module size below the level required for accurate binary sampling.

In addition, environmental factors such as non-uniform illumination, reflections, surface damage in direct part marking (DPM), or printing imperfections may further degrade symbol readability. While the Reed–Solomon error correction mechanism can compensate for a limited number of corrupted modules, decoding becomes impossible once the number of unreadable modules exceeds the error correction capability of the symbol.

In such situations classical Data Matrix decoders typically return a decoding failure rather than an incorrect output. However, in industrial environments the set of valid identifiers is often known in advance. Consequently, partially recognized or noisy outputs obtained from vision-based recognition pipelines can still be matched against existing product identifiers using similarity-based validation strategies. This additional step can improve robustness in cases where traditional decoding alone is unable to recover the encoded message.

Internally, the encoded data within a Data Matrix symbol is not stored as visually interpretable characters but as a sequence of binary codewords distributed across the module grid according to a predefined placement algorithm. In the ECC 200 specification this process follows a deterministic mapping pattern, commonly referred to as the *Utah pattern*, which distributes bits of each codeword across different spatial regions of the symbol. As a result, adjacent characters in the encoded message are not represented by localized visual patterns but by module configurations that are spatially interleaved across the symbol.

This spatial distribution has two important consequences. First, it improves robustness against localized damage, since the bits of a single character are spread across multiple regions of the symbol and can be reconstructed through Reed–Solomon error correction.

Second, it means that visually similar symbols rarely correspond to similar encoded messages, as even small changes in the encoded data typically affect many modules across the grid. Consequently, Data Matrix decoding relies on recovering the binary module pattern rather than recognizing individual characters directly.

In damaged or partially occluded symbols, several preprocessing and restoration techniques can be applied prior to decoding. These include image binarization, morphological filtering, and geometric normalization to reconstruct the underlying module grid. Even when parts of the symbol remain unreadable, partial reconstruction of the binary module matrix may still be possible. In constrained industrial environments where the set of valid identifiers is known, such partially reconstructed patterns can be compared against candidate symbols from the identifier database, enabling probabilistic identification even when classical decoding fails.

5.4 Binary and continuous module similarity

A rectified Data Matrix symbol can be represented as an ordered vector of module values. For two binary vectors $\mathbf{x}, \mathbf{R} \in \{0, 1\}^N$, the Hamming distance counts the positions at which their values differ (Hamming, 1950):

$$d_H(\mathbf{x}, \mathbf{R}) = \sum_{i=1}^N \mathbb{I}(x_i \neq R_i), \quad S_H(\mathbf{x}, \mathbf{R}) = 1 - \frac{d_H(\mathbf{x}, \mathbf{R})}{N}. \quad (1)$$

Normalized Hamming similarity is bounded by zero and one and gives equal weight to every module disagreement. Its binary formulation requires each observed cell to be classified as dark or light before comparison. This hard decision discards information about cells whose intensity lies near the separation between the two levels.

A continuous representation replaces the observed binary value x_i with a normalized darkness value $p_i \in [0, 1]$. Extending the binary mismatch indicator by absolute difference gives

$$d_{\text{softH}}(\mathbf{p}, \mathbf{R}) = \frac{1}{N} \sum_{i=1}^N |p_i - R_i|, \quad S_{\text{softH}}(\mathbf{p}, \mathbf{R}) = 1 - d_{\text{softH}}(\mathbf{p}, \mathbf{R}). \quad (2)$$

This formulation is an L_1 comparison between a continuous observation and a binary reference. When $\mathbf{p}$ is binary, $|p_i - R_i|$ equals the binary mismatch indicator, so soft-Hamming similarity reduces exactly to normalized Hamming similarity. Intermediate cell values contribute proportionally rather than being forced into a full binary agreement or disagreement. The values p_i express relative darkness on the observed symbol's own intensity scale and are not calibrated probabilities.

Jaccard similarity provides a related binary-set comparison by dividing the intersection of dark modules by their union. It emphasizes agreement among dark modules but excludes joint light-module positions from the numerator and denominator. Hamming and soft-Hamming similarity instead evaluate agreement over the complete ordered module grid. The pipeline-specific metric choice and the empirical comparison of these alternatives are separated from this theoretical definition and reported with the evaluation methodology. Reference ranking and candidate acceptance are distinct operations. A similarity measure orders the candidate reference symbols, whereas an acceptance rule determines whether the leading candidate is sufficiently distinguishable from its competitors. Relative separation can be invariant to a common scaling of all candidate scores, but it does not establish that the leading score is absolutely reliable. This distinction is required when similarity evidence is accumulated across observations and when unresolved cases must remain possible. The implemented multi-candidate rule is inspired by the normalized-gap form of Dixon’s Q statistic, but it is used only as a deterministic separation heuristic: no outlier-test interpretation, p -value, or calibrated confidence is claimed. The precise rule and its configurable thresholds are specified in Chapter 6.

5.5 Evidence aggregation across observations

Repeated observations provide multiple pieces of evidence about the same physical identifier. Candidate-wise accumulation preserves both agreement and conflict: observations supporting the same identifier reinforce its total, while observations supporting different identifiers remain in separate totals. Normalization converts these totals into relative shares and enables comparison across episodes of different lengths. However, normalized shares do not retain the magnitude of the accumulated evidence because they sum to one whenever any evidence exists. Raw and normalized support therefore represent complementary quantities.

Evidence from sequential processing stages must also remain mutually exclusive at the observation level. If primary decoding, secondary decoding, and similarity recovery were allowed to vote simultaneously for one frame, a single image could be counted several times. Selecting at most one accepted evidence source per detected instance avoids this duplication and makes temporal accumulation interpretable as aggregation across observations rather than across internal pipeline stages.

6. MODEL AND SYSTEM DESIGN

This chapter constitutes the core technical contribution of the dissertation. Its objective is to describe the design of a complete visual product identification system that operates under visually realistic conditions approximating dynamic intralogistics environments. Rather than presenting isolated algorithmic components, the chapter adopts a system-oriented perspective in which data acquisition, model development, and decision logic are treated as tightly coupled elements of a unified identification process.

The individual components of this system draw substantially on established computer vision and decoding methods. Detection relies on a standard YOLO-based segmentation architecture (Section 6.3), and primary decoding uses an existing ECC200-compliant decoding library (Section 6.4); each is adapted only through the geometric preprocessing needed to connect it within the pipeline. The contribution of this chapter lies not in these individual components but in their system-level integration: the recovery and validation mechanisms introduced in Section 6.5 and the temporal evidence-aggregation logic introduced in Section 6.6 are proposed as part of this dissertation.

To illustrate the practical context of the proposed system, consider a typical HMLV warehouse scenario. A manual or autonomous vehicle moves along storage rows while a mobile camera continuously captures visual input. As the camera advances, multiple Data Matrix symbols may enter and leave the field of view. The same symbol can appear across several consecutive frames, often under significantly different visual conditions. For instance, a code may initially be observed at a large distance or under an oblique viewing angle, later becoming more prominent as the camera approaches, and eventually disappearing from view as the operator moves forward. In such environments, reliable product identification is often influenced by transient visibility conditions. While certain frames may allow confident decoding when the symbol is clearly visible and well aligned, overall identification reliability can be significantly improved by aggregating visual evidence across multiple observations over time. This scenario serves as a conceptual deployment motivation. The empirical study reproduces selected visual conditions through mobile video acquisition but does not implement vehicle integration, spatial localization, or WMS updating.

This observation motivates a fundamental design distinction between frame-level perception and sequence-level identification. At the frame level, the system performs visual recognition tasks such as Data Matrix detection, geometric normalization, and decoding attempts. However, at the sequence level, the system integrates information across multiple frames in order to improve identification reliability under degraded or dynamically changing imaging conditions. This temporal perspective reflects key characteristics of mobile vision deployment scenarios in warehouse environments and provides a concep-

tual foundation for the overall system design.

The proposed architecture therefore combines a frame-based visual recognition pipeline with higher-level decision logic that aggregates observations across time. This design enables both methodological analysis of individual components and potential applicability in real-world deployment scenarios. Importantly, it allows the system to balance robustness and efficiency by distinguishing between immediate decoding opportunities and delayed identification decisions based on accumulated evidence.

The structure of the chapter reflects this system-oriented viewpoint. It begins with a system-level overview that defines the functional architecture of the proposed pipeline, followed by a detailed description of the dataset design and acquisition strategy. The dataset is presented as an integral part of the system design, as its structure and controlled variability directly influence model behavior, generalization capability, and evaluation outcomes. Subsequent sections introduce the deep learning-based detection model, the geometric preprocessing and decoding mechanisms, and the recovery strategies applied when direct decoding is not feasible. The chapter concludes with a description of temporal integration and the planned framework for end-to-end evaluation, linking component-level behavior to overall system performance.

The design of the system is guided by several key principles.

First, robustness is prioritized over idealized performance. The primary goal of the proposed system is not to maximize benchmark accuracy under controlled laboratory conditions, but to achieve stable and predictable behavior under visually degraded and variable acquisition scenarios. This principle motivates the use of mobile camera imagery, the inclusion of negative samples, and the explicit analysis of performance across multiple environmental dimensions.

Second, end-to-end identification reliability is emphasized over isolated component metrics. While detection and dedicated decoding performance are evaluated individually, the system is ultimately assessed based on its ability to correctly identify products when all components operate together. This principle directly informs the inclusion of a similarity-based validation stage and the focus on integrated evaluation in later chapters.

Third, controlled variability is treated as a design requirement rather than a source of noise. Environmental factors such as distance, viewing angle, illumination, and motion-induced blur are not merely tolerated but intentionally structured within the dataset. By explicitly controlling these variables, the system design enables systematic analysis of failure modes and performance degradation patterns, which is essential for answering the formulated research questions.

Fourth, modularity and extensibility are incorporated into the pipeline design. Each component of the system—detection, decoding, validation, and integration—is implemented as a separable module with clearly defined inputs and outputs. This modular structure sup-

ports ablation studies, facilitates comparative evaluation, and enables future extensions, such as the integration of additional sensors, alternative Data Matrix decoders, or external system interfaces.

Fifth, the system design explicitly distinguishes between frame-level perception and sequence-level identification. While the core recognition pipeline operates on individual image frames, the intended deployment scenario involves continuous camera motion, where the same Data Matrix symbol may be observed across multiple frames under varying visual conditions. Consequently, identification reliability is treated as an emergent property of accumulated observations rather than as a purely single-frame inference outcome. This principle informs the integration of temporal reasoning mechanisms at the system level.

Together, these elements form a transparent and extensible framework for investigating robust visual identification of Data Matrix-coded products in dynamic environments representative of HMLV warehouse conditions. The following section presents the system-level architecture that serves as the conceptual backbone of the proposed approach.

The contribution is architectural and methodological rather than a claim of a new detector or decoder. The implemented framework combines established perception and decoding components with explicit reference validation, a failure-activated recovery branch, and episode-level evidence aggregation.

The chapter documents the implemented interaction between detection, decoding, validation, recovery, and temporal decision making. Quantitative results are reported separately so that architectural specification and empirical evidence remain distinct.

6.1 System-level architecture

The evaluated framework has two connected levels. The frame-level pipeline converts each detected Data Matrix instance into at most one accepted identifier and a continuous support value, or into no evidence when the observation remains unresolved. The episode-level pipeline processes the observations assigned to one episode, accumulates support by candidate identifier, and returns an *identified*, *ambiguous*, or *unresolved* outcome. Figure 4 summarizes this information flow.

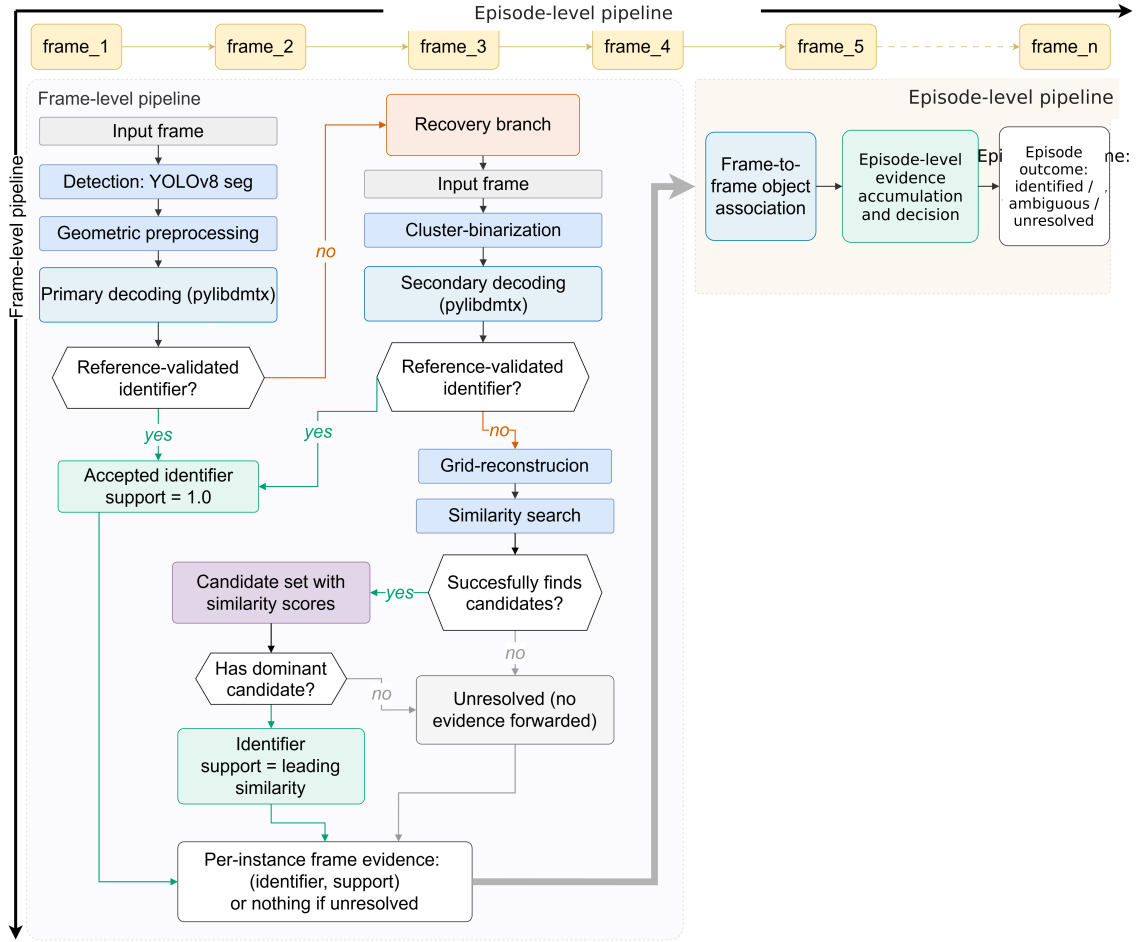


Figure 4: Proposed two-level Data Matrix identification architecture. At frame level, segmentation and geometric normalization feed primary decoding and, if needed, recovery; each observation yields at most one accepted identifier–support pair. At episode level, evidence is accumulated by candidate and classified as identified, ambiguous, or unresolved. Source: own elaboration.

At frame level, YOLOv8 instance segmentation localizes candidate symbols. Each mask polygon is converted into an ordered and validated quadrilateral, perspective-rectified, and padded with a synthesized quiet zone. A dedicated Data Matrix decoder then attempts to read the symbol, and the returned string is accepted only if it belongs to the known identifier set. Rejection by this open-set validation rule activates the recovery branch. Recovery first makes one bounded decode attempt on a re-binarized crop. If this also fails, the branch estimates the L-shaped finder geometry, reconstructs a 16×16 module representation, ranks the known digital references using soft-Hamming similarity, and forwards the leader only when it passes the dominance rule.

Every detected instance therefore contributes at most one evidence pair (identifier, support). A validated primary or secondary decode contributes support 1.0. An accepted similarity candidate contributes its raw soft-Hamming score. An unresolved instance contributes nothing, rather than a zero-valued vote. Detection confidence and intermediate geometric measurements are retained for evaluation and

diagnosis, but they are not accumulated as identifier support.

At sequence level, tracking associates quadrilaterals across frames. The annotated sequence manifest defines an episode’s temporal extent and stores the track identifiers whose observations belong to it, without supplying the expected code identity to the decision layer. The episode pipeline accumulates support for each observed candidate, normalizes the accumulated values, and applies the dominance condition. The runtime decision is thus ground-truth-free; correctness is assessed only afterwards by the evaluation modules.

Component	Role in the contribution	Interpretation
YOLOv8 segmentation and dedicated Data Matrix decoding	Adopted	Established perception and decoding components used as the primary processing path.
Mask-to-quad conversion, quiet-zone rectification, L-shape recovery, and soft-cell similarity	Adapted and integrated	Existing geometric and similarity principles are configured and connected for the study’s controlled Data Matrix setting.
Reference validation and mutually exclusive frame-evidence rule	System integration	Converts stage outputs into ground-truth-free evidence while preventing multiple stages from voting for one observation.
Episode construction, support accumulation, and shared dominance logic	Dissertation-specific system formulation	Connects frame-level evidence to an explicit sequence-level identification decision. The contribution concerns the formulation and integration, not a new tracking or statistical test algorithm.
Controlled, stage-wise evaluation design	Methodological contribution	Separates detection, decoding, recovery, identification, and operational interpretation under recorded acquisition variability.

Table 8: Positioning of the principal pipeline components within the dissertation contribution.

The implemented scope comprises frame-level detection, geometric normalization, dedicated decoding, reference validation, failure-activated recovery, and episode-level evidence aggregation. Sharpness estimates are retained as diagnostic metadata rather than used as a decoding gate. Geometric recovery terminates in direct soft-cell comparison with the reference library, without rendering the reconstructed matrix for a further decoding attempt. Identifier support is accepted through reference-validated decoding or dominance-tested similarity; no absolute similarity floor is applied. Spatial localization, inventory-

state updating, and WMS integration are outside the implemented system boundary and are treated as operational extensions.

6.2 Dataset design and data acquisition

The dataset used in this dissertation is designed as an integral component of the proposed visual identification system rather than as a standalone benchmark. Its primary purpose is to support controlled, repeatable evaluation of Data Matrix detection, decoding, and validation under conditions that are representative of mobile visual acquisition scenarios inspired by intralogistics environments. Consequently, the dataset design emphasizes controlled environmental variability, realistic negative samples, and strict separation between training, validation, and test data.

6.2.1 Acquisition setup

Image data were acquired indoors with an Apple iPhone 12 operated handheld, without fixed mounting, stabilization equipment, or manual camera-setting overrides. Videos were recorded at 1920×1080 pixels and 29.98 frames per second; the video data rate reported by the source files was approximately 8.48 Mb/s (8.63 Mb/s total stream rate).

Data Matrix codes were physically attached to surfaces and objects at predefined locations, mimicking the visual characteristics of product labeling in intralogistics contexts. The acquisition environment was intentionally kept flexible rather than highly structured, allowing variation in background appearance, surface texture, and surrounding visual clutter. Handheld operation created natural variation in viewpoint, apparent size, and motion-related blur. The setup does not represent a fully operational warehouse; it is an indoor, HMLV-inspired visual setting designed to expose the identification pipeline to variability and uncertainty without claiming deployment validation.

6.2.2 Macro-scene concept

To structure the dataset and to prevent information leakage, the concept of a *macro-scene* is introduced. A macro-scene is defined as a coherent acquisition context characterized by a fixed physical configuration of the Data Matrix code, background, and surrounding objects. Within a macro-scene, multiple image frames may be captured under varying acquisition parameters, such as distance, viewing angle, or illumination, while the underlying physical setup remains unchanged.

Macro-scenes serve as the fundamental unit of dataset partitioning. All frames belonging to the same macro-scene are assigned exclusively to either the training, validation, or test split. This design prevents the model from observing visually similar configurations across different splits and enforces a more realistic evaluation of generalization performance.

By operating at the macro-scene level, the dataset avoids overly optimistic performance estimates that can arise from random frame-level splitting, especially in scenarios where consecutive frames or nearly identical viewpoints are present.

6.2.3 Controlled environmental variability

Within each macro-scene, controlled variability is introduced along several dimensions that are relevant to visual Data Matrix recognition. These dimensions are treated as explicit design variables rather than incidental noise.

The selection of illumination, viewpoint, motion-related blur, apparent size, and partial visibility preceded model development and evaluation. These dimensions reflect expected sources of image degradation in mobile camera acquisition and were incorporated into the data-collection design rather than selected retrospectively from the reported results.

To enable systematic analysis of detection and decoding robustness, environmental variability is explicitly structured and encoded into a set of predefined categorical dimensions. The dataset design introduces controlled variation along four dimensions: illumination, viewpoint, object condition, and blur. Illumination is recorded at the frame level; viewpoint angle, object condition, and blur are recorded per Data Matrix instance, since a single frame may contain multiple codes under different conditions. These dimensions are consistently referenced throughout the empirical evaluation in Chapter 7.

Illumination conditions are structured into three discrete categories, denoted as F1 to F3, reflecting the lighting configurations used during data acquisition. These categories capture both favorable and degraded lighting scenarios commonly encountered in indoor environments with natural or mixed light sources.

F1 corresponds to natural daylight. F2 represents semi-dark conditions with the blinds approximately 90% closed and no artificial lighting or camera flash. F3 combines natural daylight with the iPhone camera flash, creating a frontal illumination condition that may introduce glare, reflections, or reduced code contrast.

This illumination categorization allows systematic analysis of how lighting degradation and unfavorable light direction affect detection and decoding performance, particularly in combination with distance and viewpoint-related variability.

Variability related to camera **viewpoint** is captured through two categories, denoted S1 and S2. An S1/S2 suggestion is derived automatically from the geometry of the annotation object’s quadrilateral and is subsequently reviewed during annotation. S1 (frontal) corresponds to a near-orthogonal viewing angle; S2 (oblique) corresponds to a non-orthogonal viewing angle. This dimension captures the effect of realistic mobile acquisition on geometric distortion.

Object condition combines a manually annotated occlusion flag with an automatically derived too-small classification based on a minimum pixel-size threshold, yielding three categories: normal (fully visible, unoccluded), occluded (partially occluded), and too small

(below the size threshold).

Blur is annotated manually per object as blurred or not blurred, capturing motion- or focus-induced image degradation.

In addition to positive samples containing visible Data Matrix codes, the dataset explicitly includes negative frames without any visible code. These negative frames are used to evaluate false-positive detection behavior and to prevent the model from learning an implicit prior that a code is always present.

By defining these dimensions explicitly, the dataset supports stratified performance analysis and enables consistent comparison of results across different environmental conditions.

Table 9: Structured environmental variability dimensions used in dataset design

Dimension	Category	Description
Illumination (F)	F1	Good, homogeneous lighting with high contrast
	F2	Reduced lighting conditions; semi-dark environment with ambient light only
	F3	Frontal light source; glare, reflections, or reduced contrast
Viewpoint (S)	S1	Near-orthogonal viewing angle
	S2	Oblique, non-orthogonal viewing angle
Object condition	Normal	Fully visible and above the minimum-size threshold
	Too small	Minimum side length does not exceed 40 pixels
	Occluded	Partial visibility determined manually; this category takes precedence over the size rule
Blur	Blurred / Not blurred	Independent manual judgment of motion- or focus-induced image blur

Representative fully visible and occluded observations are shown in Figures 46 and 47 in Appendix B. These examples are illustrative visual samples rather than an exhaustive or systematically stratified representation of all variability dimensions defined in the dataset. Furthermore, a single image frame may contain multiple Data Matrix symbols observed under different geometric and photometric conditions. Consequently, variability dimensions such as viewpoint, visibility, and effective scale are inherently object-specific and cannot be uniquely attributed to the frame as a whole.

The selected examples therefore aim to provide a representative visual overview of the types of acquisition variability encountered during data collection, including differences in illumination, motion-induced blur, partial occlusion, and oblique viewing angles. The figure should thus be interpreted as a qualitative illustration of dataset characteristics rather than as a formal categorical sampling of the defined variability dimensions.

6.2.4 Annotation protocol

Annotation was performed by one researcher, so no inter-annotator agreement statistic is reported. The full test-set manifest underwent a second manual review because its object metadata are used in the reported condition-specific evaluation. The review checked polygons and condition labels, including blur. During failure-case inspection, labels that were clearly inconsistent with the image evidence were corrected. This was a verification and correction procedure rather than an independent blinded agreement study; no agreement coefficient or error-rate estimate is claimed.

Frame-level annotation is organized by image frame and object instance. Each manifest row represents one frame, while its object field stores a dictionary indexed by the ground-truth Data Matrix identifier. Every object record contains a polygon in pixel coordinates and object-specific metadata. Validated model-assisted polygons are ordered quadrilaterals, whereas manually drawn polygons may contain three or more points. The frame-level fields indicating positivity, the number and identifiers of visible codes, and the aggregate viewpoint category are derived from the stored objects when the manifest is saved.

The annotation polygon follows the visible symbol boundary rather than the surrounding background or quiet zone. Polygon area is recorded both in pixels and as a proportion of the frame area. Absolute width, height, area, and minimum-side measurements are recomputed from the polygon, together with in-plane rotation. These geometric quantities support scale and viewpoint analysis but do not supply the identifier to the runtime pipeline. Negative frames contain no object annotation. Multiple visible Data Matrix symbols are represented as separate objects within the same frame.

Object condition is represented by the mutually exclusive categories *normal*, *too small*, and *occluded*. An object is classified automatically as *too small* when its minimum side does not exceed 40 pixels. Occlusion is judged manually and takes precedence over the automatic size rule because it records that the complete symbol is not visible. Blur is recorded as an independent manual flag, allowing it to coexist with either small scale or occlusion. The displayed sharpness measures support review but do not overwrite the manual blur judgment.

Viewpoint is derived from the four interior angles of a quadrilateral. The category S1 is assigned when every corner lies within 15 degrees of a right angle; otherwise the object is assigned to S2. At frame level, S2 is assigned when at least one visible object has an S2 viewpoint. Model-assisted suggestions are generated by YOLOv8 segmentation

followed by validated mask-to-quadrilateral conversion. Suggestions remain provisional until reviewed and saved, and any prediction for which a valid quadrilateral cannot be obtained is omitted rather than replaced by a bounding box.

Annotation quality control combines manual review with consistency checks. Stored object condition and viewpoint values are compared with values recomputed from current geometry, while the manual blur judgment can be compared with the diagnostic sharpness classification. These checks identify records requiring review without silently changing manually assigned fields.

6.2.5 Episode annotation and construction

The sequence manifest defines the ground truth for episode boundaries and post-hoc identity evaluation. An episode is one code’s continuous visibility span within one video. Visibility runs are merged when no more than 15 invisible frames occur between consecutive visible observations; a larger gap creates a new episode. The episode span extends from the first to the last frame and therefore includes intervening frames in which the object was not detected. This definition preserves temporary detection losses without equating an episode with a strictly contiguous list of detections.

Episodes are constructed in five stages. The detector first processes every video frame. Quadrilaterals are then associated across consecutive frames, after which a reviewer assigns the physical code identifier to each track or discards the track. Visible frames assigned to the same code are pooled and separated using the gap rule. Finally, the episode bounds and visibility metadata are reviewed manually. The reviewed fields include whether any fully visible frame occurs, the first such frame, and the principal reason for partial visibility when no fully visible observation exists.

A track and an episode represent different entities. Tracking provides a temporary association between observations and may divide one physical symbol into several track identifiers after a sufficiently long detection loss. Episode construction can reconcile such tracks through manual code assignment and the gap rule. Consequently, the sequence manifest stores the list of track identifiers associated with each episode. During runtime evaluation, these identifiers select the quadrilaterals belonging to the episode, while the expected Data Matrix identifier remains unavailable to the decision layer and is used only for subsequent correctness assessment.

6.2.6 Train–validation–test separation

Dataset partitioning is performed at the macro-scene level to ensure strict separation between training, validation, and test data. This separation strategy prevents leakage of visual context information and supports a realistic assessment of generalization performance.

The training split is used exclusively for model learning and hyperparameter selection. The validation split supports model selection and parameter tuning without influencing final performance reporting. The test split is reserved for all quantitative evaluations reported in the Results chapter.

This partitioning strategy reflects the intended deployment scenario, in which the system is expected to encounter previously unseen physical configurations and visual contexts rather than slight variations of known scenes.

6.2.7 Dataset statistics

The dataset is organized into **macro-scenes**: coherent acquisition contexts within which the physical configuration of symbols and surroundings remains fixed while acquisition parameters vary. The dataset comprises three macro-scenes, one per split, each defined by a distinct physical code layout and a disjoint code set, so that no code identifier or physical configuration is shared across splits. Within each macro-scene, one video was recorded per illumination condition (F1–F3). The resulting split is summarized in Table 10; the test split, used throughout the Results chapter, contains 1015 annotated Data Matrix instances. The HMLV-inspired intralogistics context informs the design of the dataset and the integration of the identification pipeline rather than the internal operation of the detector or the dedicated Data Matrix decoder. It motivates mobile-camera observations, controlled variation in illumination, viewpoint, blur, scale, and partial visibility, as well as repeated observations of the same symbol. It also motivates reference validation and episode-level evidence integration, because an identifier is interpreted across incomplete and visually variable observations rather than from an isolated ideal image. Detection, decoding, and identification correctness are nevertheless evaluated using task-level definitions and metrics that do not depend on a warehouse-management system, product master data, or inventory records. The domain context therefore defines the operationally motivated acquisition conditions and system requirements, whereas the evaluation measures the behaviour of the visual identification process itself.

The test split contains 1,015 annotated ground-truth Data Matrix objects. Section 7.1 distinguishes this count from the broader set of evaluation records used when characterizing detection outcomes.

Overall, the dataset design supports controlled experimentation while maintaining a clear link to practical visual identification challenges. By embedding environmental variability and strict data separation into the system design, the dataset provides a robust foundation for evaluating detection robustness, decoding reliability, and end-to-end identification performance.

Split name	Size
Train	1216 frames
Validation	933 frames
Test	922 frames

Table 10: Train–validation–test split sizes (frame count)

Split	Macro-scene	Data Matrix identifiers	Videos
Train	Train macro-scene	DataMatrix1, DataMatrix2, DataMatrix5, DataMatrix7	train_f1, train_f2, train_f3
Validation	Validation macro-scene	DataMatrix6, DataMatrix8	val_f1, val_f2, val_f3
Test	Test macro-scene	DataMatrix3, DataMatrix4	test_f1, test_f2, test_f3

Table 11: Macro-scene manifest and illumination-specific videos. Each split contains one distinct macro-scene and a disjoint identifier set.

Each split contains exactly one macro-scene recorded once under each illumination condition. No physical code layout or Data Matrix identifier is shared across train, validation, and test splits. This separation prevents leakage through repeated physical configurations, symbol appearance, or code identity, while retaining illumination variation within every split.

Rather than constituting a general-purpose benchmark, the dataset is conceived as an experimental instrument supporting the systematic design and analysis of the proposed identification pipeline.

6.3 Data Matrix detection model

The first computational component of the proposed pipeline is the automatic detection of Data Matrix regions within the input image. This task is formulated as a visual localization problem in which candidate regions containing Data Matrix codes must be identified under varying viewing conditions, including perspective distortion, motion blur, and partial occlusion.

6.3.1 Model selection

Detection is implemented using the YOLOv8 instance segmentation architecture. The YOLO family of models has become widely adopted in real-time visual perception systems due to its favorable trade-off between detection accuracy and computational effi-

ciency. Prior research on visual code recognition commonly employs deep learning–based object detectors, such as YOLO or Faster R-CNN, as the primary localization component within multi-stage decoding pipelines (Almeida et al., 2020, 2021; Ecula et al., 2023; Matuszczyk and Weichert, 2023; Corpuz et al., 2025; Muallim et al., 2025). As discussed in the literature review, such architectures are typically used to isolate candidate code regions prior to geometric normalization and decoding.

In the proposed system, the choice of detection model is additionally constrained by the requirements of downstream geometric processing. Since the pipeline relies on contour estimation and perspective rectification, accurate representation of the visible code boundary is required rather than coarse object localization.

For this reason, the segmentation variant of YOLOv8 is selected instead of a conventional bounding box detector. The model predicts pixel-level masks for each detected instance, enabling a more precise approximation of the code boundary under perspective distortion or partial visibility. These masks support subsequent contour extraction and quadrilateral fitting, which form the basis of geometric normalization in the decoding stage.

Compared to bounding box detection, segmentation-based localization is therefore expected to improve the stability of rectified symbol representations and reduce geometric error propagation to downstream decoding and similarity-based identification stages.

The selection was guided by the requirements of the complete identification pipeline rather than by an attempt to establish a detector benchmark. The systematic review in Chapter 2 identifies one-stage YOLO variants as the most frequently adopted detector family in the reviewed industrial and logistics corpus, particularly where visual localization is combined with operational constraints. This prevalence does not establish universal superiority, but it provides an application-relevant basis for selecting a mature single-stage detector family.

YOLOv8 provides a documented instance-segmentation implementation, pretrained weights, and a lightweight `yolov8n-seg` variant (Jocher et al., 2023). These properties support reproducible model development while keeping the learned perception component proportionate to the scope of the study. The choice follows the required interface between learned perception and explicit downstream processing, rather than a claim that YOLOv8 is intrinsically more accurate than every alternative detector.

Later YOLO variants, transformer-based detectors such as RT-DETR, and two-stage architectures such as Faster R-CNN and Mask R-CNN remain valid alternatives. A fair empirical comparison would require identical annotations, data splits, training budgets, operating points, and downstream geometric and identification evaluation. Such a detector benchmark is outside the present dissertation, whose empirical focus is the behaviour of the implemented detection–decoding–recovery–aggregation pipeline. Accordingly, no performance superiority over these untested alternatives is claimed.

6.3.2 Training configuration

The detector was trained with Ultralytics 8.3.81 under Python 3.12.3 using the pretrained yolov8n-seg segmentation model. The recorded run used an image size of 960×960 pixels, batch size 16, eight data-loading workers, 100 epochs, seed 0, deterministic execution, and cosine learning-rate scheduling. Training completed all 100 epochs; the configured patience of 30 epochs did not trigger early stopping. The framework’s automatic optimizer selection resolved to AdamW with learning rate 0.002 and momentum 0.9.

Training used device 0 on an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB memory, with PyTorch 2.5.1+cu121. Final evaluation was executed on the CPU of a 13th Gen Intel Core i9-13905H system. The test split was excluded from training, parameter selection, and checkpoint selection.

The unmodified Ultralytics validation procedure retained `best.pt`, the checkpoint with the highest validation fitness during the training run. No custom checkpoint-selection rule was introduced. The runtime environment used OpenCV 4.11.0 and `pylibdmtx` 0.1.10; the patch-level native `libdmtx` version was not recorded separately.

The associated repository release contains the complete `requirements.txt` file, the saved Ultralytics training configuration, the parameter manifest covering preprocessing and decision thresholds, and the final model checkpoint at documented locations. The reference library contains eight Data Matrix identifiers; the same payload strings derived from the physical labels are used for open-set validation and for reference-similarity search.

6.3.3 Data augmentation

No task-specific synthetic augmentation policy was designed or tuned for the reported detector. Framework-level augmentation was retained only as part of the baseline YOLOv8 training configuration and was not selected on the basis of test-set outcomes.

Robustness to the acquisition factors considered in this study was addressed primarily through the controlled collection of real video data rather than through synthetic simulation. The dataset design specified illumination, viewpoint, motion-related blur, apparent symbol size, and partial visibility before model training, and these sources of variation were represented in the macro-scene-level data collection. Framework-level augmentation therefore served as regularization, not as a separately evaluated mechanism for reproducing these conditions.

No augmentation ablation was performed, so the observed performance differences are not attributed to a particular transformation. A targeted comparison of blur, perspective, illumination, or partial-visibility augmentation would constitute a separate train-only experiment rather than a necessary extension of the present baseline. Severe occlusion and

very small apparent symbol size also remain information-limited cases that require downstream recovery and uncertainty handling in addition to detector training.

For reproducibility, the retained framework configuration used HSV adjustments (hue 0.015, saturation 0.7, value 0.4), translation 0.1, scale 0.5, horizontal flip probability 0.5, mosaic probability 1.0 with mosaic disabled during the final 10 epochs, RandAugment, and erasing probability 0.4. Rotation, shear, perspective, vertical flip, mixup, and copy-paste were disabled.

6.3.4 Detection output

For each input frame, the detection model produces a set of candidate Data Matrix regions represented by segmentation masks and associated confidence scores. These masks are subsequently used to extract localized symbol regions and to estimate geometric boundaries for downstream processing.

Since the proposed identification pipeline is strongly dependent on geometric normalization and decoding stability, detection errors directly influence the reliability of later processing stages. Consequently, the detection module is treated as a foundational component whose behavior must be analyzed in conjunction with geometric preprocessing and temporal evidence accumulation.

Representative qualitative examples of segmentation-based Data Matrix detection are presented in Figures 5–6. The examples illustrate typical detection behaviour under different illumination regimes and acquisition conditions, including variations in sharpness and viewing angle. Each example shows the original input frame alongside the corresponding segmentation output produced by the detection model.

Figure 5 presents detection examples of fully visible codes. These samples demonstrate that the segmentation model is capable of accurately localizing Data Matrix regions both under near-ideal frontal observation and in more challenging cases involving motion blur or oblique viewpoints. The qualitative results indicate that mask predictions remain spatially consistent with the visible symbol boundaries even when local image sharpness is reduced.

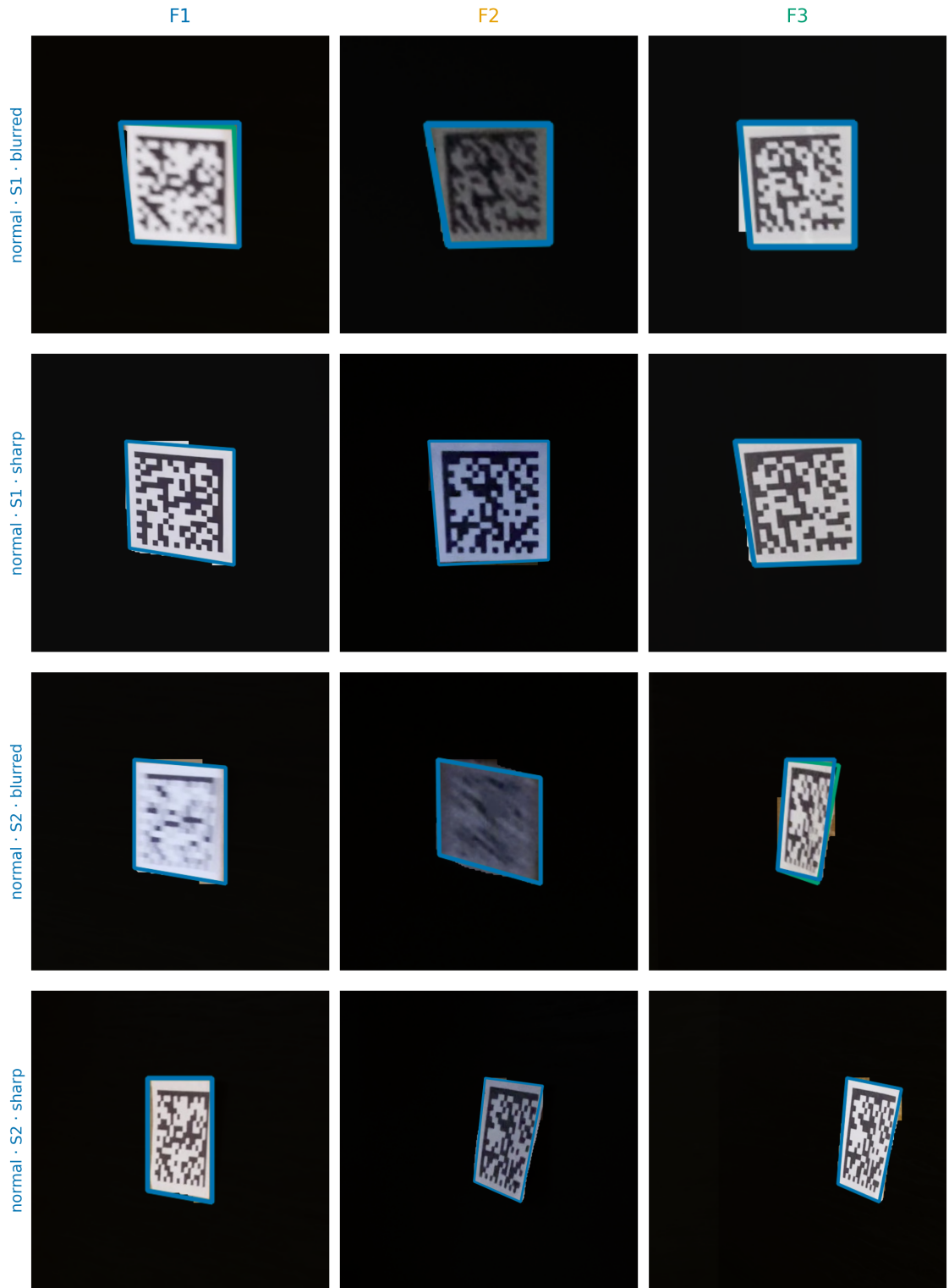


Figure 5: Segmentation outputs for fully visible Data Matrix symbols across lighting setups F1–F3, sharpness, and viewpoint. Blue contours show the predicted instance boundaries; localization remains stable under moderate blur and oblique viewing, although boundary uncertainty increases. Source: own elaboration.

Figure 6 illustrates detection output and performance under occlusion conditions. In such scenarios, decreased contrast and increased sensor noise introduce additional uncertainty

in visual perception. Despite these degradations, the segmentation model maintains the ability to identify candidate symbol regions, although mask confidence and boundary precision may vary depending on local lighting and motion effects.

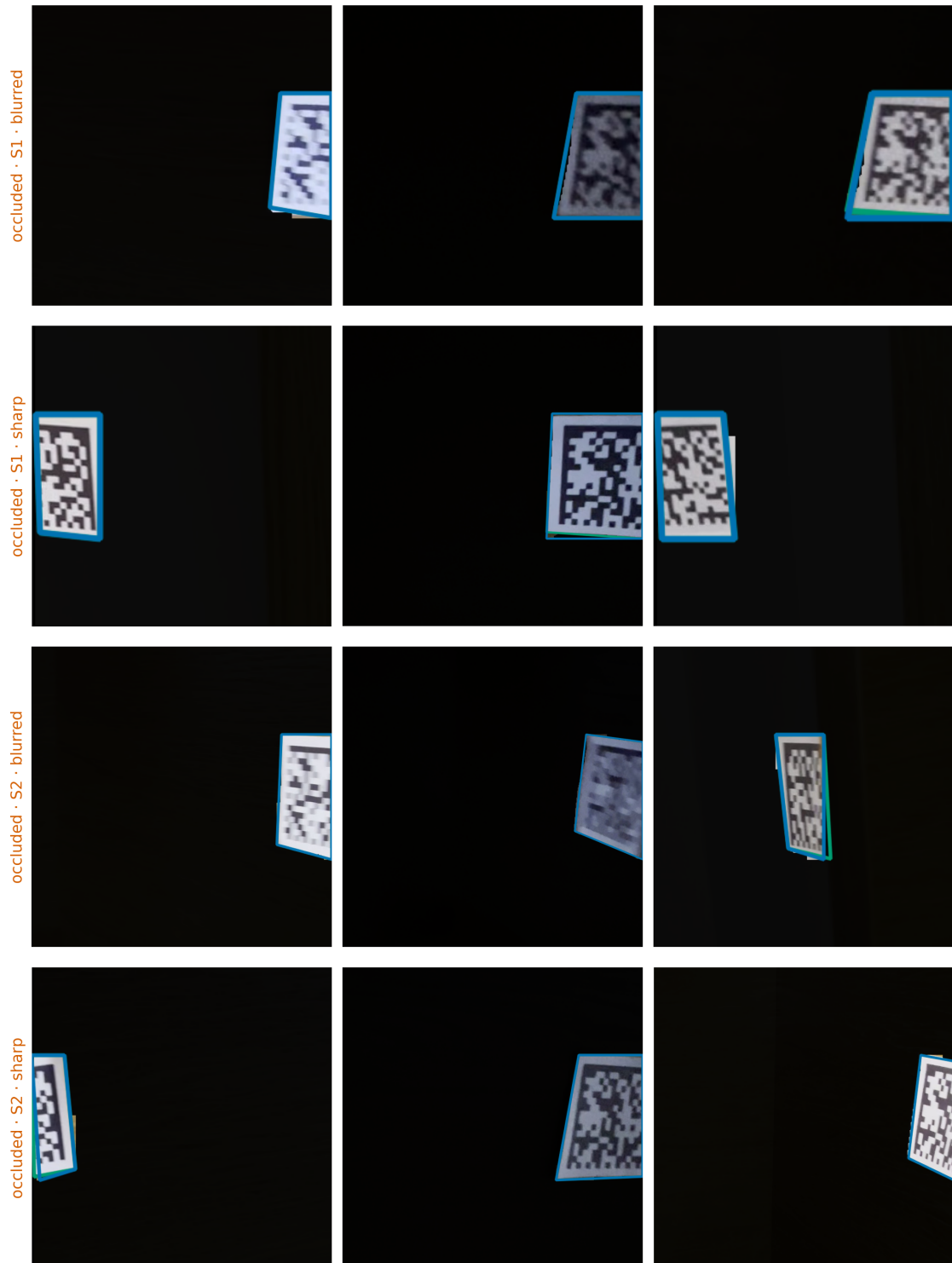


Figure 6: Segmentation outputs for occluded or insufficiently visible Data Matrix symbols across lighting setups F1–F3, sharpness, and viewpoint. The predicted contours follow the visible symbol extent, while severe occlusion and small apparent size reduce boundary completeness. Source: own elaboration.

The presented examples highlight the robustness of the segmentation-based detection approach in the presence of such photometric challenges, while also illustrating typical failure modes such as partial mask fragmentation or reduced boundary sharpness.

Overall, these qualitative examples provide visual evidence that segmentation-based localization constitutes a suitable front-end component for the subsequent geometric preprocessing and decoding stages of the proposed pipeline. The figures are intended to illustrate characteristic detection behaviour rather than to serve as a formal quantitative evaluation of model performance.

6.4 Primary decoding

Following the detection stage, the proposed pipeline performs a primary decoding attempt on each localized Data Matrix candidate region. The objective of this stage is to transform visually observed symbol patterns into machine-readable identifiers under realistic imaging conditions. While the detection model operates at the level of full-frame perception, primary decoding focuses on geometrically normalized symbol regions and aims to recover structured symbolic information from individual observations.

6.4.1 Mask-derived quadrilateral and crop extraction

For each YOLOv8 mask polygon, duplicate points are removed and a convex hull is formed. The hull is approximated using `approxPolyDP` with progressively increasing perimeter-relative tolerances until a four-point polygon is obtained. The resulting corners are ordered as top-left, top-right, bottom-right, and bottom-left, then validated for convexity, minimum area, minimum edge length, and consistency with the source-mask area. Candidates for which no reliable quadrilateral can be constructed are rejected. The validated quadrilateral preserves the projective geometry required for subsequent rectification.

The validated quadrilateral is mapped by a planar homography to a 300-pixel square crop. The symbol occupies a 200-pixel central square, while a 50-pixel white border is synthesized on each side as a quiet zone. This margin is part of the decoding procedure rather than a visual formatting choice because the annotation polygon follows the symbol boundary and therefore excludes the surrounding quiet zone required by the decoder. Image intensities are sampled from the original frame; the segmentation mask supplies geometry only.

6.4.2 Primary Data Matrix decoding and reference validation

Primary decoding is performed by `pylibdmtx`, a dedicated ECC200 Data Matrix decoder. The decoder is constrained to one result, a 16×16 symbol shape, and a 500 ms search

timeout. The fixed symbol shape reflects the present dataset rather than a general assumption about all Data Matrix symbols. These bounds reduce the cost of unsuccessful searches, which dominate runtime when image quality is poor.

The decoder returns either a string or no result. A returned string is not accepted using the ground-truth code of the evaluated object. Instead, the open-set validation rule checks whether it belongs to the known identifier set. A known identifier becomes primary frame evidence with support 1.0. An unknown string and a failed decode are both rejected and activate the recovery branch. Correctness relative to the annotated identifier is computed only after pipeline execution.

F1-normal-S1-sharp | code 3 | test_F1_test_f1_f000100.jpg | polygon source: detected

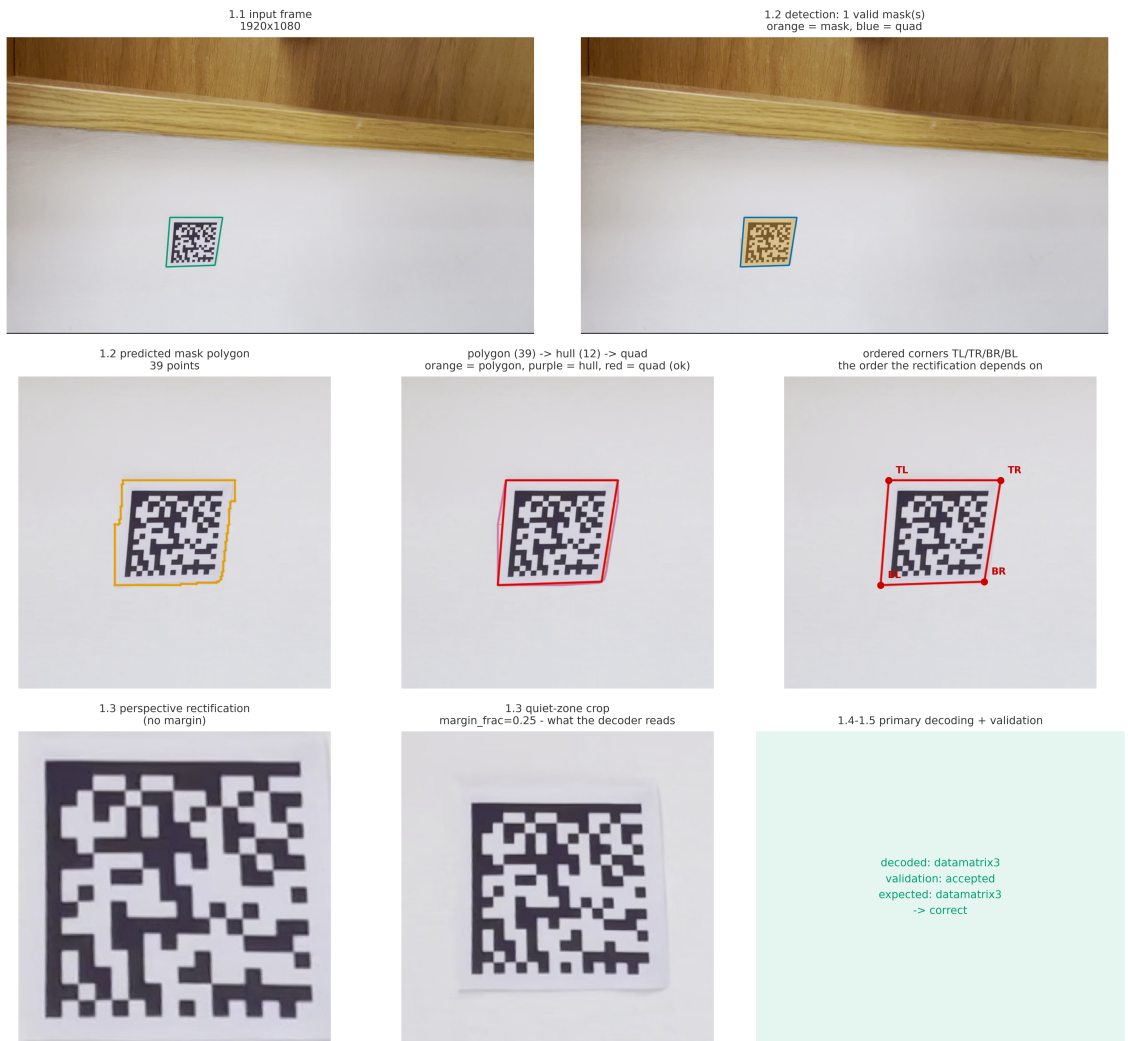


Figure 7: Successful primary decoding under F1, normal object condition, S1 viewpoint, and sharp acquisition. Instance segmentation is converted into an ordered quadrilateral, rectified and cropped with a quiet zone, after which the dedicated decoder returns a reference-validated identifier. Source: own elaboration.

This distinction produces three separate concepts: decoding success means that the library

returned a string; reference validation means that the string belongs to the allowed identifier set; identification correctness means that the accepted identifier matches ground truth. The first two are runtime events, while the third exists only in evaluation. Figures 7 and 8 show that the geometric stages can remain usable in both cases; the decisive difference is whether the rectified module pattern retains enough detail for the dedicated decoder.

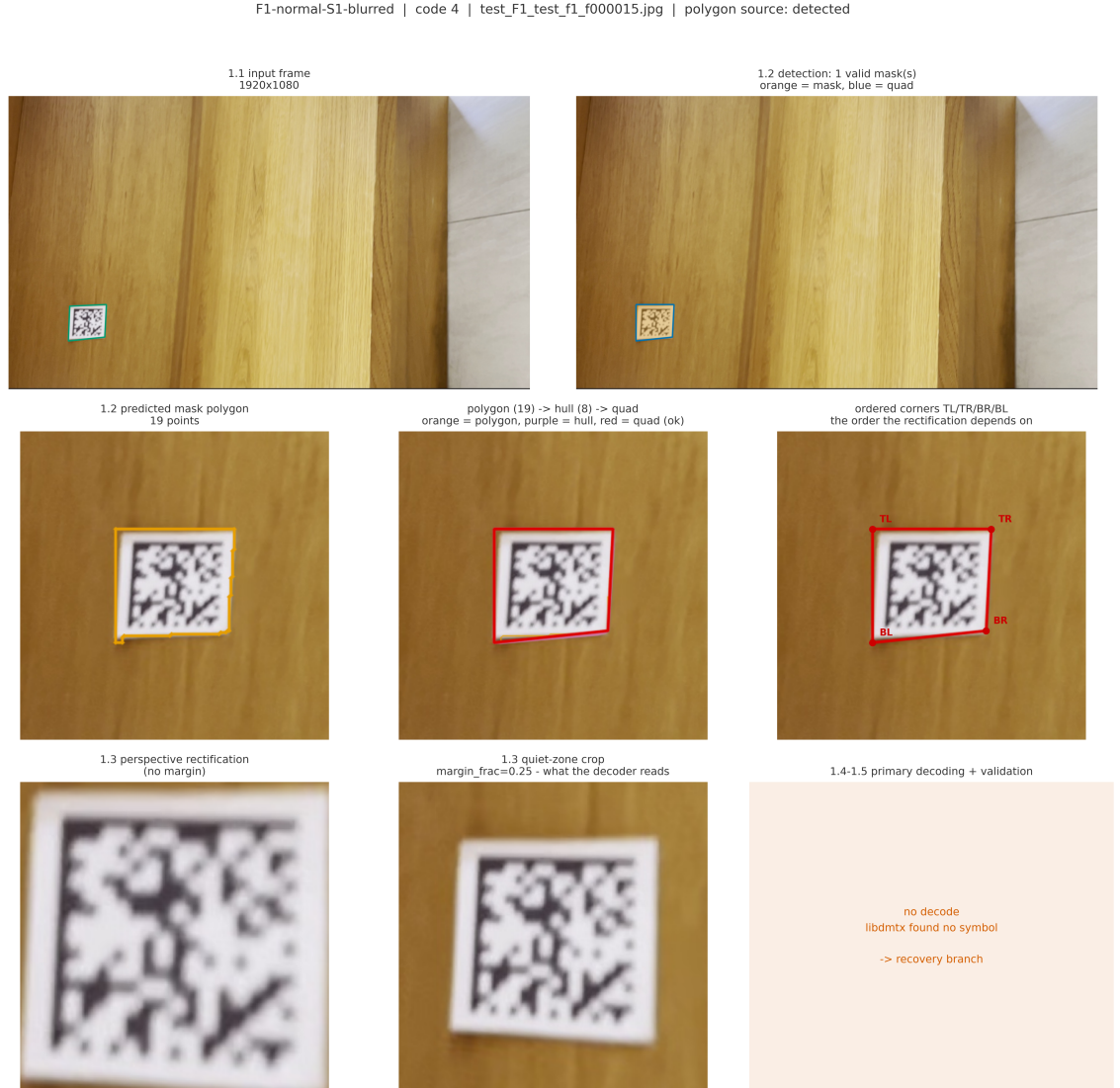


Figure 8: Primary-decoding failure under F1, normal object condition, S1 viewpoint, and blurred acquisition. Detection and geometric normalization produce a usable crop, but blur merges module boundaries and the dedicated decoder returns no symbol, so processing continues to the recovery branch. Source: own elaboration.

6.5 Recovery branch and similarity-based identification

The recovery branch is activated when primary decoding does not produce a reference-validated identifier. As shown in Figure 9, cluster binarization is followed by one bounded secondary decoding attempt. If the observation remains unresolved, geometric reconstruc-

tion produces a module representation that is compared directly with the reference library using soft-cell similarity.

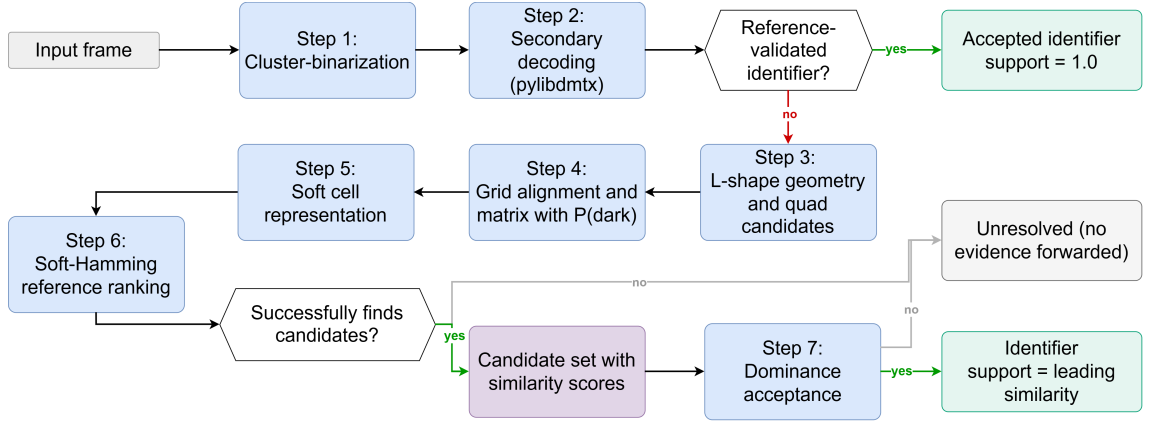


Figure 9: Recovery branch of the evaluated frame-level pipeline. A successful reference-validated secondary decode terminates the branch after Step 2; Steps 3–7 are evaluated only when decoding remains unresolved. Source: own elaboration.

Re-binarization and secondary decoding. A weighted one-dimensional k -means procedure with $k = 2$ is applied to the grayscale histogram of the crop. The darker cluster is mapped to black and the lighter cluster to white. One bounded `pylibdmtx` attempt is made on this clustered binary crop in an isolated worker process. Isolation prevents a rare native-library process termination from aborting the complete batch. A known decoded identifier resolves the frame with support 1.0; otherwise recovery continues.

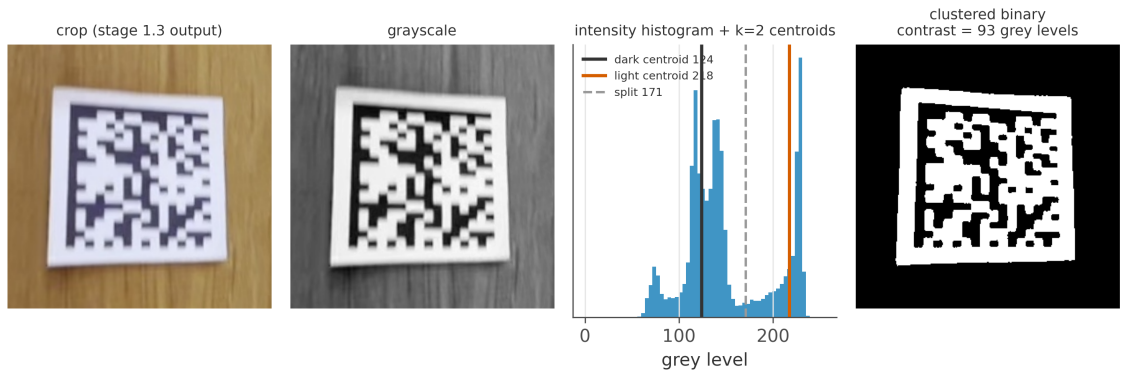


Figure 10: Cluster-based binarization for an observation resolved by secondary decoding. The crop is converted to grayscale, split using two intensity centroids, and binarized before the decoder returns the reference-validated identifier `datamatrix4`. Source: own elaboration.

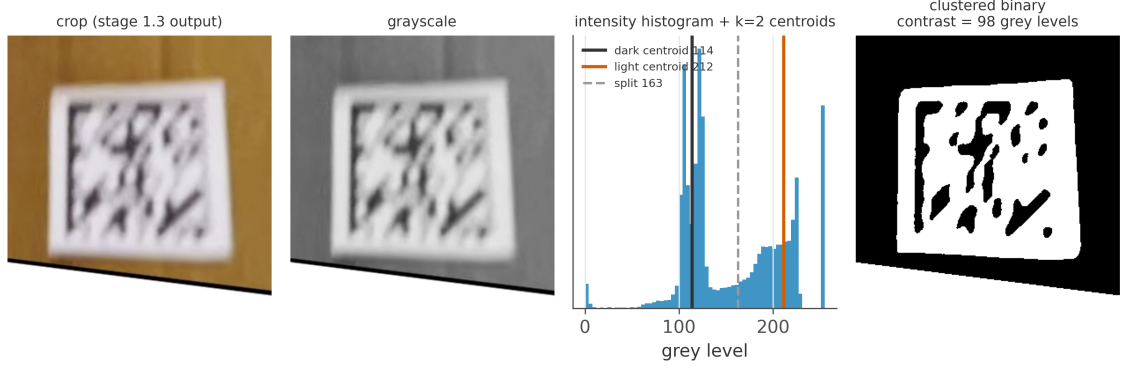


Figure 11: Cluster-based binarization for an observation that remains unresolved after primary and secondary decoding. Although the intensity separation produces a binary image, blur merges module boundaries and the secondary decoder returns no symbol. Source: own elaboration.

Figures 10 and 11 contrast the two outcomes of the bounded secondary-decoding attempt. Re-binarization can restore sufficient separation in some crops, but it cannot recover module boundaries that are no longer visually distinguishable.

L-shape geometry and candidate quadrilaterals. The geometric branch applies Canny edge detection, line-segment detection, outer-envelope filtering, merging of approximately collinear segments, connected L-pair search, RANSAC fitting of the two solid finder-pattern arms, opposite-edge estimation, and corner construction. The search is restricted to a region derived from the known crop geometry and expanded by 0.18 of its size. Up to three source-aware quadrilateral candidates are considered: the raw refined quadrilateral, a candidate bounded by the mapped source rectangle, and, for detector-derived polygons only, a candidate bounded by the expanded search rectangle. Candidate selection is based on the Data Matrix border-structure score described below. RANSAC line fitting is confined to the geometric recovery branch.

Figures 12 and 13 show that edge abundance alone is insufficient: the geometric branch requires a connected pair consistent with the solid L-shaped finder pattern before it can propose a symbol quadrilateral.

Grid alignment and module reconstruction. Each candidate is tightly rectified and rotated to canonical orientation. For a 16×16 grid, offsets are evaluated over ± 0.30 module in both directions using seven positions per axis, giving 49 grids. The binary sampler uses the central 60% of each cell. Grid selection uses the known Data Matrix border structure:

$$B = 0.35B_{\text{left}} + 0.35B_{\text{bottom}} + 0.15T_{\text{top}} + 0.15T_{\text{right}}, \quad (3)$$

where B_{left} and B_{bottom} measure the two solid finder edges and T_{top} and T_{right} measure

agreement with the alternating timing patterns. A small offset penalty favours the smaller correction when structural scores are nearly equal. The four weights and the offset-search settings are hand-set design parameters, not calibrated estimates.

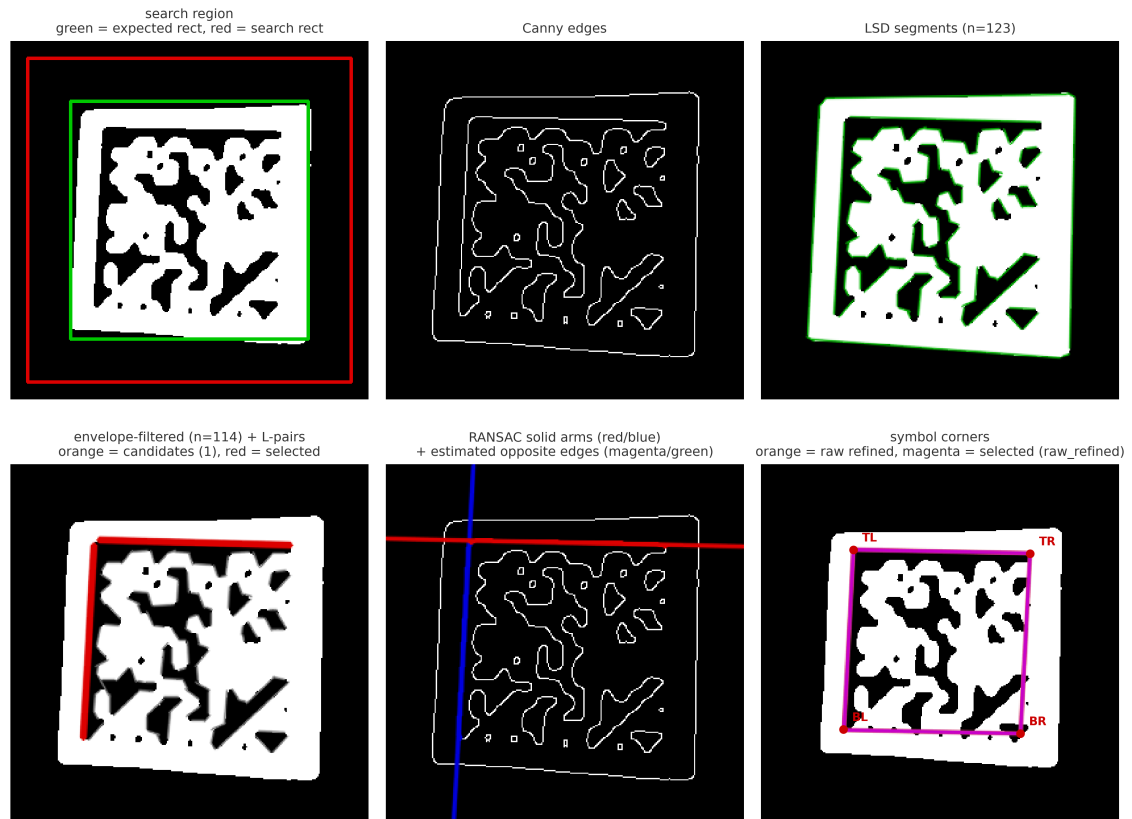


Figure 12: Successful geometric reconstruction after decoding failure. Edge extraction and line-segment filtering produce one admissible L-pair; RANSAC fits the solid finder-pattern arms, estimates the opposite boundaries, and yields a selected quadrilateral for subsequent rectification. Source: own elaboration.

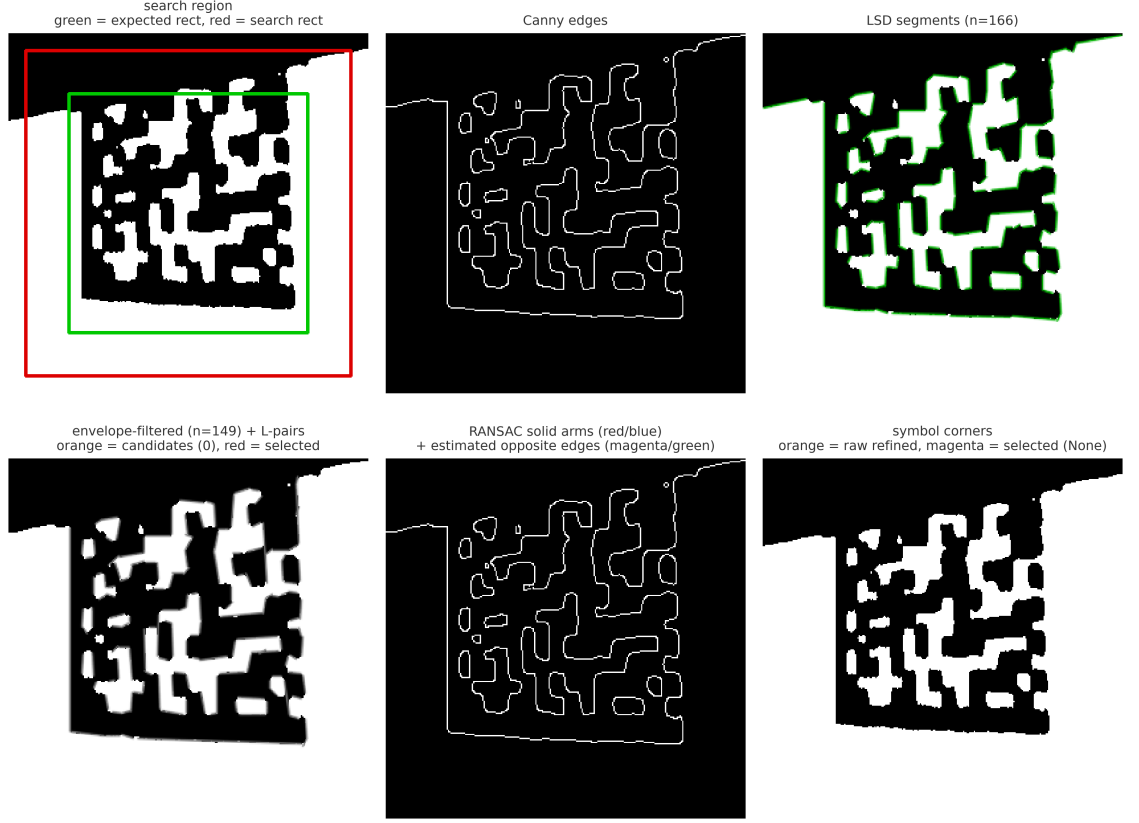


Figure 13: Unsuccessful geometric reconstruction under F3, occlusion, S2 viewpoint, and blur. Edge extraction yields many segments, but none satisfy the L-pair constraints; consequently, no quadrilateral is selected and no soft-cell evidence is produced. Source: own elaboration.

Soft-cell representation and reference ranking. After geometry is fixed, grayscale evidence is retained for reference matching. Dark and light intensity levels, I_{dark} and I_{light} , are estimated from the rectified symbol using two-cluster intensity centroids. For cell i , the mean grayscale intensity I_i over the full cell is converted to a continuous normalized darkness value:

$$p_i = \text{clip} \left(\frac{I_{\text{light}} - I_i}{I_{\text{light}} - I_{\text{dark}}}, 0, 1 \right). \quad (4)$$

The value p_i is not a calibrated probability. It preserves how strongly the observed cell resembles the symbol's dark or light intensity level. For a binary reference matrix R , the configured ranking measure is soft-Hamming similarity:

$$S_{\text{softH}}(p, R) = 1 - \frac{1}{N} \sum_{i=1}^N |p_i - R_i|, \quad N = 16^2. \quad (5)$$

Soft Hamming reduces exactly to normalized Hamming similarity when p is binary. All eight reference symbols are scored and ranked. The raw leading score is used as frame support when the candidate passes the dominance rule; no baseline rescaling and no ab-

solute score floor are applied.

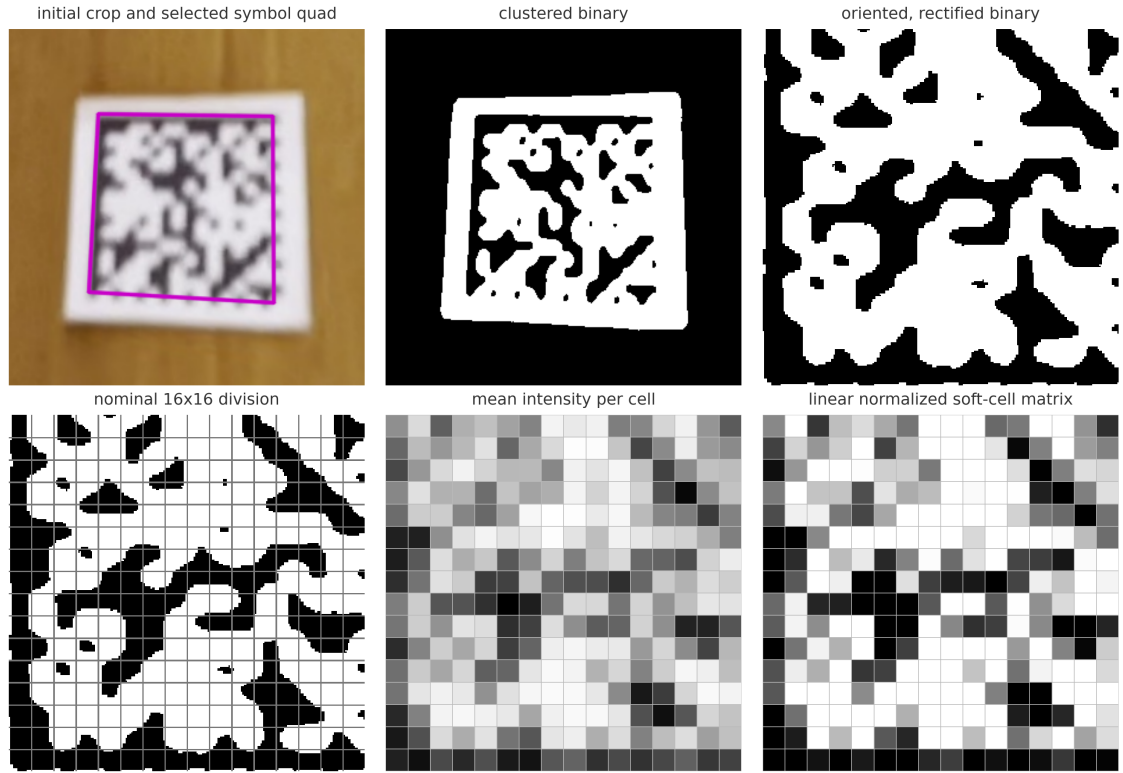


Figure 14: Soft-cell reconstruction from the selected symbol quadrilateral. The clustered binary crop is oriented and rectified, partitioned into a nominal 16×16 grid, summarized by mean intensity per cell, and linearly normalized to form the soft-cell matrix used for similarity comparison. Source: own elaboration.

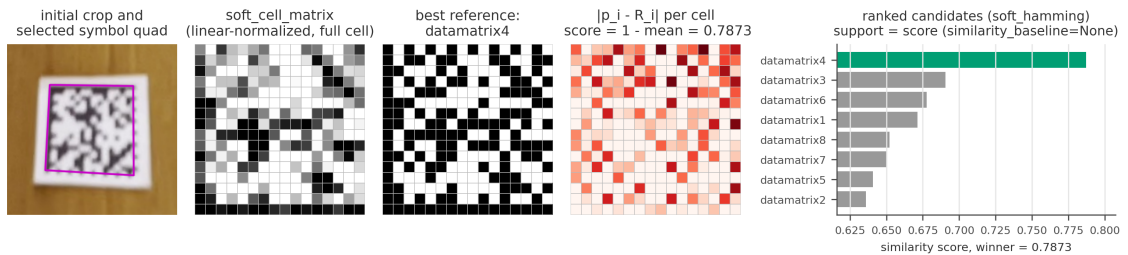


Figure 15: Soft-Hamming reference comparison for a reconstructed symbol. The linearly normalized soft-cell matrix is compared cellwise with each reference; the smallest mean absolute difference gives a similarity score of 0.7873 for *datamatrix4*, which ranks first and is forwarded to the dominance rule. Source: own elaboration.

Figures 14 and 15 connect the recovered geometry to the continuous reference evidence. The acceptance decision applied to the ranked candidates is defined in Section 6.5.1. Soft-Hamming similarity was selected over normalized Hamming and Jaccard alternatives based on a retrieval-accuracy comparison reported in Section 7.7.1.

6.5.1 Frame-level evidence and dominance rule

Rather than forwarding the ranked similarity candidates as-is, a frame-level acceptance rule is applied: the leading candidate is forwarded as the frame’s contribution to sequence-level reasoning only if it is accepted as *dominant* by the following rule. This rule is reused unchanged at the sequence level (Section 6.6). Each instance produces at most one frame-evidence pair. Stages are consulted in order: a validated primary decode, a validated secondary decode, then an accepted similarity candidate. Decode evidence has support 1.0, similarity evidence has the accepted raw soft-Hamming score, and an unresolved observation produces no pair. Consequently, primary and recovery stages cannot cast simultaneous votes for one detected instance. Let $v_1 \geq v_2 \geq \dots \geq v_k$ denote the ranked candidate values under consideration, here the similarity scores against the reference library. Candidate v_1 is accepted as dominant if

$$\text{Dominant}(v_1, \dots, v_k) = \begin{cases} \text{false}, & k = 0, \\ \text{true}, & k = 1, \\ \frac{v_1 - v_2}{v_1} \geq \theta_M, & k = 2, v_1 > 0, \\ \frac{v_1 - v_2}{v_1 - v_k} \geq \theta_Q(k), & k \geq 3, v_1 \neq v_k, \\ \text{false}, & \text{otherwise.} \end{cases} \quad (6)$$

Algorithm 1 Dominance test over ranked candidate values $v_1 \geq \dots \geq v_k$

```

1: if  $k = 1$  then
2:   return accept  $v_1$ 
3: else if  $k = 2$  then
4:    $M \leftarrow (v_1 - v_2)/v_1$ 
5:   return accept  $v_1$  if  $M \geq \theta_M$ , else reject
6: else
7:    $Q \leftarrow (v_1 - v_2)/(v_1 - v_k)$ 
8:   return accept  $v_1$  if  $Q \geq \theta_Q(k)$ , else reject
9: end if

```

The configured values are $\theta_M = 0.10$ and a 95% Dixon confidence level, giving $\theta_Q(8) = 0.526$ for the eight-reference library. Both remain uncalibrated placeholders rather than the output of a calibration pass.

Algorithm 2 formalizes the complete frame-level decision sequence for every detected instance. It complements Figures 4 and 9 by specifying the order in which mutually exclusive evidence sources are consulted. The algorithm receives the image frame, the known identifier set, and the corresponding reference matrices, and returns at most one evidence

pair for each detected instance.

Algorithm 2 Frame-level Data Matrix identification and evidence formation

Require: image frame I

Require: known identifier set $\mathcal{R}$ and reference matrices $\{R_i\}_{i \in \mathcal{R}}$

Ensure: frame-evidence list $\mathcal{E}$

```

1:  $\mathcal{E} \leftarrow$  empty list
2:  $\mathcal{D} \leftarrow \text{SegmentDataMatrixInstances}(I)$ 
3: for all detected instances  $d \in \mathcal{D}$  do
4:    $Q \leftarrow \text{ConstructValidatedQuadrilateral}(d)$ 
5:   if  $Q$  is invalid then
6:     continue
7:   end if
8:    $C \leftarrow \text{RectifyWithQuietZone}(I, Q)$ 
9:    $z \leftarrow \text{PrimaryDecode}(C)$ 
10:  if  $z \in \mathcal{R}$  then
11:     $\text{Append}(\mathcal{E}, (z, 1.0, \text{primary}))$ 
12:    continue
13:  end if
14:   $C_b \leftarrow \text{ClusterBinarize}(C)$ 
15:   $z \leftarrow \text{SecondaryDecode}(C_b)$ 
16:  if  $z \in \mathcal{R}$  then
17:     $\text{Append}(\mathcal{E}, (z, 1.0, \text{secondary}))$ 
18:    continue
19:  end if
20:   $p \leftarrow \text{RecoverSoftCellRepresentation}(C, C_b, d)$ 
21:  if  $p$  is unavailable then
22:    continue
23:  end if
24:  for all  $i \in \mathcal{R}$  do
25:     $v_i \leftarrow \text{SoftHammingSimilarity}(p, R_i)$ 
26:  end for
27:  Sort  $(i, v_i)$  pairs so that  $v_1 \geq \dots \geq v_k$ 
28:  if  $\text{Dominant}(v_1, \dots, v_k)$  then
29:     $\text{Append}(\mathcal{E}, (i_1, v_1, \text{recovery similarity}))$ 
30:  end if
31: end for
32: return  $\mathcal{E}$ 

```

The control flow enforces mutual exclusivity among evidence sources: a validated pri-

mary decode terminates processing for the instance, followed in precedence by a validated secondary decode and, finally, a dominant similarity candidate. An invalid quadrilateral, failed reconstruction, or non-dominant candidate set contributes no evidence rather than a zero-valued observation. Detection confidence remains diagnostic metadata and is not incorporated into identifier support. The resulting evidence pairs are propagated to the episode-level decision described in Section 6.6.

6.6 Temporal integration and episode-level decision

Sequence-level reasoning operates on episodes rather than treating a tracker identity as the ground-truth product identity, using the episode definition and gap rule established in Section 6.2. Consecutive visible frames remain in one episode when the number of missing frames between them does not exceed 15; a larger gap starts a new episode. Tracking may split one physical code into several track identifiers after prolonged misses, so the reviewed manifest stores all track identifiers that form each episode. These identifiers select the episode’s quadrilaterals without revealing the expected Data Matrix value to the decision layer.

The tracker associates quadrilaterals from consecutive frames using mask overlap, scale-normalized centroid proximity, and relative area consistency. For track quadrilateral Q_t and observation Q_o ,

$$d_c = \frac{\|c_t - c_o\|_2}{\sqrt{(a_t + a_o)/2}}, \quad (7)$$

$$d_a = \frac{|a_o - a_t|}{\max(a_o, a_t)}, \quad (8)$$

$$T(Q_t, Q_o) = 0.5 \text{IoU}(Q_t, Q_o) + 0.3 \max\left(0, 1 - \frac{d_c}{2.0}\right) + 0.2 \max\left(0, 1 - \frac{d_a}{1.0}\right). \quad (9)$$

A pair is matched when $T \geq 0.4$, and an unmatched track remains active for five frames. These settings are hand-set. The tracker has no appearance embedding, Kalman filter, re-identification model, or camera-motion compensation.

Unresolved observations require no separate filtering stage because they forward no support. For each candidate identifier i that does occur in frame evidence, support is accumulated across the episode and normalized:

$$A_i = \sum_{t: y_t=i} w_t, \quad s_i = \frac{A_i}{\sum_j A_j}, \quad \sum_i s_i = 1. \quad (10)$$

Here y_t is the accepted identifier and w_t its frame support. Decode observations therefore carry weight 1.0, while accepted similarity observations carry their sub-unit raw score. No additional primary-versus-recovery weighting parameter exists. The normalized episode-

level support s_i is distinct from the per-frame similarity score S_{softH} of Equation 5: the former is an across-observation share that sums to one over candidates, the latter a single observation’s raw match quality.

An episode is accepted precisely when one normalized candidate dominates:

$$\text{Dominant}(s_1, \dots, s_k). \quad (11)$$

The categorical result is therefore governed by dominance alone: *identified* when a candidate dominates, *ambiguous* when candidates exist but none dominates, and *unresolved* when no frame produced evidence. In particular, $k = 1$ is accepted regardless of absolute similarity, which remains a documented limitation. An empirically motivated minimum-evidence or absolute-score floor has not been adopted.

Algorithm 3 expresses the episode-level procedure independently of tracking and ground-truth identity. Its input is the chronological evidence produced by Algorithm 2 for the observations assigned to one episode. The procedure retains the accumulated and normalized support values together with the categorical decision so that the amount and separation of evidence remain available for subsequent evaluation.

Algorithm 3 Episode-level evidence aggregation and identification decision

Require: chronological frame evidence $\mathcal{O} = \{(y_t, w_t)\}$

Ensure: episode status and accepted identifier, if available

Ensure: accumulated support A and normalized support s

```
1: Initialize an empty mapping  $A$ 
2: for all observations  $(y_t, w_t) \in \mathcal{O}$  in temporal order do
3:   if  $y_t$  and  $w_t$  are available then
4:      $A_{y_t} \leftarrow A_{y_t} + w_t$ , with an initial value of zero
5:   end if
6: end for
7: if  $A$  is empty then
8:   return (unresolved,  $\emptyset$ ,  $A$ ,  $\emptyset$ )
9: end if
10:  $W \leftarrow \sum_j A_j$ 
11: for all candidate identifiers  $i$  in  $A$  do
12:    $s_i \leftarrow A_i / W$ 
13: end for
14: Sort  $(i, s_i)$  pairs so that  $s_1 \geq \dots \geq s_k$ 
15:  $d \leftarrow \text{Dominant}(s_1, \dots, s_k)$ 
16: if  $d$  then
17:   return (identified,  $i_1$ ,  $A$ ,  $s$ )
18: else
19:   return (ambiguous,  $\emptyset$ ,  $A$ ,  $s$ )
20: end if
```

The algorithm distinguishes absence of evidence from disagreement among observed candidates. An episode is unresolved when no frame supplies evidence. It is ambiguous when evidence exists but the normalized leader is not dominant, and identified when it is. Figure 16 provides the corresponding information-flow representation.

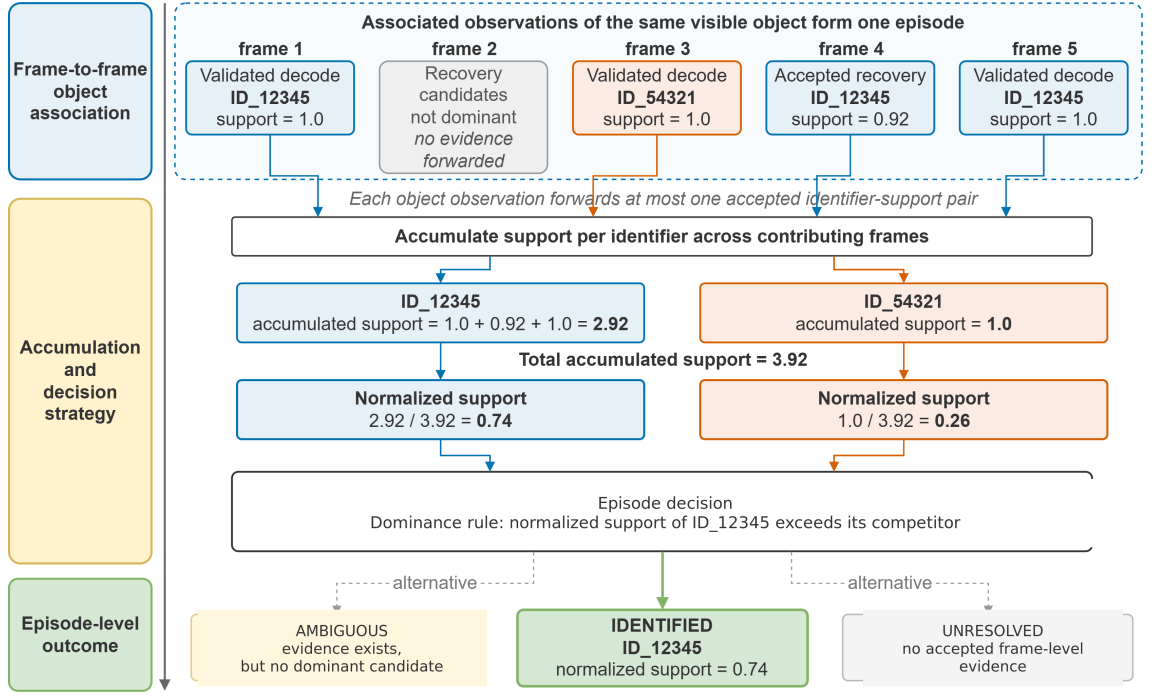


Figure 16: Episode-level evidence accumulation across associated observations. Each frame contributes at most one accepted identifier–support pair; supports are summed by candidate and normalized before the dominance rule returns identified, ambiguous, or unresolved. In the illustrated episode, conflicting frame-level candidates are resolved in favour of ID_12345. Source: own elaboration.

The episode-level process accumulates mutually exclusive frame evidence by candidate identifier, retains both accumulated and normalized support, and maps the resulting evidence state to the categorical outcomes defined in Equations 10–11.

6.7 Operational interpretation of episode-level identification outcomes

While the primary objective of the proposed system is reliable visual identification of Data Matrix-coded products, its episode-level decision outputs also admit a bounded operational interpretation. In mobile acquisition scenarios representative of HMLV-inspired intralogistics environments, an accepted identifier, an ambiguous outcome, and an unresolved outcome represent different states of evidence rather than equivalent product assertions.

The sequence-level reasoning layer introduced in Section 6.6 aggregates frame-level observations into an episode-level decision. An accepted identifier may be passed to a downstream process as evidence of observed product presence; an ambiguous or unresolved episode withholds that assertion and may instead trigger review or leave the inventory representation unchanged. This distinction provides an interface between visual identification and higher-level decision processes without defining how any downstream system

should update its state.

The present dissertation does not develop or empirically validate spatial localization, inventory-state updating, or warehouse-management integration. The operational interpretation is therefore a conceptual extension of the identification framework: it specifies the information and uncertainty states that a future integration could consume, rather than claiming a completed cyber–physical inventory subsystem.

Consequently, the proposed architecture may be viewed not only as a perception pipeline, but as a foundational layer for future integration of visual identification with spatial reasoning and operational decision support. This interpretation reinforces the system-level perspective adopted throughout the dissertation, in which robust identification is treated as a prerequisite for higher-level warehouse intelligence rather than as an isolated computer vision task.

7. RESULTS

This chapter reports the empirical evaluation of the implemented Data Matrix identification framework on the held-out test data. It is organized as an argument that builds from the components to the system. Section 7.1 fixes the evaluation units, denominators, and reporting conventions. Sections 7.2 to 7.4 then evaluate the three components in the order the pipeline applies them, establishing that detection is close to saturated, that the decoder is exact but covers only part of the data, and that the recovery branch extends coverage under a selective acceptance rule.

Section 7.5 combines the components and reports what a single frame achieves, together with how that result varies across the controlled acquisition and object-level conditions defined in Section 6.2. Section 7.6 moves to the episode, the unit at which the system actually returns an identifier, and quantifies what accumulating evidence across observations adds. Section 7.7 closes the chapter by ablating the recovery stages at both levels, so that the contribution of each mechanism, and the point at which the residual risk enters, can be read from one comparison.

Throughout, coverage and correctness are reported as separate quantities. A system that returns an identifier for most objects and a system that returns few but reliable identifiers describe different operational propositions, and the distinction is maintained deliberately in every table below.

7.1 Evaluation scope and reporting conventions

Two evaluation units are used throughout this chapter, and they answer different questions. The *annotated object* is one Data Matrix symbol visible in one frame; it is the unit for detection, decoding, and frame-level identification. The *episode* is one physical code followed through its continuous span of visibility within one video; it is the unit at which the system actually returns an identifier to a downstream process. Frame-level results therefore describe how much information a single observation carries, whereas episode-level results describe how reliably the system answers the operational question.

Accordingly, a correct identification means that the identifier returned by the pipeline matches the ground-truth item-level identifier associated with the observed symbol. It does not imply that an SKU, quantity, warehouse location, or inventory record has been inferred or updated.

Table 12 gives the composition of the frame-level evaluation set. The held-out test split contains 798 annotated frames holding 1,015 annotated Data Matrix objects, unevenly distributed across the controlled conditions: 72.7% of the objects are marked as blurred, 43.2% are seen from the oblique viewpoint category S2, and 15.0% are annotated as occluded or too small. The evaluation set is therefore weighted toward difficult acquisition

conditions rather than balanced across them, which must be kept in mind when the aggregate rates are read.

Factor	Level	Frames	Objects	Share of objects	Detection (%)
Overall	all	798	1015	100.0	98.5
Lighting	F1	169	190	18.7	98.9
Lighting	F2	341	454	44.7	98.9
Lighting	F3	288	371	36.6	97.8
Viewpoint	S1	467	577	56.8	99.5
Viewpoint	S2	355	438	43.2	97.3
Object condition	Normal	701	863	85.0	99.2
Object condition	Occluded	140	150	14.8	94.7
Object condition	Too small	2	2	0.2	100.0
Is blurred	No	241	277	27.3	99.3
Is blurred	Yes	587	738	72.7	98.2

Table 12: Composition of the frame-level evaluation set. Counts are annotated ground-truth objects; the detection column is the share of those objects matched by the detector at $\text{IoU} \geq 0.50$.

Table 13 gives the corresponding sequence-level scope. The same three test videos yield 34 episodes, distributed as 13 under lighting setup F1, 12 under F2, and 9 under F3. Across these episodes the detector produced a quadrilateral on 98.3% of the frames in which a code was annotated as visible, so an episode is rarely interrupted by missed localization. Only 22 of the 34 episodes contain at least one frame in which the annotator judged the symbol to be fully visible, which means that a third of the episodes must be resolved, if at all, from partially visible observations alone.

Lighting	Videos	Episodes	Frames	Observed frames	With a fully visible frame	Observed (%)
F1	1	13	1181	1159	7	98.1
F2	1	12	3174	3113	8	98.1
F3	1	9	2546	2510	7	98.6
All	3	34	6901	6782	22	98.3

Table 13: Composition of the sequence-level evaluation set. An episode is one product instance followed through consecutive frames. The observation rate is the share of an episode’s frames on which the detector produced a quadrilateral, and the penultimate column counts episodes containing at least one fully visible annotated frame.

A distinction between two counts is needed before any rate is read. The frame-level evaluation output contains 1,145 records, whose composition is given in Table 14. Only the 1,015 records that correspond to an annotated object, that is the true positives and the

false negatives together, carry an expected identifier, so only those form the denominator of every identification rate reported below. Unmatched predictions and empty-frame observations have no ground-truth identifier to be judged against; they are used when characterizing detection behaviour and are excluded everywhere else.

Record type	<i>n</i>	Share of records (%)
True positive (matched object)	1000	87.3
False negative (missed object)	15	1.3
False positive (unmatched prediction)	17	1.5
True negative (empty frame)	113	9.9

Table 14: Composition of the frame-level record. True positives and false negatives together form the annotated-object denominator used by every identification rate; false positives and empty-frame rows carry no reference identifier and are excluded from those rates.

The ground-truth target of this study is the unique Data Matrix identifier encoded on an annotated physical symbol instance. It is therefore a code-level identity, not a stock keeping unit (SKU), a product master-data record, a warehouse location, or an inventory-state record. A reference identifier may be linked to a product in a later operational system, but this association is not represented or evaluated in the present dataset. The implemented pipeline receives visual observations and a permitted reference-identifier set; it does not access the expected identifier, product metadata, or an inventory record at runtime.

The evaluation units follow from this boundary. For detection, the unit is an annotated Data Matrix object in one frame and correctness concerns its localization. For decoding, the unit is one localized symbol observation and the outcome is a returned character string or no result. Frame-level identification evaluates whether an accepted identifier for an annotated object matches its encoded ground-truth identifier. Episode-level identification evaluates the final decision formed from all observations assigned to one continuous visibility episode of the same annotated code. Product-level inventory inference remains an operational interpretation beyond the evaluated scope: no WMS record was queried or updated, and no correctness measure is reported for such an inference.

Two geometric input conditions are also distinguished. The principal results are obtained from quadrilaterals derived from the detector output and therefore describe the implemented pipeline as it would run. A second analysis replaces those quadrilaterals with the annotated polygons while leaving the downstream decoding and recovery logic unchanged. That second condition is a diagnostic used in Section 7.5.9 to estimate how much of the residual loss is attributable to localization; it is never reported as pipeline performance. Where a number in this chapter is not explicitly marked as the ground-truth-polygon diagnostic, it comes from the detector-derived condition.

Finally, three outcome categories recur at both evaluation levels and are kept separate throughout. An *identified* unit received an accepted identifier, which may afterwards prove correct or incorrect. An *unresolved* unit received no identifier. At episode level a third category, *ambiguous*, records that evidence accumulated for several candidates without any of them dominating. Coverage and correctness are reported separately, because a system that answers often but wrongly and a system that answers rarely but reliably fail in different ways operationally.

7.1.1 Evaluation measures

The measures used in this chapter are defined below in the same order in which the pipeline is evaluated. Unless stated otherwise, a rate is a proportion on the interval $[0, 1]$ and is multiplied by 100 when displayed as a percentage. Every conditional measure names its denominator explicitly, because detection, decoding, recovery, frame-level identification, and episode-level identification operate on different evaluation units.

1. Detection and instance-segmentation measures. A predicted instance is matched to at most one previously unmatched ground-truth instance of the Data Matrix class. Matching is based on intersection over union (IoU), which measures the overlap between a predicted region P and an annotated region G :

$$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}. \quad (12)$$

An IoU of zero means that the regions do not overlap, while an IoU of one means that they coincide exactly. For mask IoU, P and G are the predicted and annotated pixel sets; for box IoU, they are the corresponding axis-aligned bounding boxes. The main detection counts use a one-to-one match at $\text{IoU} \geq 0.50$. A matched prediction is a true positive (TP), an unmatched prediction is a false positive (FP), and an unmatched annotated object is a false negative (FN). Object detection has no natural object-level true-negative count because the number of possible background regions is not finite. The true negatives in Table 14 are therefore defined at frame level and denote negative frames on which no prediction was raised.

Detection precision, detection recall, and their harmonic mean are

$$\begin{aligned} \text{Precision}_{\text{det}} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, & \text{Recall}_{\text{det}} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F_1 &= \frac{2 \text{ Precision}_{\text{det}} \text{ Recall}_{\text{det}}}{\text{Precision}_{\text{det}} + \text{Recall}_{\text{det}}}. \end{aligned} \quad (13)$$

Precision answers how often a reported detection corresponds to an annotated symbol. Recall answers how many annotated symbols were found and is therefore also called the *detection rate* in this dissertation. F_1 is high only when both are high. For the separately reported empty-scene check, the negative-frame rejection rate and its complement are

$$R_{\text{neg}} = \frac{N_{\text{negative, no prediction}}}{N_{\text{negative}}}, \quad R_{\text{FP, frame}} = \frac{N_{\text{negative, at least one prediction}}}{N_{\text{negative}}} = 1 - R_{\text{neg}}. \quad (14)$$

Average precision (AP) summarizes the complete precision–recall curve rather than one operating point. Predictions are ordered by confidence, the confidence threshold is swept, and matching is repeated at a fixed IoU threshold τ . With the 101 recall levels $\mathcal{R}_{101} = \{0, 0.01, \dots, 1\}$, interpolated precision and AP are

$$p_{\tau}^{\text{int}}(r) = \max_{\tilde{r} \geq r} p_{\tau}(\tilde{r}), \quad \text{AP}_{\tau} = \frac{1}{101} \sum_{r \in \mathcal{R}_{101}} p_{\tau}^{\text{int}}(r). \quad (15)$$

For K detection classes, mean average precision averages AP across classes. The two variants reported here are

$$\text{mAP}_{50} = \frac{1}{K} \sum_{k=1}^K \text{AP}_{k,0.50}, \quad \text{mAP}_{50-95} = \frac{1}{10K} \sum_{k=1}^K \sum_{\tau \in \{0.50, 0.55, \dots, 0.95\}} \text{AP}_{k,\tau}. \quad (16)$$

The detector has one class, Data Matrix, so the class average is numerically identical to that class’s AP. mAP_{50} uses the relatively permissive 0.50 overlap criterion and primarily reflects whether symbols are found. mAP_{50-95} also includes increasingly strict localization criteria and therefore falls when mask boundaries or box positions are imperfect even if detection at IoU 0.50 is successful. Mean matched IoU is a different summary,

$$\overline{\text{IoU}}_{\text{match}} = \frac{1}{N_{\text{TP}}} \sum_{j=1}^{N_{\text{TP}}} \text{IoU}(P_j, G_j), \quad (17)$$

and is computed only over matched true positives. It describes localization quality conditional on successful matching and does not penalize false positives or missed objects. Consequently, it must be read together with recall and precision. In the stratified detector table, precision, F_1 , and mAP are fully interpretable for the overall and lighting partitions. Object condition, viewpoint, and blur are ground-truth-object properties to which an unmatched prediction cannot be assigned; their false-positive counts are therefore zero by construction, and only recall and matched IoU are interpreted for those rows.

2. Frame-level decoding, recovery, and identification measures. Let N_o be the number of annotated objects in the relevant frame-level population, C_o the number given a correct accepted identifier, I_o the number given an incorrect accepted identifier, and U_o the number left unresolved. Thus $N_o = C_o + I_o + U_o$. The principal frame-level measures are

$$\begin{aligned}
R_o &= \frac{C_o + I_o}{N_o}, & A_o &= \frac{C_o}{N_o}, & P_o &= \frac{C_o}{C_o + I_o}, \\
F_o &= \frac{I_o}{N_o}, & U_o^{(\text{rate})} &= \frac{U_o}{N_o}.
\end{aligned} \tag{18}$$

R_o is the frame-level identification or resolution rate: it measures coverage, irrespective of whether the returned identifier is correct. A_o is the correct-identification rate over all annotated objects. P_o is identification precision, that is, correctness conditional on the pipeline returning an identifier. F_o is the false-identification rate over all annotated objects, and $U_o^{(\text{rate})}$ is the unresolved share. Precision is undefined when no identifier is returned and is shown as -- in that case. This distinction also governs decoder reporting: the primary decode rate is the share of applicable objects for which reference validation accepts a decoded string, whereas decoder correctness is the share of those accepted strings that equal the ground-truth identifier. Silence reduces decode coverage but does not count as an incorrect accepted decode.

For any stage s , let I_s be the number of inputs reaching that stage and O_s the number producing the output named in the table. Its conditional rate and its contribution relative to a stated reference population N_{ref} are

$$q_s = \frac{O_s}{I_s}, \quad c_s = \frac{O_s}{N_{\text{ref}}}. \tag{19}$$

The conditional rate answers how often the stage succeeds once invoked. The contribution answers how much it adds to a larger population. In the recovery-stage table, N_{ref} is the recovery input; in the complete frame-level table it is all annotated objects. The denominator is therefore not inferred from the metric name but stated in each caption. Precision is reported only for stages that output an identifier; geometric reconstruction outputs a module grid and has no identification precision of its own.

The recovery diagnostics separate reconstruction, retrieval, and acceptance. With $N_{\text{rec,in}}$ reconstruction attempts, N_{grid} successful grids, N_{rank1} grids whose correct reference is ranked first, N_{acc} accepted leading candidates, and $N_{\text{acc,corr}}$ correct accepted candidates,

$$R_{\text{recon}} = \frac{N_{\text{grid}}}{N_{\text{rec,in}}}, \quad R_{\text{rank1}} = \frac{N_{\text{rank1}}}{N_{\text{grid}}}, \quad R_{\text{acc}} = \frac{N_{\text{acc}}}{N_{\text{grid}}}, \quad P_{\text{acc}} = \frac{N_{\text{acc,corr}}}{N_{\text{acc}}}. \tag{20}$$

R_{recon} measures whether usable geometry was recovered, not whether the symbol was identified. R_{rank1} measures reference retrieval before the dominance rule. R_{acc} measures the selectivity of that rule, and P_{acc} measures the reliability of what it forwards. The leading soft-Hamming similarity of Equation 5 is called *support*. A value of one denotes exact agreement with the reconstructed module grid; lower values denote increasing disagreement. The score is not a calibrated probability of correctness. Where reported, the

margin is the difference between the leading and second-ranked supports, $m = s_1 - s_2$; it quantifies separation between candidates, not absolute reconstruction quality.

Two image-space descriptors are used for stratification:

$$A_{\text{rel}}(\%) = 100 \frac{A_{\text{polygon}}}{A_{\text{frame}}}, \quad d_{\text{short}} = \min \left(\frac{\ell_1 + \ell_3}{2}, \frac{\ell_2 + \ell_4}{2} \right). \quad (21)$$

Here $\ell_1, \dots, \ell_4$ are the ordered quadrilateral side lengths and opposite sides form the two averages. A_{rel} is the percentage of the image occupied by the annotated polygon, while d_{short} is the average pixel length along the shorter quadrilateral dimension. Both describe apparent image-space scale; neither is a calibrated physical distance or object size.

3. Episode-level and temporal-decision measures. Let N_e be the number of test episodes, C_e the number ending in a correct identifier, I_e the number ending in an incorrect identifier, B_e the number ending as ambiguous, and U_e the number ending unresolved. Thus $N_e = C_e + I_e + B_e + U_e$. The episode measures are

$$\begin{aligned} R_e &= \frac{C_e + I_e}{N_e}, & A_e &= \frac{C_e}{N_e}, & P_e &= \frac{C_e}{C_e + I_e}, \\ F_e &= \frac{I_e}{N_e}, & B_e^{(\text{rate})} &= \frac{B_e}{N_e}, & U_e^{(\text{rate})} &= \frac{U_e}{N_e}. \end{aligned} \quad (22)$$

R_e is episode resolution, the share for which the decision returns an identifier. A_e is episode identification accuracy over all episodes. P_e is correctness conditional on an identifier being returned. F_e , $B_e^{(\text{rate})}$, and $U_e^{(\text{rate})}$ are the false-identification, ambiguous, and unresolved shares over all episodes. An ambiguous episode is deliberately not counted as resolved, because the system has not selected one identifier. Accuracy and precision coincide only when every episode is resolved; in the present study they differ because some episodes return no identifier.

Detection continuity within an episode is summarized by the observation rate

$$R_{\text{obs}} = \frac{N_{\text{visible frames with a detected quadrilateral}}}{N_{\text{annotated visible frames}}}. \quad (23)$$

This is a frame-level detection-continuity measure aggregated over episode frames; it is not episode identification accuracy. The composition of accumulated evidence is described by the number of observations that forward an evidence pair and by each source's support share, namely the support contributed by that source divided by total accumulated support. A validated primary or secondary decode contributes support 1.0, while accepted similarity evidence contributes its raw soft-Hamming support.

The retrospective single-frame decoding indicator used as a baseline is

$$R_{\text{ever decode}} = \frac{1}{N_e} \sum_{e=1}^{N_e} \mathbf{1} [\text{episode } e \text{ contains at least one correct primary decode}]. \quad (24)$$

This quantity is a diagnostic proxy ceiling for primary decoding, not an operational selection rule, because identifying the correct frame retrospectively requires ground truth. The contribution of temporal integration is reported as the percentage-point difference $A_e - R_{\text{ever decode}}$, together with episode-by-episode agreement, rather than as a relative percentage increase.

For an episode that contains a correct frame-level identification, latency to the first correct observation is measured in frames as

$$L_e = t_{e,\text{first correct}} - t_{e,\text{start}}. \quad (25)$$

The mean, median, and maximum of L_e are calculated only over episodes in which such an observation exists. This is the time at which correct evidence first becomes available; it is not the commitment time of the episode decision. Decision-horizon curves address the latter by re-applying the complete decision rule to the evidence contained in each prefix of length h and reporting $R_e(h)$ and $A_e(h)$ over all episodes.

For sampling stride s , every s -th observed frame is retained and the complete episode decision is re-applied. The processed-frame share is

$$C_s = \frac{N_{\text{frames retained at stride } s}}{N_{\text{frames retained at stride 1}}}. \quad (26)$$

The corresponding resolution and accuracy use Equation 22 on the subsampled decisions. C_s is an observation-count proxy for computation, not a measured reduction in wall-clock time: per-frame cost varies because some observations invoke recovery and others do not.

4. Uncertainty, descriptive summaries, and implementation cost. For a measured quantity $x_1, \dots, x_n$, let $F_n(x) = n^{-1} \sum_i \mathbf{1}[x_i \leq x]$ be its empirical cumulative distribution. The arithmetic mean is $\bar{x} = n^{-1} \sum_i x_i$, the empirical quantile is $Q_p = \inf\{x : F_n(x) \geq p\}$, the median is $Q_{0.50}$, P90 is $Q_{0.90}$, and the maximum is $\max_i x_i$. The mean describes average cost or location, while the median and P90 expose skew and upper-tail behaviour. These summaries are descriptive and do not by themselves quantify sampling uncertainty.

Uncertainty for a binomial proportion $\hat{p} = x/n$ is reported with a two-sided Wilson 95% interval. With $z = 1.96$, its limits are

$$\text{CI}_{\text{Wilson}} = \frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}. \quad (27)$$

The interval is used for proportions such as identification rate, accuracy, and precision. Its denominator follows the measure: all annotated objects or episodes for rates such as R_o and A_e , but only returned identifiers for P_o and P_e . A Wilson interval quantifies uncer-

tainty due to the finite evaluated sample; it does not correct dataset shift, dependence between frames of the same episode, label error, or model miscalibration. Episode-level intervals are consequently interpreted cautiously because only 34 episodes from three videos are available.

Runtime is measured as wall-clock duration per evaluated instance. For a stage that may be bypassed, bypassed cases are assigned zero in the unconditional profile. If the stage is invoked for n_s of n instances, its unconditional and conditional mean costs are related by

$$\bar{t}_{s,\text{uncond}} = \frac{1}{n} \sum_{i=1}^n t_{s,i} = \frac{n_s}{n} \bar{t}_{s|\text{invoked}}. \quad (28)$$

The unconditional mean measures the stage’s average contribution to total workload, while the conditional mean measures its cost when it actually runs. Mean stage times can be added when they refer to the same instances, giving $\bar{t}_{\text{end-to-end}} = \bar{t}_{\text{detection}} + \bar{t}_{\text{downstream}}$. Percentiles and maxima are not additive and are reported only where the combined duration was measured directly. Under the timed workload of approximately one evaluated instance per processed frame, sustained serial throughput is estimated as

$$Q_{\text{serial}} = \frac{1}{\bar{t}_{\text{end-to-end}}} \quad \text{processed frames per second.} \quad (29)$$

This reciprocal-of-mean estimate applies to the stated hardware, serial implementation, and observed instance density. It is a capacity estimate, not a latency guarantee and not evidence of real-time performance on other hardware, at other object densities, or under parallel execution.

7.2 Object detection

Detection determines which annotated objects enter the identification pipeline at all, so it is reported first. Performance was evaluated on the held-out test split described above, using the trained YOLO-based segmentation model of Section 6.3. Table 15 reports precision, recall, F1, mean average precision, and mean matched IoU, overall and within each controlled acquisition factor.

The detector localized 1,000 of the 1,015 annotated objects, a recall of 98.5%, at a precision of 98.3% and an mAP50 of 0.98. Performance is close to uniform across the controlled conditions. Recall stays at or above 0.98 under every lighting setup and under both blur levels, and the two clear exceptions are modest: recall falls to 0.95 for occluded objects and to 0.97 for the oblique viewpoint category S2. Twelve of the fifteen missed objects come from S2 and thirteen from blurred frames, which is the expected direction but a small absolute effect.

Factor	Level	n(gt)	n(pred)	TP	FP	FN	Precision	Recall	F1	mAP50	mAP50-95	Mean IoU
Overall	all	1015	1017	1000	17	15	0.98	0.99	0.98	0.98	0.93	0.97
Lighting	F1	190	195	188	7	2	0.96	0.99	0.98	0.98	0.92	0.97
Lighting	F2	454	457	449	8	5	0.98	0.99	0.99	0.98	0.94	0.98
Lighting	F3	371	365	363	2	8	0.99	0.98	0.99	0.97	0.92	0.97
Object condition	Normal	863	856	856	0	7	1.00	0.99	0.99	0.99	0.94	0.97
Object condition	Occluded	150	142	142	0	8	1.00	0.95	0.97	0.94	0.91	0.98
Object condition	Too small	2	2	2	0	0	1.00	1.00	1.00	1.00	0.60	0.78
Viewpoint	S1	577	574	574	0	3	1.00	1.00	1.00	1.00	0.94	0.98
Viewpoint	S2	438	426	426	0	12	1.00	0.97	0.99	0.97	0.91	0.97
Blur	Not blurred	277	275	275	0	2	1.00	0.99	1.00	0.99	0.94	0.98
Blur	Blurred	738	725	725	0	13	1.00	0.98	0.99	0.98	0.93	0.97

Table 15: Object-detection performance overall and under each controlled acquisition factor. Precision, recall and F1 are evaluated at $\text{IoU} \geq 0.50$; mAP50-95 averages 101-point interpolated average precision over IoU 0.50 to 0.95 in steps of 0.05.

One caveat applies to how Table 15 should be read. Precision, F1, and mAP are meaningful for the overall population and for the lighting groups, because those are frame-level partitions to which an unmatched prediction can still be assigned. Object condition, viewpoint category, and blur are properties of an annotated object, so an unmatched prediction cannot be placed in one of their levels and no false positive is counted there. The precision column consequently reads 1.00 for those rows by construction. Recall and mean matched IoU are the interpretable measures in that part of the table.

Localization quality matters to the stages that follow, because the decoder receives a crop cut from the predicted geometry. Two views of that quality are reported. Figure 17 shows average precision as a function of the IoU matching threshold: it stays close to 1.0 up to a threshold of approximately 0.80 and only collapses as the criterion approaches 0.95. Figure 18 shows the distribution of the achieved overlap directly. The median matched object attains an IoU of 1.000, 90.4% are localized at $\text{IoU} \geq 0.90$, and 99.7% at $\text{IoU} \geq 0.75$. The gap between mAP50 (0.98) and mAP50-95 (0.93) therefore reflects imperfect mask boundaries under the strictest criteria rather than misplaced detections.

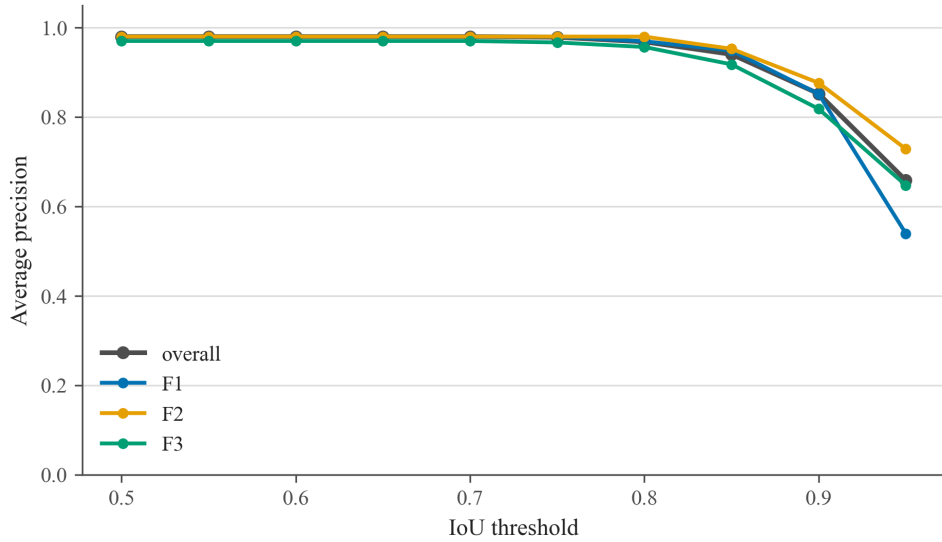


Figure 17: Detection average precision as a function of the IoU matching threshold. Source: own elaboration.

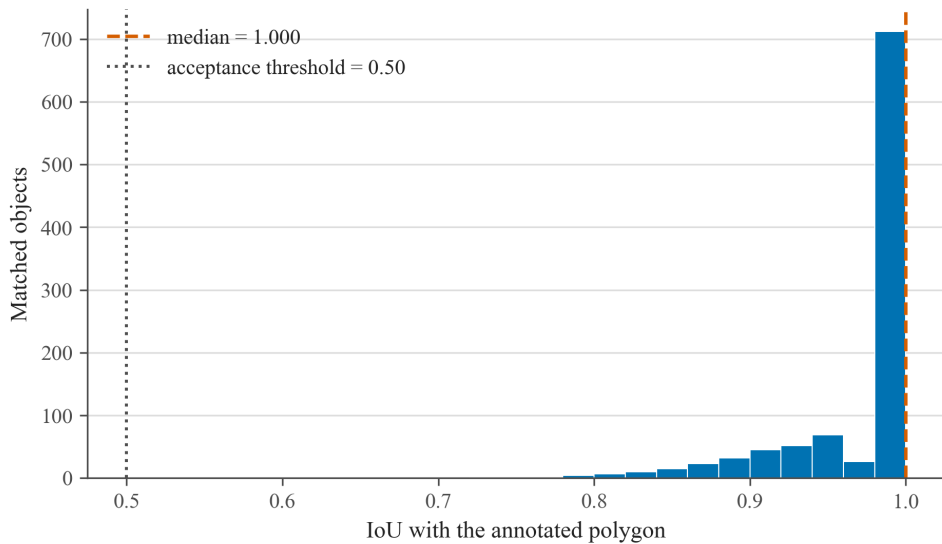


Figure 18: Distribution of the mask IoU achieved between each matched prediction and its annotated polygon. The dashed line marks the median and the dotted line the acceptance threshold used for matching. Source: own elaboration.

Because the mask supplies the quadrilateral used for rectification, it is worth asking whether the segmentation output is more accurate than the bounding box that a box detector would have produced. Over the 1,000 matched objects the two overlaps are equivalent: mask IoU and box IoU both average 0.974 with a median of 1.000, and the mean difference between them is -0.001 . The mask attains $\text{IoU} \geq 0.90$ for 90.4% of objects against 93.3% for the box. Segmentation therefore costs nothing in overlap accuracy relative to a box, while supplying the oriented boundary that the projective rectification stage requires, which is the reason the architecture uses it.

The negative scenes are reported separately, because a spurious detection in a frame containing no symbol is a different failure from a mislocalized one. Of the 922 frames in the test split, 798 contain at least one annotated Data Matrix and 124 contain none. Among the 124 negative frames, 113, or 91.1%, produced no detection at all, while 11 produced at least one spurious detection. The 17 unmatched predictions in the evaluation therefore divide into 11 raised on negative frames and 6 raised on frames that do contain an annotated symbol. Because reference validation subsequently rejects any decoded value outside the known identifier set, a spurious detection can enter the pipeline but cannot by itself introduce an identifier into the inventory logic.

Detection is therefore close to saturated on this dataset, and it is largely unaffected by blur, occlusion, viewpoint, or lighting. This is a useful result for the chapters that follow, because it means the identification losses reported next cannot be attributed to the detector: almost every annotated symbol reaches the decoding stage, and it reaches it with an accurate outline.

7.3 Primary decoding

Primary decoding was evaluated on crops cut from the annotated polygons of the same test split, which isolates the decoder from any residual localization error. A decoded value counts as correct only if it matches the object’s expected reference identifier, following the reference-validation rule of Section 6.4. Figure 19 reports the decode rate overall and within each controlled factor.

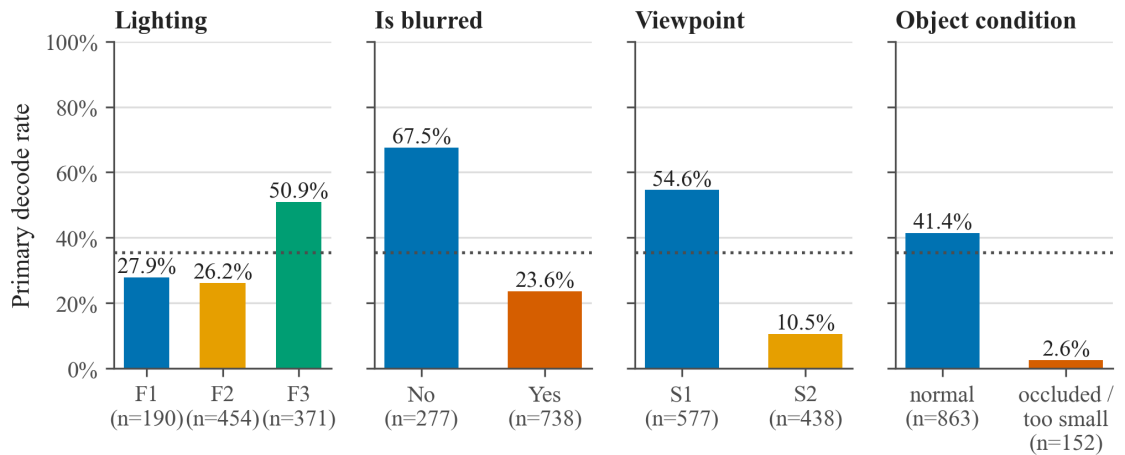


Figure 19: Primary decode rate within each controlled factor, over the 1,015 annotated objects. The dotted line marks the overall rate of 35.6%. Every value returned and accepted by reference validation was correct, so decode accuracy was 100% in every group shown. Source: own elaboration.

The decoder is exact when it succeeds and silent when it does not. Every value it returned and reference validation accepted was the correct identifier, giving a decode accuracy of

100% across all 1,015 objects. This is the property the architecture relies on: a validated decode can be treated as certain evidence, and the burden of handling uncertainty falls entirely on the stages built around the decoder rather than on the decoder itself.

Coverage is the limitation. The decoder returned a value for 35.6% of the annotated objects overall, and the marginal comparisons show where the losses concentrate. Blur is the single strongest factor, separating sharp observations at 67.5% from blurred ones at 23.6%. Viewpoint is comparable in magnitude, with 54.6% under S1 against 10.5% under S2. Restricted visibility is close to disqualifying: objects annotated as occluded or too small were decoded in 2.6% of cases. Lighting matters least of the four, although setup F3 at 50.9% is clearly ahead of F1 and F2 at 27.9% and 26.2%.

These factors are not independent, and their marginal effects should not be added. The relevant point for the design is that a dedicated Data Matrix decoder operating on a single frame recovers roughly a third of the symbols present under the acquisition variability of this dataset, while never returning a wrong identifier. The remaining two thirds define the problem that the recovery branch and the temporal decision are built to address.

The absolute distribution of reference-validated primary decodes across controlled-factor combinations is provided in Figure 48 in Appendix B; the main result is retained here as the marginal comparison in Figure 19.

7.4 Recovery branch

The recovery branch operates on the objects that the detector localized but the primary decoder did not resolve. This operational denominator excludes missed detections, because an object without a crop cannot enter the implemented recovery path. The branch is reported in its implemented order: secondary decoding of the re-binarized crop, geometric reconstruction of the module grid when secondary decoding fails, and dominance-based acceptance of a similarity candidate after successful reconstruction. Coverage and acceptance reliability are kept separate throughout, because a reconstructed symbol is not yet an identification.

Table 16 reports the stage-wise result. Of the 640 detected objects that entered recovery, secondary decoding resolved 50, and every one of these matched the expected reference identifier. The remaining 590 objects proceeded to geometric reconstruction, which produced a module grid for 574 of them, a conditional rate of 97.3%. The 16 reconstruction failures share a single recorded cause: no connected L-shape candidate inside the search region yielded four plausible symbol corners. Reconstruction is therefore not what limits the branch.

Stage	Input n	Output n	Conditional rate (%)	Contribution (%)	Precision (%)
Secondary decoding	640	50	7.8	7.8	100.0
Geometric reconstruction	590	574	97.3	89.7	—
Similarity acceptance	574	190	33.1	29.7	95.8

Table 16: Stage-wise performance of the operational recovery branch. The denominator is the set of detected objects without an accepted primary identifier. The reconstruction rate is conditional on secondary decoding having failed, while the contribution column expresses each output as a share of the recovery input. Precision is reported only for stages that output an identifier.

Table 17 collapses the branch into the four states in which an object can leave it. Recovery returned an identifier for 240 of its 640 inputs, or 37.5%, and 232 of those 240 were correct, a precision of 96.7%. Expressed against the recovery input, the branch produced a correct identifier in 36.3% of cases and an incorrect one in 1.2%, while 62.5% left the branch unresolved. The asymmetry between the correct and incorrect shares is the practically relevant property: recovery adds coverage at a small and measurable cost in false identifiers.

Terminal outcome	n	Share of recovery input (%)
Secondary identification	50	7.8
Similarity identification, correct	182	28.4
Similarity identification, incorrect	8	1.2
Unresolved	400	62.5

Table 17: Terminal outcomes of the operational recovery branch, as a share of the 640 objects entering it.

What limits the branch is candidate acceptance rather than reconstruction. Figure 20 shows where the correct reference lands in the similarity ranking of the 574 reconstructed objects. It occupies the first rank in 350 of them, or 61.0%, and is spread thinly over the lower ranks in the rest. Because the dominance test evaluates only the leading candidate, an object whose correct reference is outranked cannot be recovered even though its reconstruction succeeded. Retrieval quality and acceptance quality are therefore distinct properties, and both bound the branch.

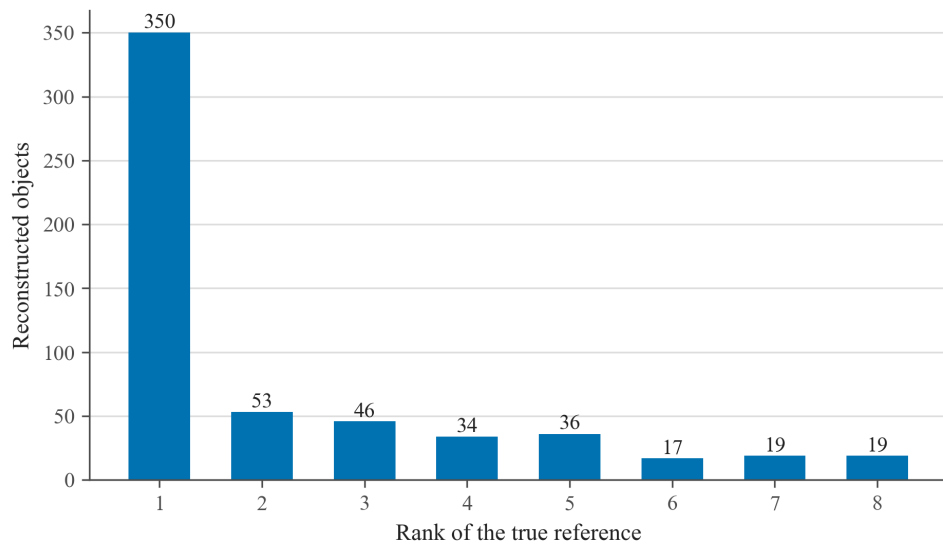


Figure 20: Rank of the ground-truth identifier among the reference candidates for the 574 successful soft-cell reconstructions. Source: own elaboration.

The dominance rule accepted 190 of those 574 leading candidates, a rate of 33.1%, of which 182 were correct and eight were incorrect. Its effect is shown in Figure 21. Forwarding every reconstructed leader without a separation test would have been correct in 61.0% of cases; requiring the leader to dominate its competitors raises correctness among the forwarded candidates to 95.8%. The rule discards roughly two thirds of the candidates in exchange for a thirty-five point gain in the reliability of what it keeps, which is the trade the architecture intends: unresolved evidence is recoverable later in the episode, whereas a wrong identifier is not.

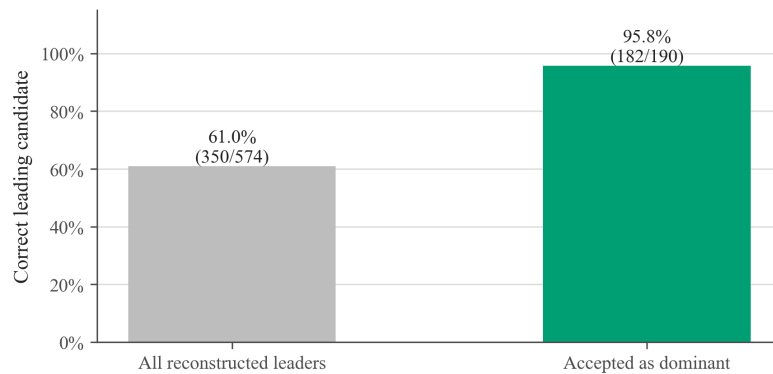


Figure 21: Correctness of the leading similarity candidate before and after the dominance test. Source: own elaboration.

Figure 22 shows the support of the leading candidate for accepted and rejected cases. The distributions overlap, but their centres are well separated: the median support was 0.721 among accepted candidates and 0.539 among rejected ones. Accepted support ranged from 0.443 to 0.875, and no accepted candidate carried zero support. The separation is

a consequence of the dominance test rather than of any fixed support threshold, since acceptance depends on the distance between the leading and the competing candidates. Support is used to weight frame-level evidence during accumulation and should not be read as a calibrated probability of correctness.

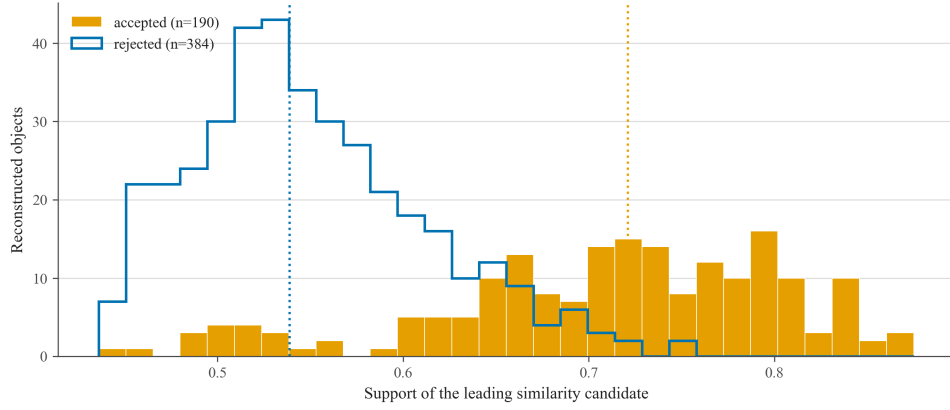


Figure 22: Support of the leading similarity candidate among reconstructed objects, separated into candidates accepted and rejected by the dominance test. Dotted lines mark the group medians. Source: own elaboration.

Factor	Level	<i>n</i>	Secondary	Similarity correct	Similarity incorrect	Unresolved
Lighting	F1	133	13.5	12.8	0.8	72.9
Lighting	F2	334	4.5	31.4	1.8	62.3
Lighting	F3	173	9.8	34.7	0.6	54.9
Is blurred	No	92	9.8	22.8	2.2	65.2
Is blurred	Yes	548	7.5	29.4	1.1	62.0
Viewpoint	S1	259	11.6	27.4	0.8	60.2
Viewpoint	S2	381	5.2	29.1	1.6	64.0
Object condition	Normal	500	9.8	35.6	0.6	54.0
Object condition	Occluded	138	0.7	2.9	3.6	92.8
Object condition	Too small	2	0.0	0.0	0.0	100.0

Table 18: Terminal recovery outcomes under controlled acquisition variability, as percentages within each subgroup of objects entering the operational recovery branch.

Table 18 and Figure 23 report the terminal outcomes within each acquisition and object-level group. Recovery was most productive under lighting setups F2 and F3, where it resolved 37.7% and 45.1% of its inputs, against 27.1% under F1. The dominant division is again visibility: recovery resolved 46.0% of the normal objects entering it but only 7.2% of the occluded ones, which remained unresolved in 92.8% of cases. Blur and viewpoint produced comparatively small differences at this stage, which is consistent with the composition of the recovery input: objects that reach recovery have already failed primary

decoding, so the most blur-affected and most oblique objects are concentrated here in both groups.

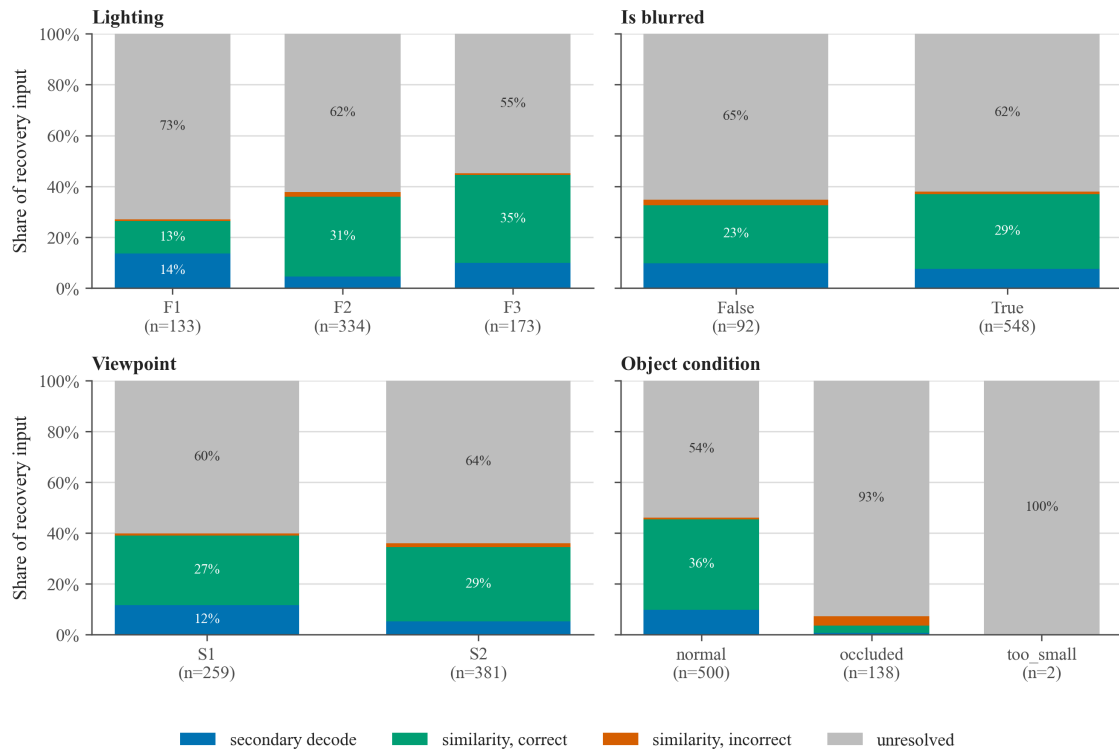


Figure 23: Terminal recovery outcomes within each controlled factor. Each bar covers the objects of one subgroup that entered the recovery branch and sums to 100%. Source: own elaboration.

The eight incorrect acceptances are examined separately in Table 19, because a wrong identifier propagates into sequence-level accumulation whereas an unresolved frame does not. They are concentrated in one group: five of the nine similarity identifiers accepted for occluded objects were incorrect, against three of the 181 accepted for normal objects. The counts are small and the table is therefore descriptive, but the pattern is consistent and it identifies where the acceptance rule is least reliable. A partially visible symbol can still yield a reconstructed grid whose leading reference separates cleanly from its competitors, even when the underlying visual evidence is insufficient to support that separation.

Factor	Level	Accepted n	Incorrect n	Incorrect among accepted (%)
Lighting	F1	18	1	5.6
Lighting	F2	111	6	5.4
Lighting	F3	61	1	1.6
Is blurred	No	23	2	8.7
Is blurred	Yes	167	6	3.6
Viewpoint	S1	73	2	2.7
Viewpoint	S2	117	6	5.1
Object condition	Normal	181	3	1.7
Object condition	Occluded	9	5	55.6

Table 19: Marginal distribution of the incorrect dominance-accepted similarity identifications. Rates are conditional on the accepted similarity identifications within each factor level.

Repeating the branch with annotated polygons instead of detector-derived quadrilaterals produces a closely comparable picture. Under ground-truth geometry, 654 objects entered recovery, secondary decoding resolved 49, and reconstruction succeeded for 590, or 90.2% of the recovery input. The correct reference ranked first in 365 of these reconstructions, or 61.9%, and the dominance rule accepted 203 candidates of which 193 were correct, a precision of 95.1%. The two geometric conditions agree on the ordering of the stages and on the position of the bottleneck, which supports reading this behaviour as a property of the reconstruction and ranking logic rather than of localization.

Figures 24–26 contrast correct, incorrect, and rejected similarity outcomes after geometric reconstruction. Correct acceptance requires the expected reference to lead and satisfy the dominance rule. Incorrect acceptance occurs when degradation makes another reference dominant, whereas unresolved cases retain evidence but fail the acceptance condition. Additional examples of reference-validated secondary decoding and failed geometric reconstruction are provided in Figures 49 and 50 in Appendix B.

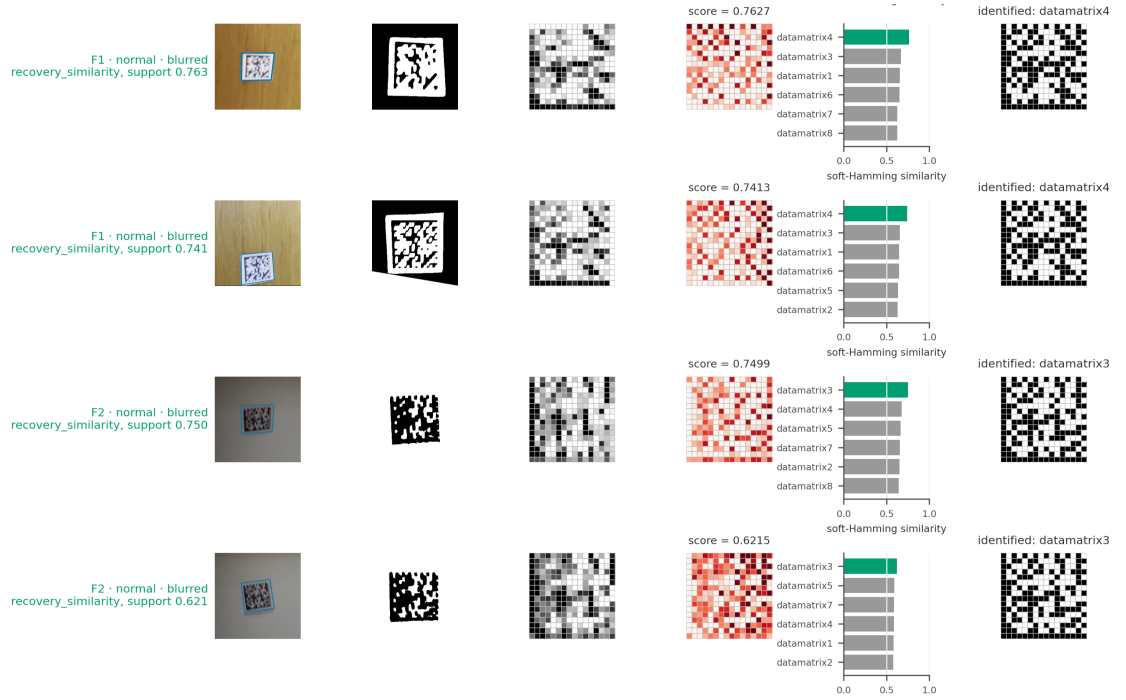


Figure 24: Examples of correct similarity-based identification after successful geometric reconstruction. Each row follows the observation from the detector-derived crop through reconstruction and reference comparison; the expected identifier ranks first and passes the dominance rule. Source: own elaboration.

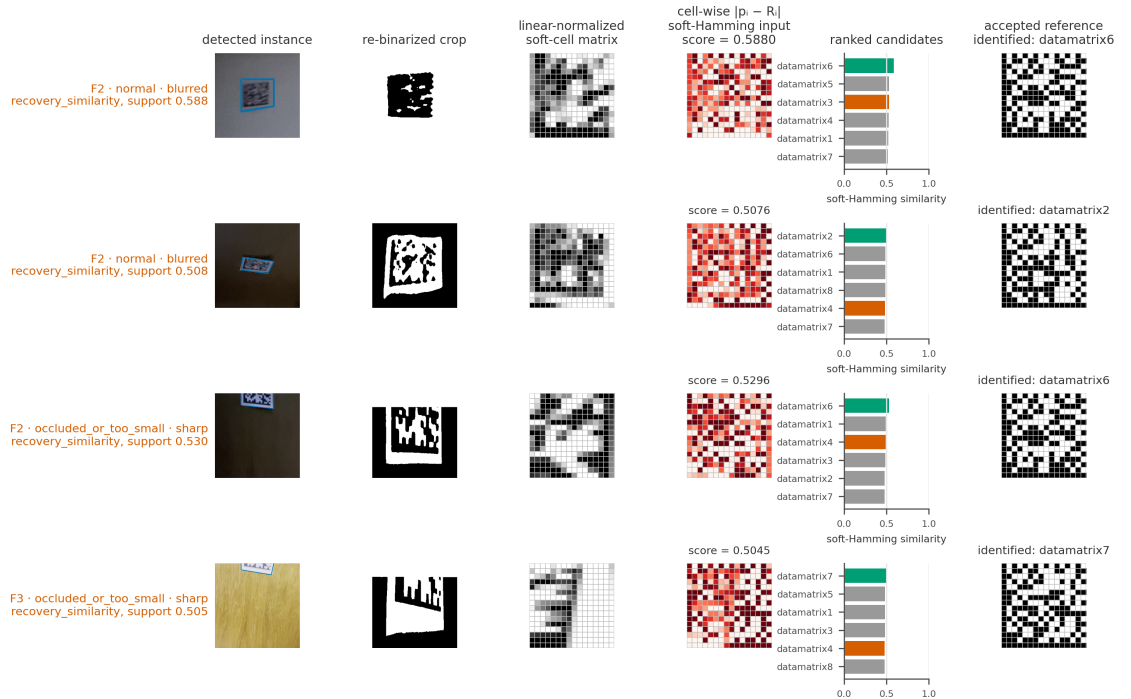


Figure 25: The eight incorrect frame-level identifications accepted by similarity search. Geometric reconstruction succeeds, but the expected reference does not rank first and an incorrect leader passes the dominance rule; these cases are concentrated in degraded, low-support observations. Source: own elaboration.

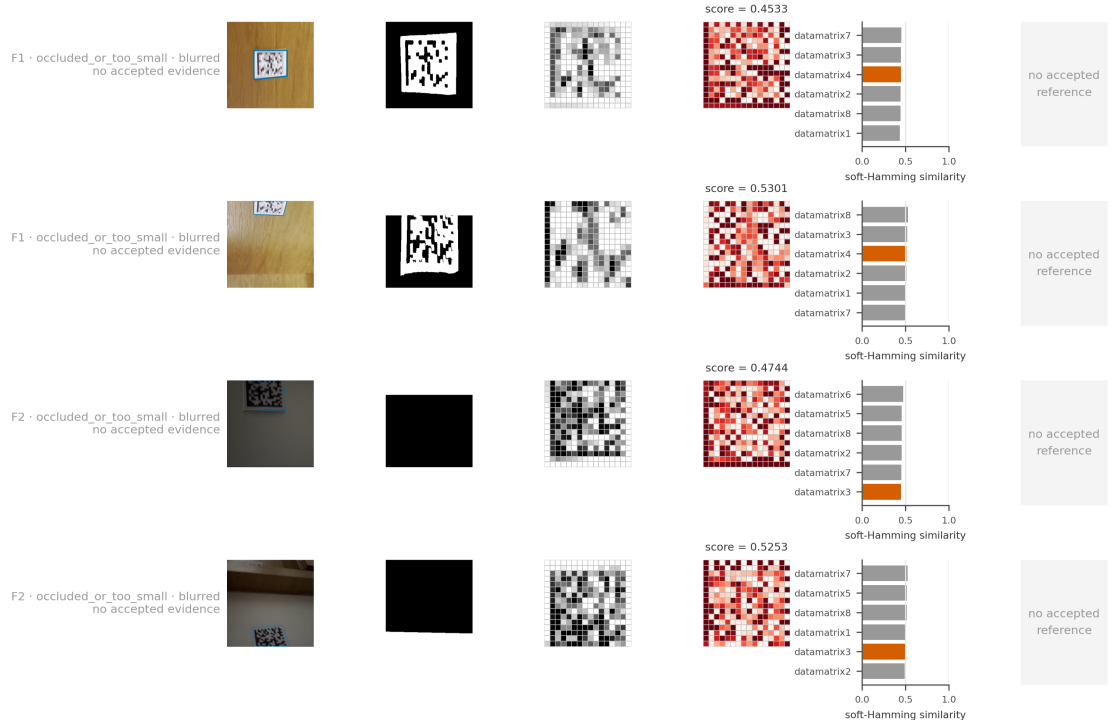


Figure 26: Examples that remain unresolved after successful geometric reconstruction. The candidate similarities are retained for analysis, but the leading reference does not satisfy the configured dominance condition, so no identifier is forwarded. Source: own elaboration.

Figure 25 visualizes the eight frame-level errors, which are listed individually in Table 20. Three properties are common to all of them. Every one was accepted through the Dixon Q branch of the dominance test, that is with three or more competing candidates rather than through the single-candidate or two-candidate branches. Every one carried a support between 0.443 and 0.588, against a median of 0.721 among accepted candidates, so all eight sit in the lower tail of the accepted range. And in none of them was the correct reference the runner-up: it occupied rank three, five, or eight in the similarity ranking, while the separation between the two leading, both incorrect, candidates averaged only 0.030.

Frame	Expected	Accepted	Conditions	Support	Margin	True rank	Branch
test_f1_f001165	datamatrix3	datamatrix1	F1, occluded, S2, blurred	0.522	0.016	3	Dixon <i>Q</i>
test_f2_f000155	datamatrix4	datamatrix6	F2, occluded, S1, sharp	0.530	0.038	3	Dixon <i>Q</i>
test_f2_f000930	datamatrix4	datamatrix2	F2, normal, S2, blurred	0.508	0.021	5	Dixon <i>Q</i>
test_f2_f001405	datamatrix4	datamatrix1	F2, occluded, S2, blurred	0.443	0.019	8	Dixon <i>Q</i>
test_f2_f001835	datamatrix4	datamatrix1	F2, occluded, S2, blurred	0.490	0.033	8	Dixon <i>Q</i>
test_f2_f002145	datamatrix4	datamatrix1	F2, normal, S2, blurred	0.461	0.031	3	Dixon <i>Q</i>
test_f2_f002670	datamatrix3	datamatrix6	F2, normal, S1, blurred	0.588	0.057	3	Dixon <i>Q</i>
test_f3_f001990	datamatrix4	datamatrix7	F3, occluded, S2, sharp	0.505	0.020	5	Dixon <i>Q</i>

Table 20: The eight incorrect frame-level identifications. Support is that of the accepted candidate, margin is its separation from the runner-up, and true rank is the position of the correct reference in the similarity ranking.

The last property is the most informative one. These are not near-misses in which the correct identifier narrowly lost. They are reconstructions so degraded that the correct reference was never competitive, and in which two equally wrong candidates happened to separate by a small margin that the relative test accepted. The dominance rule tests separation but not absolute quality, and these cases are exactly where that distinction matters.

7.4.1 Sensitivity to an absolute support floor

The observation above suggests a direct check: what would happen if acceptance additionally required the leading candidate’s support to exceed a fixed value, rather than only to dominate its competitors? Table 21 reports the effect of such a floor applied on top of the existing dominance test, holding everything else in the pipeline unchanged.

Support floor	Accepted	Correct	Incorrect	Frame identification (%)	Frame precision (%)
none	190	182	8	59.1	98.7
0.45	189	182	7	59.0	98.8
0.50	185	180	5	58.6	99.2
0.55	173	172	1	57.4	99.8
0.59	170	170	0	57.1	100.0
0.65	147	147	0	54.9	100.0
0.70	115	115	0	51.7	100.0

Table 21: Effect of an absolute support floor applied in addition to the dominance test. Accepted, correct, and incorrect counts refer to similarity-based identifications; the two rightmost columns are computed over all 1,015 annotated objects.

The trade is favourable over a wide range. A floor of 0.55 removes seven of the eight errors at a cost of ten correct identifications, and a floor of 0.59 removes all eight at a cost of twelve. Frame-level identification falls from 59.1% to 57.1%, a loss of 2.0 percentage

points, while precision among accepted identifiers rises from 98.7% to 100%. Beyond 0.59 the cost grows quickly with no further benefit, since no errors remain to remove.

Two qualifications are essential. The floor was selected by inspecting the evaluation data, so it is an in-sample value and its apparent optimality is not evidence that 0.59 would generalize; establishing a defensible threshold would require a separate calibration set. The analysis is also reported as a sensitivity check on the acceptance rule rather than as an adopted configuration: the results throughout this chapter describe the implemented pipeline, which applies no absolute floor. What the check does establish is that the residual frame-level error is concentrated in a region of the support scale that a single additional criterion can isolate, which is a more actionable finding than the error rate alone.

7.5 Frame-level pipeline

The preceding sections evaluated the stages in isolation. This section combines them and reports what the complete frame-level pipeline achieves on a single observation, using detector-derived geometry throughout. An object counts as identified when the pipeline forwards an identifier from any stage; whether that identifier was right is reported separately as precision.

Figure 27 gives the resulting composition over all 1,015 annotated objects. Primary decoding resolved 35.5%, and the recovery branch raised the identified proportion to 59.1%, leaving 40.9% unresolved. The pipeline therefore identifies roughly two thirds more objects per frame than the decoder alone.

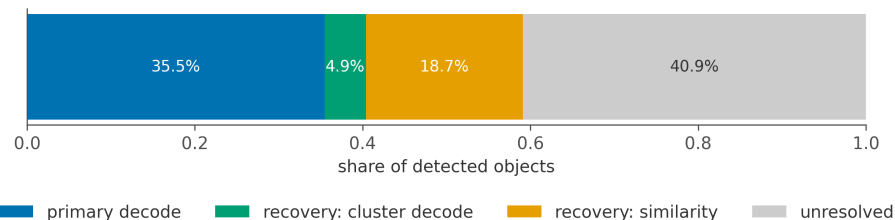


Figure 27: Composition of frame-level outcomes over the 1,015 annotated Data Matrix objects. Source: own elaboration.

Table 22 states the end result as shares of the annotated objects, and Figure 28 traces the same objects stage by stage. Detection passed 1,000 of the 1,015 objects onward. Primary decoding resolved 360 of those, a conditional rate of 36.0%. Of the 640 objects it left, secondary decoding resolved a further 50, a conditional rate of 7.8%, and the similarity stage received 574 reconstructed candidates and accepted 190, a conditional rate of 33.1%. Each stage therefore works on a progressively harder residue, which is why the conditional rates fall while the cumulative coverage rises.

Outcome or stage	<i>n</i>	Share of annotated objects (%)	Precision (%)
Annotated objects	1,015	—	—
Detected	1,000	98.5	—
Primary identification	360	35.5	100.0
Recovery contribution	240	23.6	—
Final identification	600	59.1	98.7
False identification	8	0.8	—
Unresolved	415	40.9	—

Table 22: Overall frame-level identification performance. Detection denotes entry into the downstream identification pipeline, and the recovery contribution combines secondary decoding with accepted similarity-based identification. Precision is conditional on an accepted identifier.

The diagram reports each transition both against its parent stage and against the complete set of annotated objects, which makes the terminal states directly comparable: the width of the unresolved path against the widths of the three identifying paths, and the width of the incorrect path against all of them.

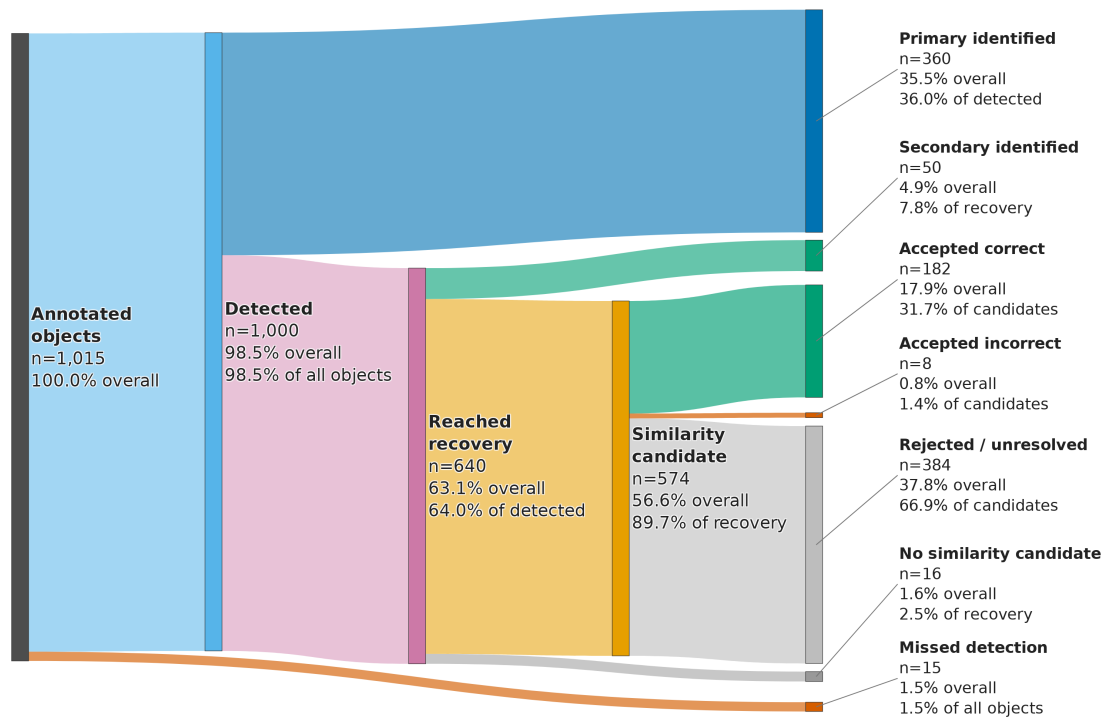


Figure 28: Flow of the 1,015 annotated objects through detection, primary decoding, secondary decoding, reconstruction, and similarity acceptance. Recovery raises identified coverage from 35.5% to 59.1%; eight accepted similarity candidates are incorrect, while 40.9% of annotated objects remain unresolved. Source: own elaboration.

Over the complete annotated-object denominator the pipeline produced an identifier for

600 objects, a frame-level identification rate of 59.1%. Of those outputs 592 were correct and eight were incorrect, so precision among identified objects was 98.7% and the false-identification rate over all annotated objects was 0.8%. The remaining 415 objects, the 15 missed detections included, forwarded no identifier and are treated as unresolved.

Two properties of this result matter for the design. First, the recovery branch raised identification from 35.5% to 59.1%, an absolute gain of 23.6 percentage points, and it did so while precision remained above 98%. Coverage was therefore bought at a small and quantified cost rather than by relaxing the acceptance criterion. Second, the unresolved share of 40.9% shows that a single frame frequently remains insufficient under the acquisition variability of this dataset. That residue is not a failure of the frame-level design so much as a statement of what one observation can carry, and it is the evidence gap the episode-level decision is built to close.

Figure 29 isolates the ceiling. Under the most favourable combination of conditions, lighting setup F3 with normal object condition, viewpoint category S1 and no blur, primary decoding alone resolved 98.2% of the 110 objects and recovery resolved the remainder, leaving nothing unresolved. The pipeline is therefore not limited by its decoding or reasoning logic; under favourable acquisition it saturates. What the aggregate rates measure is the cost of the acquisition variability that the dataset deliberately contains.

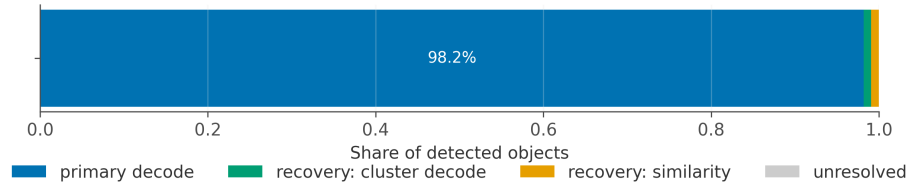


Figure 29: Composition of frame-level outcomes for the most favourable condition combination: lighting setup F3, normal object condition, viewpoint category S1, and no blur ($n = 110$). Source: own elaboration.

The next subsections quantify that cost, reporting how the outcome composition changes across the controlled conditions.

7.5.1 Performance under controlled variability

Table 23 reports primary decoding and final frame-level identification rates across the joint combinations of lighting, viewpoint, blur, and object condition. Reading the table by its margins gives a consistent ordering. Identification was highest under lighting setup F3, for normal objects, under viewpoint category S1, and for observations not marked as blurred. It was lowest, at 9.2%, for objects classified as occluded or too small. Because these conditions co-occur and the joint cells are neither balanced nor experimentally randomized, the table provides conditional descriptive comparisons rather than estimates of independent effects or factor interactions.

Object condition	Lighting	Viewpoint	Sharp		Blurred		Total	
			<i>n</i>	P (R)	<i>n</i>	P (R)	<i>n</i>	P (R)
Normal	F1	S1	19	100.0% (100.0%)	68	26.5% (52.9%)	87	42.5% (63.2%)
		S2	22	59.1% (81.8%)	41	7.3% (34.1%)	63	25.4% (50.8%)
	F1 Total		41	78.0% (90.2%)	109	19.3% (45.9%)	150	35.3% (58.0%)
	F2	S1	29	93.1% (100.0%)	175	44.0% (73.7%)	204	51.0% (77.5%)
		S2	13	23.1% (69.2%)	163	4.3% (42.3%)	176	5.7% (44.3%)
	F2 Total		42	71.4% (90.5%)	338	24.9% (58.6%)	380	30.0% (62.1%)
	F3	S1	110	98.2% (100.0%)	123	52.0% (74.0%)	233	73.8% (86.3%)
		S2	30	36.7% (80.0%)	70	8.6% (54.3%)	100	17.0% (62.0%)
	F3 Total		140	85.0% (95.7%)	193	36.3% (66.8%)	333	56.8% (79.0%)
	Normal Total			223	81.2% (93.7%)	640	27.3% (58.9%)	863
Occluded or too small	F1	S1	3	33.3% (33.3%)	13	0.0% (0.0%)	16	6.2% (6.2%)
		S2	4	0.0% (0.0%)	20	5.0% (15.0%)	24	4.2% (12.5%)
	F1 Total		7	14.3% (14.3%)	33	3.0% (9.1%)	40	5.0% (10.0%)
	F2	S1	5	0.0% (20.0%)	25	4.0% (4.0%)	30	3.3% (6.7%)
		S2	17	0.0% (0.0%)	27	0.0% (11.1%)	44	0.0% (6.8%)
	F2 Total		22	0.0% (4.5%)	52	1.9% (7.7%)	74	1.4% (6.8%)
	F3	S1	1	0.0% (0.0%)	6	0.0% (16.7%)	7	0.0% (14.3%)
		S2	24	4.2% (16.7%)	7	0.0% (0.0%)	31	3.2% (12.9%)
	F3 Total		25	4.0% (16.0%)	13	0.0% (7.7%)	38	2.6% (13.2%)
	Occluded or too small Total			54	3.7% (11.1%)	98	2.0% (8.2%)	152
Grand Total			277	66.1% (77.6%)	738	24.0% (52.2%)	1015	35.5% (59.1%)

Table 23: Decoding performance under controlled acquisition conditions. P denotes the primary decoding rate and the bold value in parentheses (R) denotes the final resolved rate after application of the recovery branch. All rates are reported as percentages.

Frame-level identification therefore remained sensitive to visibility, scale, viewpoint, illumination, and blur even after the recovery branch. Ranked by the size of the gap they produce, the factors order as follows. Object condition separates most strongly: normal objects were identified at 67.9% against 9.2% for objects classified as occluded or too small, a difference of 58.7 percentage points. Viewpoint follows, with 72.4% under S1 against 41.6% under S2, a difference of 30.8 points. Blur is close behind, at 77.6% against 52.2%, a difference of 25.4 points. Lighting produces the smallest spread of the four, from 47.9% under F1 to 72.2% under F3.

Table 24 isolates the interaction between the strongest factor and blur, and it shows that the two act differently. Among normal objects, blur reduced identification from 93.7% to 58.9%, a difference of 34.8 percentage points, while precision stayed at or near 100%. Among objects annotated as occluded or too small, identification was low irrespective of blur, at 11.1% for sharp and 8.2% for blurred observations, and precision fell to 66.7% and

62.5%. Blur therefore mainly removes coverage, whereas insufficient visibility degrades coverage and the reliability of the few identifiers that are still accepted. That second effect is the operationally important one, because it is the condition under which the pipeline is most likely to answer and be wrong.

Object condition	Is blurred	Objects	Primary (%)	Recovery (%)	Identified (%)	False ID (%)	Precision (%)	Unresolved (%)
Normal	No	223	81.2	12.6	93.7	0.0	100.0	6.3
Normal	Yes	640	27.3	31.6	58.9	0.5	99.2	41.1
Occluded or too small	No	54	3.7	7.4	11.1	3.7	66.7	88.9
Occluded or too small	Yes	98	2.0	6.1	8.2	3.1	62.5	91.8

Table 24: Frame-level identification by object condition and blur. Rates are percentages of annotated objects within each subgroup, except precision, which is conditional on an accepted frame-level identifier.

These comparisons describe conditional performance within the evaluation dataset and should not be read as isolated causal effects. Several acquisition characteristics co-occurred, and the number of observations differed across groups. Blur, viewpoint, apparent symbol area, and object condition may jointly determine whether the decoder or the recovery branch receives sufficient visual structure, and the design does not permit those contributions to be separated.

The subsections that follow present each factor as an outcome composition rather than as a single rate, so that the share contributed by primary decoding, by recovery, and by the unresolved outcome can be compared within every group. They are ordered as lighting, blur, object condition, viewpoint, and relative symbol area.

7.5.2 Effect of lighting

The three lighting setups produced distinct outcome compositions, as shown in Figure 30. The final identification rate was 47.9% under F1, 53.1% under F2, and 72.2% under F3. Primary decoding accounted for a progressively larger share of outcomes from F1 to F3, whereas unresolved observations declined accordingly. Recovery increased coverage under every lighting setup, but did not remove the performance differences between them. The lighting-specific rates reflect the complete set of observations acquired under each setup rather than a balanced factorial comparison. Their interpretation must therefore account for the associated distribution of blur, viewpoint, object condition, and relative symbol area. Within this limitation, the results show that the acquisition setup remained materially associated with both direct decodability and final frame-level resolution.

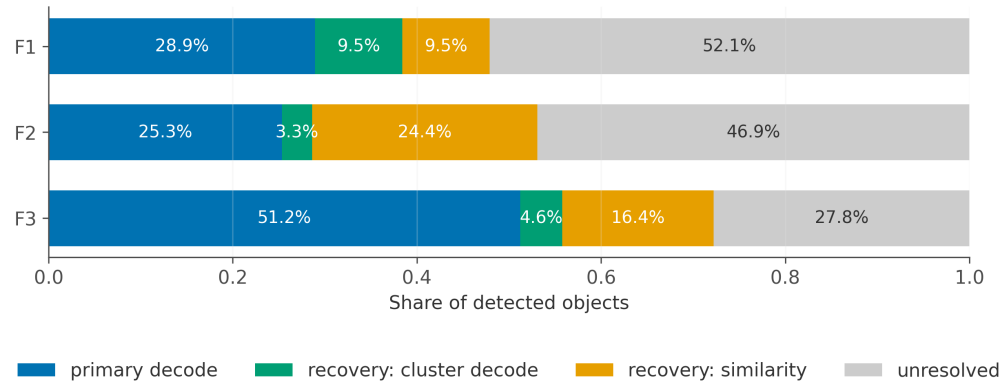


Figure 30: Composition of frame-level outcomes by lighting setup. Each bar represents the annotated objects acquired under the corresponding setup and sums to 100%. Source: own elaboration.

7.5.3 Effect of blur

Figure 31 separates observations according to the manually reviewed blur annotation. Sharp observations achieved a final identification rate of 77.6%, compared with 52.2% among observations marked as blurred. The difference was already pronounced at the primary decoding stage, which resolved 66.1% of sharp observations and 24.0% of blurred observations. Recovery narrowed this gap by adding a larger absolute contribution among blurred observations, but more than two fifths of the blurred objects remained unresolved. This pattern is consistent with the dependence of both dedicated decoding and soft-cell reconstruction on recoverable module structure. The result describes the association observed in the test set; it does not isolate motion blur from correlated factors such as view-point, apparent symbol area, or partial visibility.

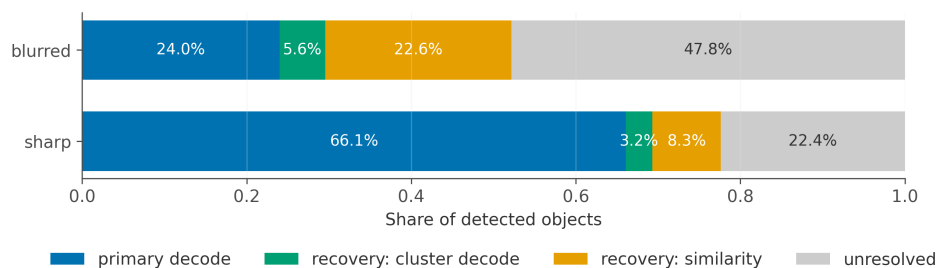


Figure 31: Composition of frame-level outcomes for sharp and blurred observations. Blur status was assigned through manual review of the annotation suggestion. Source: own elaboration.

7.5.4 Effect of object condition

Object condition produced the largest separation among the reported binary groupings. As shown in Figure 32, normal objects reached a final identification rate of 67.9%, whereas

objects classified as occluded or too small reached 9.2%. Primary decoding resolved only 2.6% of the latter group, and recovery increased the final rate by a further 6.6 percentage points. Consequently, 90.8% of the observations in this condition group remained unresolved.

The combined category represents cases in which the visible symbol information was judged insufficient because of occlusion or scale. The low recovery rate indicates that the branch can exploit degraded but still structured observations, but it cannot compensate when too little of the Data Matrix pattern is available for reconstruction and reference comparison.

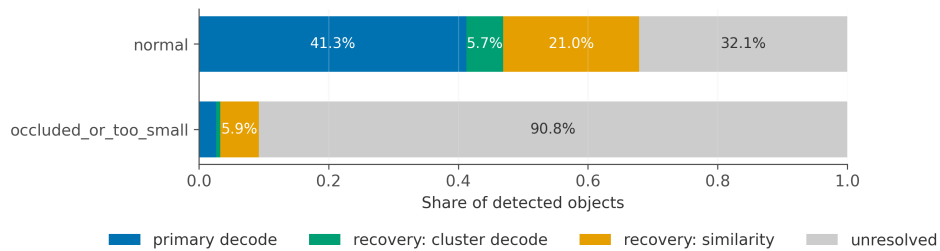


Figure 32: Composition of frame-level outcomes by object condition. The occluded-or-too-small category combines observations with insufficient visible symbol information. Source: own elaboration.

7.5.5 Effect of viewpoint angle

The viewpoint grouping also showed a substantial difference in identification performance. Figure 33 reports a final identification rate of 72.4% for S1 and 41.6% for S2, a difference of 30.8 percentage points. Recovery contributed to both groups, but the S2 observations retained a markedly larger unresolved share.

Viewpoint category was suggested from the geometry of the annotated object quadrilateral and subsequently reviewed manually. The grouping therefore provides a reproducible descriptive indicator of perspective severity rather than a calibrated physical angle. Its association with outcome composition supports the role of geometric distortion as an important source of difficulty in the evaluated frame-level pipeline.

7.5.6 Effect of relative symbol area

The distribution of final outcomes by relative symbol area is reported in Table 25. Primary decoding accounted for an increasing share of outcomes as the symbol occupied a larger image area, rising from 5.9% in the smallest category to 83.7% in the largest. Similarity-based recovery contributed most strongly in the two smallest area categories and in the 1.00–1.99 category.

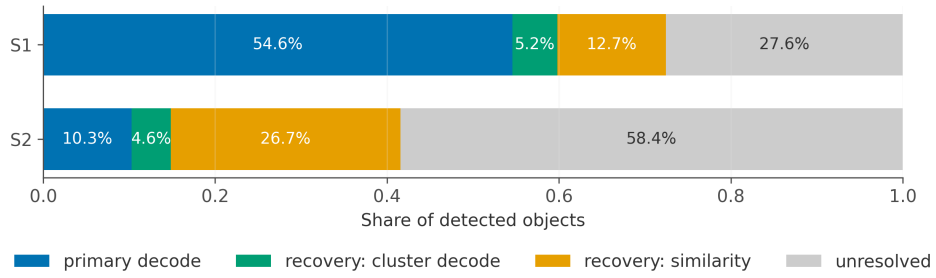


Figure 33: Composition of frame-level outcomes by viewpoint category. S1 and S2 denote the manually reviewed viewpoint groups derived from annotated quadrilateral geometry. Source: own elaboration.

Table 25: Composition of frame-level outcomes by relative Data Matrix area.

Relative area category	<i>n</i>	Primary	Secondary	Similarity	Unresolved
< 0.25	85	5.9%	0.0%	31.8%	62.4%
0.25–0.49	285	24.2%	6.3%	23.5%	46.0%
0.50–0.99	280	37.9%	5.0%	13.6%	43.6%
1.00–1.99	168	31.5%	6.0%	23.8%	38.7%
2.00–4.99	154	59.1%	4.5%	11.7%	24.7%
≥ 5.00	43	83.7%	2.3%	0.0%	14.0%

Figure 34 complements the aggregated values by showing the outcome composition overall and within each lighting setup. Relative area is calculated as the annotated Data Matrix polygon area divided by the complete frame area and is therefore an image-space scale measure rather than a calibrated physical distance.

The unresolved proportion generally declined as relative area increased, from 62.4% in the smallest category to 14.0% in the largest. The progression was not strictly monotonic across the intermediate categories, which is consistent with the interaction between apparent symbol size and the other controlled conditions. The results therefore support relative area as an important descriptive factor, but do not establish a single area threshold that independently determines identification success.

7.5.7 Absolute symbol size and the decoding operating range

Relative area describes how much of the frame a symbol occupies, which depends on the camera as much as on the symbol. The operationally useful question is different: how many pixels the symbol must span before it can be read. This is answered here using the shorter side of the annotated quadrilateral in pixels, measured as the average length of the shorter pair of opposite edges, which is the same quantity the annotation uses for its automatic size rule. Across the 1,015 annotated objects it ranges from 19 to 534 pixels,

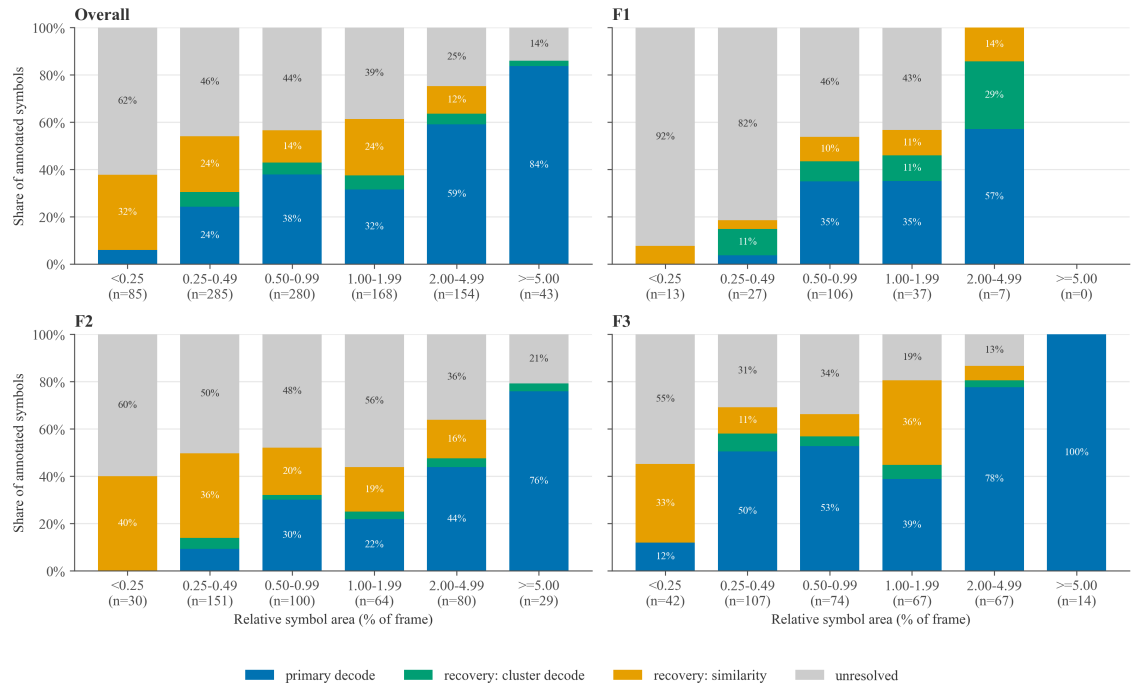


Figure 34: Frame-level identification outcome by relative symbol area, overall and within each lighting setup. Each bar presents the outcome percentages within one area bin; empty bins contain no annotated objects. Source: own elaboration.

with a median of 104.

Table 26 reports the outcome composition by this measure, both over all objects and over the normal objects alone. The second view matters because occlusion and small size are separate causes of failure that the combined view conflates: an occluded symbol may be large in pixels and still unreadable.

Shorter side (px)	All annotated objects					Normal objects only			
	<i>n</i>	Primary	Recovery	Identified	Unresolved	<i>n</i>	Primary	Identified	Unresolved
< 40	18	0.0	5.6	5.6	94.4	0	—	—	—
40–59	86	4.7	23.3	27.9	72.1	67	6.0	35.8	64.2
60–79	160	5.6	34.4	40.0	60.0	133	6.8	45.9	54.1
80–99	211	37.0	25.1	62.1	37.9	188	40.4	68.1	31.9
100–139	246	37.4	24.4	61.8	38.2	211	43.1	69.7	30.3
140–199	131	44.3	25.2	69.5	30.5	117	48.7	76.9	23.1
≥ 200	163	73.0	11.0	84.0	16.0	147	81.0	92.5	7.5

Table 26: Frame-level outcome composition by the shorter side of the annotated symbol in pixels. All values are percentages within the row. The right-hand block restricts the same analysis to objects annotated as normal, separating apparent size from occlusion.

The clearest feature is a threshold in the behaviour of primary decoding rather than a gradual decline. Among normal objects, the dedicated decoder resolves 6.0% and 6.8%

of symbols in the two bins below 80 pixels, and 40.4% in the bin immediately above. The decoder is therefore close to inoperative below roughly 80 pixels on the shorter side and becomes productive above it, which corresponds to about five pixels per module for the 16 by 16 symbols used here. Above that point its rate continues to climb with size, reaching 81.0% for symbols spanning 200 pixels or more.

Recovery behaves in the opposite direction, and this is what keeps the pipeline useful in the range where the decoder is not. Its contribution is largest between 40 and 100 pixels, where it resolves between a quarter and a third of the objects that primary decoding cannot, and it falls to 11.0% above 200 pixels simply because little is left for it to do. The two stages therefore cover complementary parts of the size range rather than the same part.

A lower bound is also visible. Below 40 pixels on the shorter side, that is fewer than about 2.5 pixels per module, 17 of the 18 objects were unresolved and the single identifier produced was incorrect. This is the only size band in which the pipeline is more likely to be wrong than right when it answers at all, and it provides empirical support for the annotation rule that classifies objects below this size as too small to be treated as decodable. In every band above 40 pixels precision remained at or above 95%.

These values describe an operating range for the present acquisition setup and symbol size, and they should not be transferred to other configurations without re-measurement. Pixel extent is a proxy for the resolution actually available on the symbol surface, which also depends on optics, sensor, and motion. What the analysis supports is the narrower and more useful statement that within this dataset the practical requirement for direct decoding is approximately 80 pixels across the symbol, and that the recovery branch extends usable identification roughly 40 pixels below that.

7.5.8 Frames containing more than one symbol

Several test frames show two Data Matrix symbols at once, which raises a distinct question: whether the pipeline can keep two simultaneously visible identities apart, or whether evidence from one leaks into the other. Of the 798 frames containing at least one annotated symbol, 581 contain one and 217 contain two, so 434 of the 1,015 annotated objects were observed alongside a second symbol.

Table 27 compares the two groups. Identification was lower on frames containing two symbols, at 54.8% against 62.3%, which is consistent with the acquisition geometry rather than with confusion between the symbols: when two codes are in view, each tends to be smaller and more obliquely presented than when one fills the working area. Primary decoding accounts for the whole of the difference, falling from 41.5% to 27.4%, while the recovery branch contributes more on the two-symbol frames, at 27.4% against 20.8%.

Frame content	Objects	Primary (%)	Recovery (%)	Identified (%)	Precision (%)	False ID (%)
One symbol visible	581	41.5	20.8	62.3	98.3	1.0
Two symbols visible	434	27.4	27.4	54.8	99.2	0.5

Table 27: Frame-level identification for objects observed alone and for objects observed alongside a second Data Matrix symbol.

Reliability does not degrade in the presence of a second symbol. Precision was 99.2% on the two-symbol frames against 98.3% on the single-symbol frames, and two of the eight incorrect identifications occurred on two-symbol frames against six on single-symbol frames. Errors are therefore not concentrated where confusion between simultaneously visible symbols would place them.

Two further observations support the same conclusion. Across the 527 frames on which the pipeline produced at least one identifier, there was no frame in which two objects were assigned the same identifier, so the association never collapsed two distinct symbols onto one identity. More tellingly, none of the eight incorrect identifiers was the identifier of the other symbol present in the frame: every one named a reference that was not physically present in the test scene at all. The errors are therefore failures of reconstruction against the reference library rather than failures to distinguish one visible symbol from another. This evidence is specific to the two-symbol case that the dataset contains, and it does not establish behaviour for scenes with many symbols or for symbols whose printed content is nearly identical. Within the evaluated scope, however, simultaneous visibility affected how often the pipeline could answer, not whether its answers were right.

7.5.9 Effect of localization on identification

The ground-truth-polygon analysis identified 60.4% of the annotated objects, compared with 59.1% for detector-derived quadrilaterals, as shown in Table 28. Replacing the implemented localization output with the annotated geometry therefore increased the identification rate by 1.3 percentage points. The corresponding false-identification rates were 1.0% and 0.8%, respectively.

Table 28: Comparison of downstream identification results by polygon source.

Polygon source	Identification rate	False-identification rate	Precision among identified objects
Detector-derived quadrilateral	59.1%	0.8%	98.7%
Ground-truth polygon	60.4%	1.0%	98.4%

In conjunction with the 98.5% detection rate, the limited difference between the two polygon sources suggests that localization was not the principal source of unresolved frame-level cases in this evaluation. Most of the remaining loss occurred after localization, where visual degradation continued to limit direct decoding, reconstruction, or the acceptance of a sufficiently dominant reference candidate. This interpretation is restricted to the evaluated data and does not imply that polygon quality is inconsequential under acquisition conditions outside the present dataset.

7.5.10 Summary of frame-level findings

The complete frame-level pipeline raised identification from 35.5% under primary decoding alone to 59.1% after recovery, an absolute gain of 23.6 percentage points, while precision among accepted identifiers remained at 98.7%. Under the most favourable acquisition conditions the pipeline saturated: primary decoding alone resolved 98.2% of objects and recovery accounted for the remainder. Across the stratified factors, object condition produced the largest separation (67.9% for normal objects against 9.2% for occluded or too-small objects), followed by viewpoint (30.8 percentage points), blur (25.4 percentage points), and lighting (the smallest spread). Below roughly 80 pixels on the symbol's shorter side primary decoding became largely inoperative, and recovery extended usable identification down to about 40 pixels. Replacing the detector-derived quadrilateral with the annotated ground-truth polygon changed the identification rate by only 1.3 percentage points, indicating that localization was not the principal source of unresolved cases.

The residual weakness concentrates where visibility is already lowest. Precision among accepted identifiers fell to 62.5–66.7% for objects annotated as occluded or too small, against 99.2–100.0% for normal objects, so the few identifiers produced under severe visibility loss are also the least reliable. Below 40 pixels on the shorter side only one of eighteen objects was identified, and that identifier was incorrect, making this the only size band in which the pipeline is more likely wrong than right when it answers. Multiple simultaneously visible symbols reduced coverage through acquisition geometry rather than through confusion between identities: precision did not degrade on two-symbol frames, and no accepted identifier matched the other symbol present in the same frame. These values characterize the implemented pipeline on the present dataset and should not be read as an unconditional operating specification.

7.6 Sequence-level pipeline

The sequence-level evaluation assesses the identification decision produced by accumulating frame evidence over an episode, as defined in Section 6.6. The unit of analysis is the episode, not the frame, and the reported outcome is the categorical decision returned at the end of the episode. Detection is not re-evaluated here, and the presence of a single

correct frame is not treated as a sequence-level success. Episode boundaries and the assignment of a physical code identifier to each track relied on manual review during dataset construction (Section 6.2.5). The results reported below therefore characterize the identification decision given verified episodes, not the performance of an autonomous end-to-end tracking-and-identification pipeline.

7.6.1 Evaluation scope and overall outcome

The 34 test episodes (Section 7.1) contain 6,855 frame-level observations, of which 2,959 forwarded identification evidence to the accumulation stage. Episode length varied considerably, from 11 to 867 observed frames, with a median of 99, which reflects the differing durations for which a code remained continuously visible.

Figure 35 reports the overall outcome. The decision rule returned an identifier for 31 of the 34 episodes, a resolution rate of 91.2%. Of these, 29 were correct and two were incorrect, giving an identification accuracy of 85.3% over all episodes and a precision of 93.5% conditional on an identifier being returned. The false-identification rate was 5.9%, and three episodes, or 8.8%, were unresolved. No episode produced an ambiguous outcome. With 34 episodes these proportions carry substantial uncertainty, and it is reported here rather than left implicit. The Wilson 95% interval for the identification accuracy of 85.3% runs from 69.9% to 93.6%, that for the resolution rate of 91.2% from 77.0% to 97.0%, and that for the precision of 93.5% from 79.3% to 98.2%. Intervals of this width mean that the episode-level values should be read as evidence that the decision rule works on this dataset, not as precise performance estimates. The frame-level results rest on 1,015 objects and are correspondingly tighter: the interval for the 59.1% identification rate runs from 56.1% to 62.1%, and that for the 98.7% precision from 97.4% to 99.3%. All intervals in this chapter are Wilson intervals for a binomial proportion, which remain valid at the small counts and near-boundary rates that occur here.

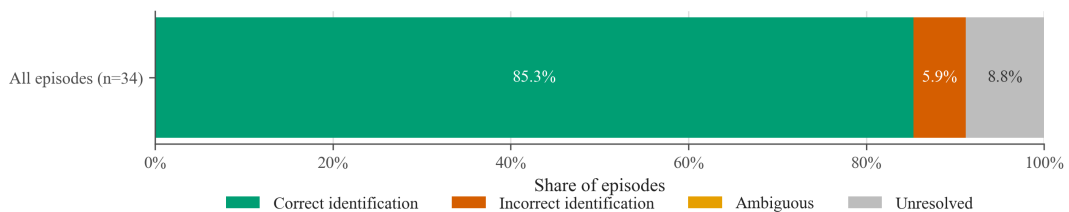


Figure 35: Composition of the sequence-level decision outcomes across the 34 test episodes. Precision conditional on an identifier being returned was 93.5%. Source: own elaboration.

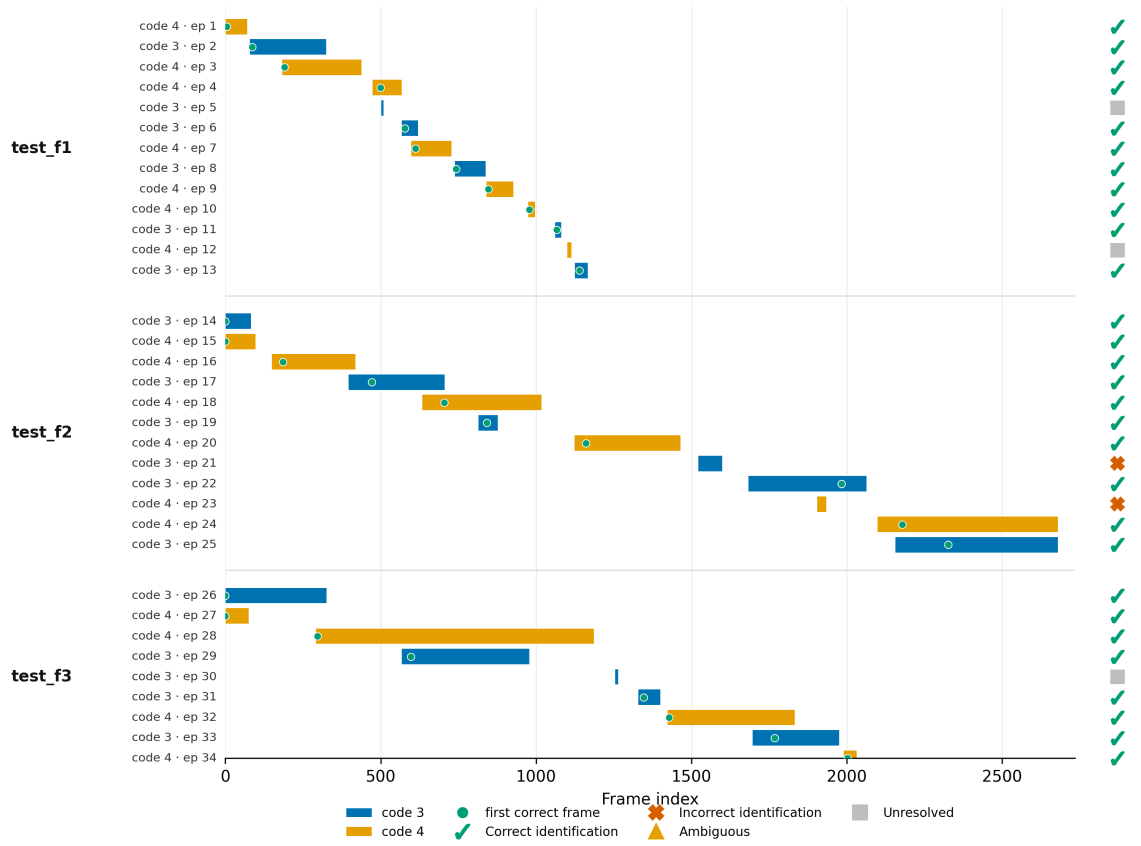


Figure 36: Timeline of the 34 test episodes grouped by lighting setup. Horizontal bars show the frame span of each code-3 or code-4 episode, green circles mark the first correct frame-level identification, and symbols at right show the final episode decision. Most episodes are identified correctly, while two F2 episodes are incorrect and three episodes remain unresolved. Source: own elaboration.

Figure 36 places the aggregate outcome in temporal context. The first correct observation usually appears near the start of an episode, but several F2 episodes require substantially longer accumulation; the two incorrect outcomes also occur in F2, whereas the unresolved outcomes contain no accepted correct frame-level evidence.

The three unresolved episodes received no identification evidence at all: no frame within them produced an accepted primary decode, secondary decode, or dominant similarity candidate. An episode remains unresolved only when the accumulated evidence set is empty, so the unresolved outcomes reflect a complete absence of usable frame-level evidence.

7.6.2 Outcomes under controlled lighting variability

Figure 37 reports the same outcome categories within each lighting setup. Identification accuracy was similar across the three setups, at 84.6% under F1 (11 of 13 episodes), 83.3% under F2 (10 of 12), and 88.9% under F3 (8 of 9). The failure modes differed. Under F1 and F3, every returned identifier was correct and the losses consisted entirely of unresolved

episodes, at 15.4% and 11.1% respectively. Under F2, all episodes were resolved, but two of the twelve returned identifiers were incorrect, reducing precision to 83.3%.

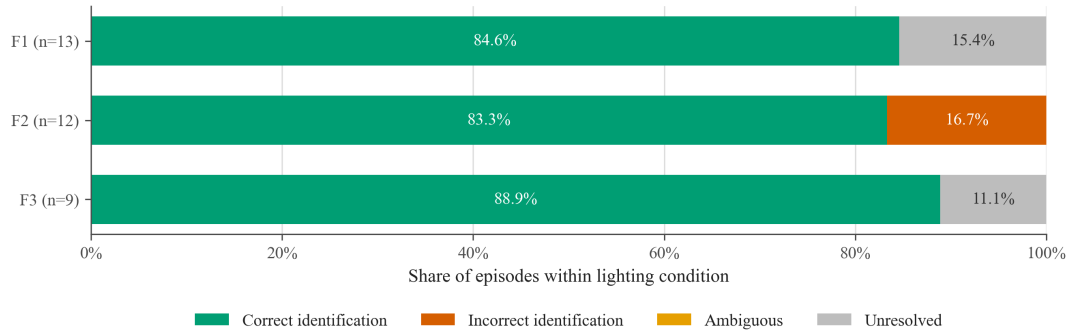


Figure 37: Sequence-level decision outcomes within each lighting setup. Each bar contains the episodes acquired under the corresponding setup and sums to 100%. Source: own elaboration.

The group sizes are small, between nine and thirteen episodes, and the corresponding intervals overlap almost completely: 57.8% to 95.7% for F1, 55.2% to 95.3% for F2, and 56.5% to 98.0% for F3. These values therefore support a descriptive comparison and not a ranking of the acquisition setups. The relevant observation is not the difference in accuracy, which is modest and not separable at this sample size, but the difference in how the residual error is distributed between unresolved and incorrect outcomes. An unresolved episode withholds an identifier and can be escalated, whereas an incorrect identifier enters downstream inventory logic without any indication that it is wrong.

7.6.3 Evidence basis of the sequence decisions

Figure 38 separates the 31 identified episodes according to whether the accepted identifier was supported by at least one validated decode or rested on recovery similarity alone. Twenty-six episodes were decode-supported, and all of them were correct. Five episodes were resolved from similarity evidence alone, of which three were correct and two were incorrect. Both sequence-level errors therefore arose in the similarity-only group.

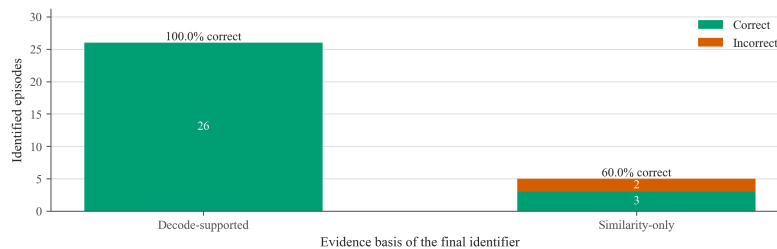


Figure 38: Correctness of the identified episodes by the evidence basis of the accepted identifier. Decode-supported episodes contain at least one validated primary or secondary decode; similarity-only episodes rest on accepted recovery candidates alone. Source: own elaboration.

The composition of the accumulated evidence is given in Table 29. Primary decoding contributed 2,057 of the 2,959 contributing observations and 74.0% of the accumulated support, secondary decoding 258 observations and 9.3%, and recovery similarity 644 observations and 16.7%. The similarity share of accumulated support is smaller than its share of observations because each validated decode contributes unit support, whereas a similarity candidate contributes its raw soft-Hamming score. Validated decoding therefore dominates the accumulated evidence whenever it is available, and similarity evidence becomes decisive only in episodes where decoding never succeeded.

Evidence source	Observations	Support share
Primary decode	2057	74.0
Secondary decode	258	9.3
Recovery similarity	644	16.7

Table 29: Composition of accumulated sequence-level evidence by source.

Table 30 lists the two incorrect decisions. Both occurred under lighting setup F2, both rested on similarity evidence alone, and both were decided by the single-candidate branch of the dominance rule, which accepts a leading identifier without a separation test when only one candidate has accumulated support. The two episodes contributed two and one observations respectively out of 80 and 33 observed frames, and their mean supports were 0.523 and 0.574, close to the lower end of the accepted support range reported in Section 7.4. The errors are therefore associated with a specific and identifiable weakness: an episode in which almost no evidence accumulates can still be resolved, because a single candidate is dominant by construction.

Episode	Lighting	Expected	Accepted	Mean support	Contributing obs.	Dominance branch
test_f2_3_3	F2	datamatrix3	datamatrix2	0.523	2	single candidate
test_f2_4_4	F2	datamatrix4	datamatrix1	0.574	1	single candidate

Table 30: The two incorrect sequence-level identifications and the evidence on which they were decided.

7.6.4 Contribution of temporal integration

The value of accumulating evidence across observations is assessed against a retrospective frame-level indicator. This indicator records whether any frame in the episode produced a correct primary decode. It is a proxy ceiling rather than an operational decision, because identifying such a frame requires ground truth and no runtime mechanism selects it. On the 34 test episodes the indicator was satisfied for 25 episodes, or 73.5%.

Figure 39 compares the two measures episode by episode. The temporal decision was correct for 29 episodes, or 85.3%, against the 73.5% proxy ceiling. Twenty-five episodes

were correct under both measures. Four episodes were decided correctly although no frame in them produced a correct primary decode, and these were resolved through accumulated secondary and similarity evidence. No episode contained a correct decode that the final decision failed to retain. The remaining five episodes were correct under neither measure.

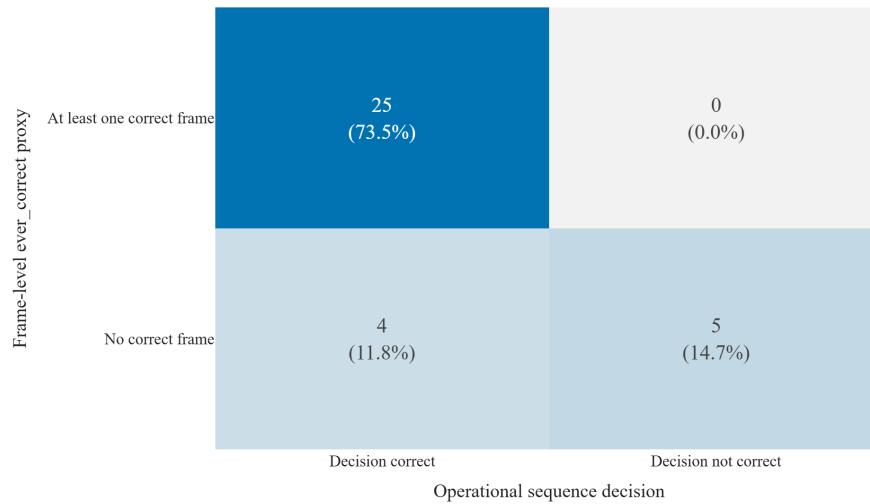


Figure 39: Episode-level agreement between the operational sequence decision and the retrospective frame-level indicator of at least one correct primary decode. Source: own elaboration.

This comparison supports two statements. First, accumulation did not discard evidence that was already available, because no episode with a correct decode was decided incorrectly or left unresolved. Second, accumulation added identifications that single-frame decoding could not produce, because four episodes were resolved correctly without any correct decode. The mechanism is therefore an evidence-combination step rather than a repetition of frame-level decoding, which is consistent with the architectural separation between perception and decision set out in Section 6.1.

Comparing the two evaluation units directly requires care, because they have different denominators. At frame level the pipeline identified 59.1% of 1,015 annotated objects with 98.7% precision; at episode level it identified 91.2% of 34 episodes with 93.5% precision. The gain in coverage is substantial, since an episode is resolved when any of its observations contributes sufficient evidence, and the median episode contains 99 observed frames. The accompanying loss of precision is equally informative: a single incorrect frame-level identifier can determine the outcome of an episode in which almost no other evidence accumulates. Accumulation therefore converts frame-level coverage into episode-level coverage, but it does not by itself suppress isolated false evidence.

7.6.5 Latency to the first correct observation

The preceding subsections treat an episode as a whole. Operationally it also matters how quickly the evidence arrives, since a system does not observe a code indefinitely. Table 31 reports the number of frames from the start of an episode to its first correct frame-level identification. A correct observation occurred in 29 of the 34 episodes, after a median of 14 frames and a mean of 36.7, with a maximum of 300. The distribution is strongly skewed: most episodes yield correct evidence within a few frames, while a small number require prolonged observation.

Episodes	With a correct observation	Mean frames	Median frames	Maximum frames
34	29	36.7	14	300

Table 31: Frames elapsed from the start of an episode to its first correct frame-level identification.

Figure 40 shows that the skew is concentrated in one acquisition setup. Under F1 the first correct observation arrived after a median of 8 frames and never later than 27. Under F3 the median was 10.5 frames, with a single episode at 72. Under F2 the median rose to 55 frames and two episodes required 172 and 300. The lighting setup that produced no unresolved episodes is therefore also the one that took longest to produce usable evidence, which is consistent with F2 supplying weaker per-frame evidence that accumulation compensates for over time.

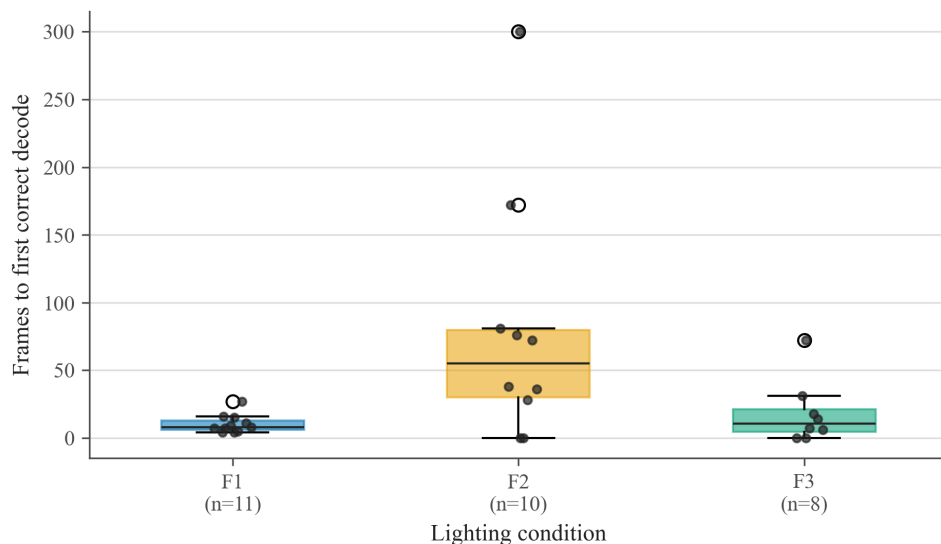


Figure 40: Frames to the first correct frame-level identification, by lighting setup. Boxes span the interquartile range, and each point is one episode. Only the 29 episodes containing a correct observation are shown. Source: own elaboration.

This measure describes when correct evidence first became available, not when the deci-

sion rule would commit to an identifier. The following subsection addresses the second question directly.

7.6.6 How much of an episode the decision requires

To establish when the decision itself becomes reliable, the accumulation and dominance rules were re-applied to every prefix of each episode. The rule is unchanged; it is simply asked at earlier stopping points, and re-deciding on the full observation list reproduces the reported episode outcomes exactly. Figure 41 reports the share of episodes identified, and the share identified correctly, as a function of the number of frames observed.

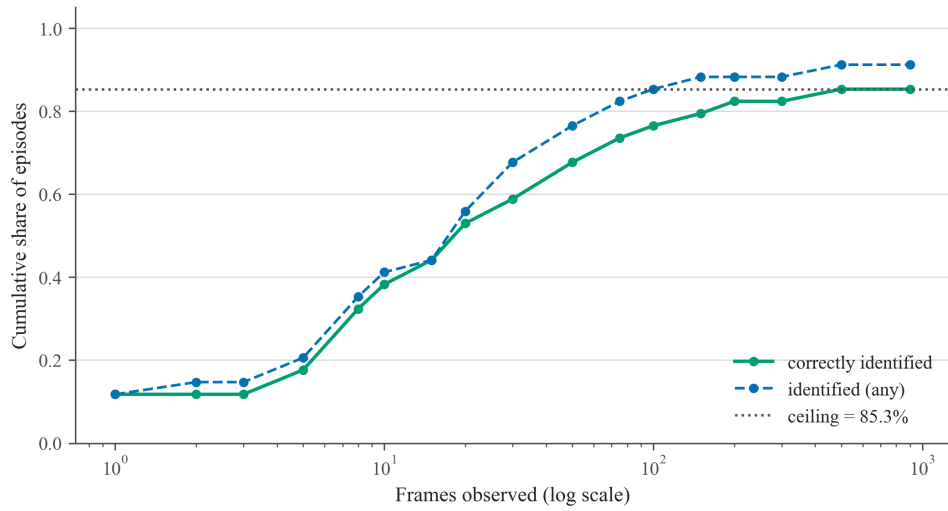


Figure 41: Share of episodes identified and correctly identified as a function of the number of frames observed, obtained by re-applying the episode decision rule to each prefix. The dotted line marks the accuracy reached on the complete episodes. Source: own elaboration.

The curve rises steeply through the first few dozen frames and then flattens. After 20 observed frames the rule had correctly identified 52.9% of the episodes, after 50 frames 67.6%, and after 100 frames 76.5%. It reached 82.4% at 200 frames and its final 85.3% at 500. Roughly nine tenths of the achievable accuracy is therefore obtained within the first 100 frames, while the remainder is contributed by a small number of long episodes in which evidence accumulates slowly. The identified and correctly identified curves stay close together across the whole range, which indicates that stopping early would mainly cost coverage rather than introduce additional errors.

7.6.7 Cost of frame sampling

A related question is how much of the frame stream needs to be processed at all. The decision rule was re-applied to uniformly subsampled observation streams, keeping every

n -th frame. Figure 42 and Table 32 report the resulting accuracy against the share of frames processed.

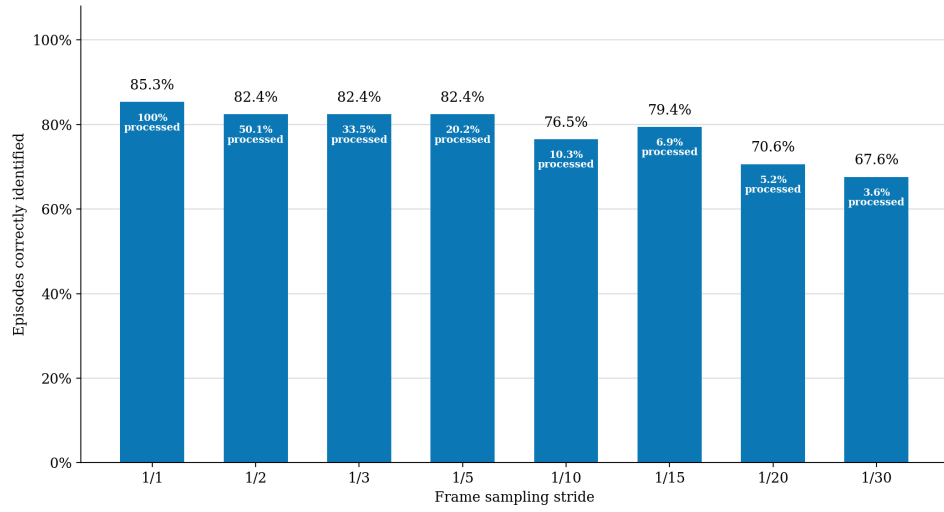


Figure 42: Episode-level accuracy against frame-sampling stride. Each bar keeps every n -th observed frame before the episode decision is re-applied; the second annotation gives the resulting share of the full observation cost. Source: own elaboration.

Stride	Frames processed	Share of full cost (%)	Resolution (%)	Accuracy (%)
1	6855	100.0	91.2	85.3
2	3437	50.1	88.2	82.4
3	2298	33.5	85.3	82.4
5	1385	20.2	82.4	82.4
10	703	10.3	76.5	76.5
15	472	6.9	79.4	79.4
20	359	5.2	70.6	70.6
30	246	3.6	67.6	67.6

Table 32: Episode-level decision performance against frame-sampling stride. Every n -th observed frame is retained before the episode decision is re-applied.

Accuracy degrades gracefully. Processing every fifth frame retains 82.4% accuracy against the full-stream 85.3%, a loss of 2.9 percentage points for one fifth of the observation cost. Only beyond a stride of ten does accuracy fall substantially, reaching 67.6% when 3.6% of the frames are processed. The progression is not strictly monotonic, and the stride of 15 outperforming the stride of 10 is a reminder that with 34 episodes a single case changes the rate by 2.9 points. Figure 43 separates the same analysis by lighting setup and shows that the tolerance is not uniform: under F3 accuracy remained at 88.9% at every stride tested, whereas under F1 it declined from 84.6% to 53.8% as the stride increased to 30. Redundant observation matters most where the per-frame evidence is weakest.

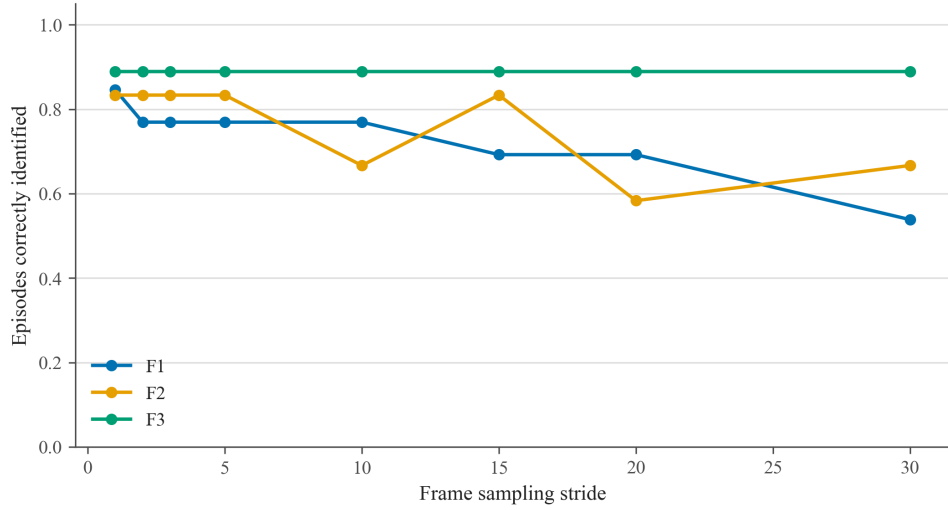


Figure 43: Episode-level accuracy against frame-sampling stride within each lighting setup. Source: own elaboration.

These measurements characterize the evidence requirement of the decision rule, not the throughput of an implementation. They indicate that the temporal decision does not depend on processing every frame, which is relevant to deployment feasibility, but they do not substitute for a timing measurement on specified hardware.

7.6.8 Measured implementation cost and throughput scope

Processing time was measured for 268 evaluated instances on the evaluation computer, which contained an Intel Core i9-13905H CPU, an NVIDIA GeForce RTX 4060 Laptop GPU, and 32 GB RAM. The measurement records the wall-clock time of the serial implementation for each pipeline stage. The timing subset contained approximately one evaluated instance per processed frame. Stages bypassed by control flow were assigned a duration of zero; therefore, the stage-wise mean values in Table 33 represent unconditional contributions to the processing cost per evaluated instance.

Table 33: Stage-wise processing-time profile of the serial implementation across 268 evaluated instances.

Stage	Mean (ms)	P90 (ms)	Maximum (ms)
Quadrilateral estimation	0.4	0.6	2.0
Crop extraction and rectification	2.2	2.9	6.0
Primary decoding	234.3	506.4	516.2
Reference validation	<0.1	<0.1	<0.1
Recovery contribution	674.3	1,588.1	3,931.4
Frame-level decision	<0.1	0.1	0.1
Downstream identification pipeline	911.2	2,053.1	4,448.5
Detection	23.0	—	—
End-to-end mean	934.2	—	—

The recovery branch was invoked for 156 of the 268 timed instances (58.2%). Its conditional mean duration was 1,158.5 ms when invoked. When averaged over all timed instances, including the 112 cases resolved before recovery, it contributed 674.3 ms to the downstream processing cost. The recovery branch therefore accounted for most of the observed end-to-end processing time, whereas detection contributed 23.0 ms on average. The measured downstream mean of 0.911 s per evaluated instance, together with the 0.023 s mean detection time, yields an end-to-end mean of 0.934 s. Under the approximately 29.9 fps source-video rate and the observed instance density, this corresponds to an average serial throughput of approximately 1.07 processed frames per second. The implementation was therefore not evaluated as a real-time frame-by-frame system for the source stream. The 2.053 s downstream 90th-percentile cost and the 4.449 s maximum further show that the average rate alone does not characterize the variability of processing time.

The stride analysis above identifies how much observation can be removed before episode-level accuracy is materially affected, whereas the timing profile identifies the capacity of the current implementation. Processing every fifth frame retained 82.4% episode-level accuracy but would require processing approximately 6.0 frames per second from the source stream, which exceeds the measured serial capacity. A stride of 30 reduces the required rate to approximately 1.0 frame per second, close to the measured mean capacity, but yields 67.6% episode-level accuracy. These findings describe the computational scope of the current serial implementation on the stated hardware; they do not establish the real-time capability of the proposed system architecture or of a parallelized implementation.

7.6.9 Effect of confirmed symbol visibility

The annotation records, for each episode, whether it contains at least one frame in which the symbol was judged fully visible. This is a property of the acquisition rather than of the pipeline, and it separates the episode outcomes sharply. As Figure 44 shows, 21 of the 22 episodes containing a fully visible frame were identified correctly, a rate of 95.5%. Among the 12 episodes without one, 8 were identified correctly, a rate of 66.7%.

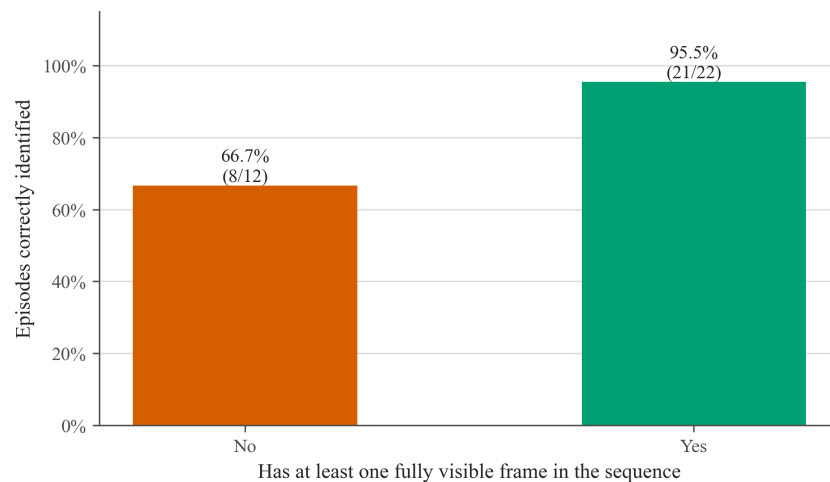


Figure 44: Share of episodes correctly identified, separated by whether the annotation confirms at least one fully visible frame within the episode. Source: own elaboration.

Both parts of this comparison are informative. A single confirmed clear view of the symbol is close to sufficient for a correct episode-level decision, which supports the operational reading that the system needs one good look rather than continuous good conditions. At the same time, two thirds of the episodes that never offered a clear view were still resolved correctly, which is precisely the contribution of accumulating partial evidence. The groups are small, at 22 and 12 episodes, and their intervals run from 78.2% to 99.2% and from 39.1% to 86.2% respectively, so the direction of the difference is clear while its size is not precisely determined.

7.6.10 Summary of sequence-level findings

Accumulating mutually exclusive frame evidence over an episode raised the share of resolved units from 59.1% at frame level to 91.2% at episode level, and produced a correct identification for 85.3% of episodes against a 73.5% single-frame decoding ceiling. Every decision backed by a validated decode was correct. The decision rule reached most of its accuracy within the first hundred observed frames and retained 82.4% accuracy when only every fifth frame was processed, and an episode containing one confirmed clear view of the symbol was identified correctly in 95.5% of cases.

The residual weakness is narrow and specific. Both errors occurred in episodes resolved

from similarity evidence alone, with one and two contributing observations, and both were admitted by the single-candidate branch of the dominance rule, where a lone candidate dominates by construction and no separation is tested. This is a property of the acceptance condition rather than of accumulation, and it suggests that a minimum-evidence requirement alongside the relative-separation test would address it. The evaluation covers 34 episodes from three videos, so these values characterize the implemented decision rule on the present dataset rather than deployment-level performance.

7.7 Ablation of the recovery stages

The two evaluation levels have so far been reported separately. This section combines them into a single ablation that asks what each recovery stage contributes, and whether that contribution is the same once evidence is accumulated over an episode. The pipeline is decomposed in the order it applies its stages: primary decoding alone, then primary decoding with secondary decoding, then the complete branch including similarity acceptance. At frame level each configuration restricts the evidence sources the identification rule may use. At episode level the same restriction is applied to the observation stream and the episode decision rule is re-applied, so both levels are ablated against the same evidence and by the same rule rather than by separate approximations.

The reported configurations are internal architectural baselines rather than comparisons between unrelated recognition systems. Each configuration begins from the same detector-derived quadrilateral and differs only in the downstream evidence sources that are enabled: primary decoding alone, primary decoding followed by bounded secondary decoding, and the complete pipeline including similarity-based recovery. At episode level, the same restrictions are applied to the observation stream before the same decision rule is re-applied. The resulting differences therefore quantify the incremental contribution of the implemented recovery stages while holding the detector output, reference library, evaluation set, and evaluation units fixed. A full-frame decoder or an external commercial reader is not evaluated here, so the ablation does not claim superiority over such systems.

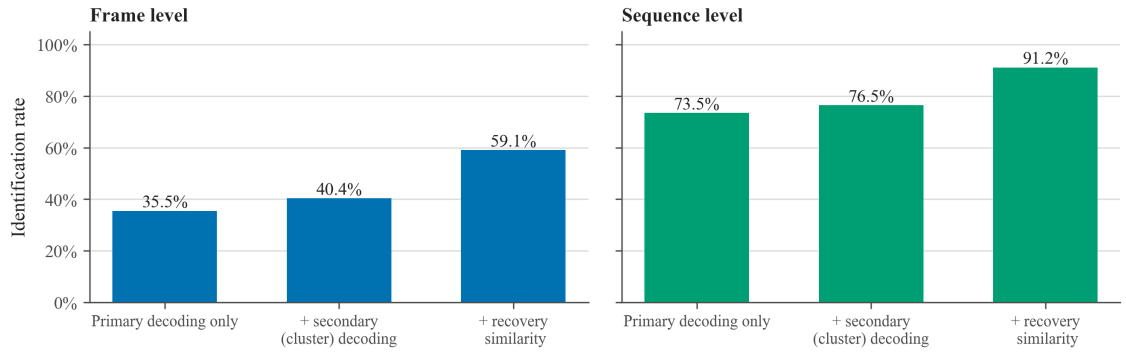


Figure 45: Cumulative internal ablation of the implemented recovery stages at frame and episode level. Each successive configuration retains the evidence sources of the preceding configuration and adds one further stage. Source: own elaboration.

Figure 45 shows that the two levels respond differently to the same additions. At frame level the three configurations identified 360, 410, and 600 of the 1,015 annotated objects, or 35.5%, 40.4%, and 59.1%, so similarity-based recovery supplies the largest single increment, 18.7 percentage points, and secondary decoding a modest 4.9. At episode level the same configurations identified 25, 26, and 31 of the 34 episodes, or 73.5%, 76.5%, and 91.2%. Two observations follow.

First, temporal accumulation is worth more than any single frame-level stage. Primary decoding alone resolves 35.5% of individual objects but 73.5% of episodes, because an episode succeeds if any of its observations decodes. Accumulation converts a modest per-frame rate into a substantially higher operational rate without any additional recovery machinery.

Second, the two mechanisms are complementary rather than redundant. If accumulation simply compensated for weak frames, adding recovery to an already-accumulated decision would yield little. Instead, similarity-based recovery raises episode-level identification by a further 14.7 percentage points, from 76.5% to 91.2%. Recovery reaches objects that repetition alone does not, because it addresses observations that no amount of repeated viewing would have decoded under the same conditions.

The cost of the final step appears in precision rather than in coverage. The two configurations restricted to validated decoding are exact at both evaluation levels, returning no incorrect identifier at all. Adding similarity acceptance lowers precision to 98.7% at frame level and to 93.5% at episode level. The larger drop at episode level reflects how concentrated the episode outcomes are: eight incorrect identifiers among 600 objects carry limited weight, whereas two incorrect decisions among 31 episodes do not. The ablation therefore locates the principal gain and the principal risk of the architecture in the same stage, which is consistent with the stage-level and error analyses reported earlier.

7.7.1 Ablation of the similarity measure

Comparing a reconstructed module matrix with a reference admits several formulations, and the choice matters because it determines which reference is ranked first. Three were considered: normalized Hamming similarity over the binarized matrix, Jaccard similarity over the sets of dark modules, and the soft-Hamming measure of Equation 5. The first two are the natural baselines, the first because it is the standard measure for binary codes and the second because a Data Matrix symbol can equally be read as a set of dark modules rather than as a labelled grid.

The three measures were compared on the objects that actually reach this stage, using the candidate scores recorded during the reported evaluation run. All three are computed from the same reconstructed matrices against the same eight references, so the comparison isolates the measure from every other part of the pipeline. Table 34 reports how often each ranks the correct reference first, and where it places the correct reference on average.

Similarity measure	Representation	Top-1 (detector)	Top-1 (annotated)	Mean true rank
Soft Hamming (implemented)	continuous per-module	61.0%	61.9%	2.26
Normalized Hamming	binarized modules	45.8%	45.8%	3.07
Jaccard	dark-module sets	39.7%	39.5%	3.44

Table 34: Retrieval performance of the three candidate similarity measures over the reconstructed matrices of the reported run. Top-1 is the share of objects for which the correct reference receives the highest score; the mean true rank is its average position among the eight references.

The separation is substantial and consistent across both geometric conditions. Retaining the continuous per-module value rather than binarizing it is worth roughly fifteen percentage points of retrieval accuracy, and the set-overlap formulation is a further six percentage points behind binary Hamming. The reason is visible in what each measure discards. Binarization forces a module whose reconstructed value is near the decision boundary to contribute as though it were certain, so the least reliable evidence is given the same weight as the most reliable. Jaccard additionally discards the agreement between modules that are light in both the reconstruction and the reference, which for a Data Matrix symbol is roughly half of the available evidence. The soft measure keeps both, which is why it separates the correct reference more often and, when it fails, still places it closer to the top.

A fourth variant was considered and rejected on structural grounds rather than empirically. It is tempting to exclude the perimeter modules from the comparison, since the finder pattern and the timing pattern carry no payload. For symbols of a fixed size these modules are prescribed by the Data Matrix specification and are therefore identical across every reference in the library. Excluding a set of modules that contributes the same number

of agreements to every candidate subtracts one constant from the numerator and another from the denominator, which is a monotone transformation of the score. A monotone transformation cannot reorder candidates, so such a mask changes the numerical value of the similarity but never the identifier that is selected. Masking is useful only where what it removes varies across candidates, which the structural pattern does not.

Two boundaries of this comparison should be stated. It evaluates retrieval, that is which reference ranks first, and not acceptance; whether a leading candidate is forwarded as evidence is decided separately by the dominance rule of Section 6.5.1. It is also a comparison among the three measures considered, not a search over all possible measures. Learned or calibrated alternatives, which would require a separate training and calibration set, are treated as future work rather than as candidates for the implemented system.

7.7.2 Ablation of the accumulation rule

The preceding ablation removes recovery stages while keeping the episode decision rule. The complementary question is what the decision rule itself contributes, holding the frame-level pipeline fixed. Two simpler aggregation rules were applied to the same observation stream. The *first-valid* rule commits to the identifier of the first observation in the episode that forwards any evidence, which is the natural baseline for a system that stops as soon as it has an answer. The *majority* rule counts the forwarded identifiers and returns the most frequent one, ignoring the support attached to each. Table 35 reports all three.

Aggregation rule	Episodes	Identified	Resolution (%)	Accuracy (%)	Precision (%)
First valid evidence	34	31	91.2	70.6	77.4
Unweighted majority	34	31	91.2	85.3	93.5
Support accumulation with dominance	34	31	91.2	85.3	93.5

Table 35: Episode-level performance of three aggregation rules applied to the same frame-level evidence stream. The last row is the implemented rule.

All three rules resolve the same 31 episodes, because an episode is answerable whenever any observation forwards evidence, and that condition does not depend on how the evidence is combined. The rules differ only in which identifier they return, so the comparison isolates the aggregation logic from the coverage of the frame-level pipeline.

Accepting the first available answer is markedly worse. It identifies 24 of the 34 episodes correctly, an accuracy of 70.6% at a precision of 77.4%, against 85.3% and 93.5% for the implemented rule. Five episodes are decided incorrectly by the first-valid rule and correctly by accumulation. In each of them the first forwarded observation was an accepted similarity candidate carrying the wrong identifier, which later observations outweighed. Accumulation contributes 14.7 percentage points of accuracy at identical coverage, and that contribution comes entirely from not committing early.

The comparison with unweighted majority voting is more restrained. It matches the implemented rule exactly on these 34 episodes, resolving and misresolving the same cases. The support weighting and the dominance test therefore did not improve on simple counting in this evaluation, and the result should be reported as such. Two considerations qualify it rather than overturn it. The episode sample is small, and the two rules can only differ on episodes where the most frequent identifier is not also the highest-supported one, which occurs rarely when most episodes contain many validated decodes. The dominance test additionally provides something a majority count does not, namely the ambiguous outcome, which allows an episode to be reported as undecided rather than resolved by a plurality; no episode in this dataset exercised that path. Establishing whether the weighting earns its complexity would require episodes with more evenly divided evidence than the present data contain.

8. DISCUSSION

This dissertation has investigated the problem of reliable visual product identification in HMLV warehouse environments through the design of an integrated computer vision pipeline, its implementation, and its evaluation under controlled acquisition variability. The research problem was formulated as a gap between the methodological focus of existing vision-based identification approaches and the operational requirements of dynamic intralogistics systems, where environmental variability, incomplete observations, and real-time decision constraints play a central role. In this sense, the dissertation did not treat Data Matrix recognition merely as a decoding task, but as a broader identification problem situated within a dynamic decision context.

The discussion is structured around the three research questions and their related objectives. The interpretation draws on the frame-level and sequence-level evidence reported in Chapter 7 and states, for each question, both what the evidence supports and where it stops. The dissertation reports results from one controlled dataset comprising 1,015 annotated objects and 34 episodes across three acquisition videos, so the claims below concern the behaviour of the implemented system under the evaluated conditions rather than validated performance in an operating warehouse.

8.1 Discussion of RQ1: system-level design principles

The first research question asked: *What system-level design enables robust visual product identification with integrated prior knowledge?* This question was primarily architectural in nature and sought to identify the conceptual structure required for reliable vision-based identification under uncertainty. In relation to this question, the dissertation addressed Objectives O1–O3, namely: (O1) *to conceptualize a modular system architecture for vision-*

based Data Matrix identification that integrates detection, decoding, and validation; (O2) to incorporate structured prior knowledge through a reference identifier list and validation mechanism; and (O3) to define architectural principles that support robustness under environmental variability and reduce uncontrolled error propagation.

RQ1 has been addressed at a conceptual and system-design level through the development of the architecture presented in Chapter 6. The proposed pipeline formalizes visual product identification as a multi-stage reasoning process in which detection, geometric normalization, decoding, recovery, validation, and temporal integration operate as explicitly separated but interacting modules. This modularity is not only a software design preference, but a methodological response to the research problem itself. In visually variable environments, reliable identification depends not only on whether a symbol can be detected or decoded in isolation, but also on how intermediate uncertainty is propagated, constrained, or corrected within the system.

Objective O1 was addressed through the formal definition of a complete pipeline structure that integrates frame-level perception and sequence-level identification. This is an important distinction, because the dissertation argues that product identification in mobile intralogistics scenarios should not be reduced to single-frame recognition performance. By separating perception from decision, the proposed architecture makes explicit that frame-level outputs constitute evidence rather than final operational conclusions. This system-level interpretation represents one of the central conceptual contributions of the dissertation.

Objective O2 was addressed through the introduction of structured prior knowledge into the identification process. More specifically, the proposed architecture does not accept decoding outputs uncritically, but evaluates them against a predefined identifier set. This validation step constrains the hypothesis space and reduces the risk that noisy or partially corrupted decoding outputs propagate directly into downstream inventory logic. From an operational perspective, this is particularly relevant in HMLV settings, where even low-frequency identification errors may have disproportionate consequences for inventory consistency or item retrieval. The contribution here is therefore not an algorithmic novelty in decoding itself, but the architectural integration of prior knowledge as a reliability-enhancing mechanism.

Objective O3 was addressed through the formulation of explicit design principles that guided the overall system structure. These include prioritizing robustness over idealized performance, emphasizing end-to-end identification reliability over isolated component metrics, treating controlled variability as a design requirement, and distinguishing frame-level perception from sequence-level reasoning. Together, these principles establish a coherent architectural position: robust identification in dynamic environments is better understood as an emergent property of a structured system than as the direct output of a

single recognition model.

The empirical results support the architectural argument at two specific points. Reference validation admitted no incorrect identifier from the dedicated decoder: every primary and secondary decode accepted in the frame-level evaluation matched the expected reference, so the constrained hypothesis space behaved as the reliability mechanism O2 describes. Separating perception from decision proved consequential rather than presentational, since the same frame evidence that resolved 59.1% of annotated objects resolved 91.2% of episodes once accumulated. The residual errors also located themselves where the architecture predicts they would: in the one stage that admits an identifier without a validated decode, namely dominance-based acceptance of a similarity candidate.

At the same time, the answer to RQ1 should be interpreted within the actual scope of the dissertation. The research provides a conceptually grounded and internally coherent system design, evaluated as one instantiation on one controlled dataset. Alternative architectures were not implemented and compared. Accordingly, RQ1 may be regarded as addressed in terms of architectural formulation, conceptual justification, and internal empirical consistency, rather than in terms of definitive empirical proof of superiority over alternative designs.

8.2 Discussion of RQ2: implementation and empirical evaluation under controlled variability

The second research question asked: *How can the proposed system be implemented and empirically validated under controlled environmental variability?* It is linked to Objectives O4–O6: constructing the structured dataset, implementing the detection–decoding–validation pipeline with instance segmentation, dedicated Data Matrix decoding, reference validation, recovery, and episode-level aggregation, and evaluating component-level and end-to-end performance under macro-scene-level data separation.

RQ2 has been addressed through a complete implementation and an executed evaluation on the held-out test split. The pipeline was run end to end over 922 test frames containing 1,015 annotated Data Matrix objects, and the episode-level decision rule was applied to all 34 test episodes. The contribution related to RQ2 therefore comprises both the instantiated system and the empirical characterization of its behaviour under the controlled variability dimensions defined in Section 6.2.

Objective O4 was addressed through the design of a structured dataset tailored to the requirements of the proposed pipeline. As discussed in Section 6.2, the dataset was not conceived as a general-purpose benchmark, but as an experimental instrument supporting controlled analysis of system behaviour under realistic acquisition variability. The explicit structuring of distance, illumination, viewpoint, and negative samples enabled the

dissertation to define variability dimensions that are operationally meaningful for Data Matrix-based identification. In addition, macro-scene-level separation was introduced in order to prevent contextual leakage between training, validation, and test sets. This design decision is methodologically important because it aligns evaluation more closely with the intended deployment logic, where the system encounters previously unseen scene configurations rather than minor frame-level variants of known ones.

Objective O5 was addressed through the technical implementation of the detection–decoding–validation pipeline. The dissertation specified and executed the use of a YOLOv8 segmentation-based detector, mask-driven geometric normalization, primary decoding, recovery branches, and sequence-level evidence aggregation.

The evaluation shows how the stages divide the identification work. Detection localized 98.5% of annotated objects, so entry into the pipeline was rarely the limiting factor. Primary decoding resolved 35.5% of objects, and the recovery branch raised the final frame-level rate to 59.1%, an absolute contribution of 23.6 percentage points. Replacing the detector output with annotated polygons increased the final rate by only 1.3 percentage points, which locates the remaining loss after localization rather than within it. The dissertation does not claim that the implementation is complete in an engineering sense, nor that the modules have been exhaustively optimized; several parameters, including the relative margin and the confidence level of the dominance test, remain hand-set rather than calibrated.

Objective O6 was addressed by evaluating component-level and end-to-end performance under macro-scene-level separation. The two evaluation units answer different questions. At frame level the pipeline identified 59.1% of annotated objects with 98.7% precision among accepted identifiers; at episode level it returned an identifier for 91.2% of episodes and was correct for 85.3%, with 93.5% precision. The comparison against the retrospective single-frame decoding indicator is the more informative result: temporal aggregation was correct for 85.3% of episodes against a 73.5% ceiling, resolved four episodes in which no frame produced a correct decode, and lost none in which such a frame existed. In particular, the distinction between frame-level and sequence-level evaluation, together with the layered analysis of detection, decoding, recovery, and temporal integration, provides a coherent experimental logic for assessing system reliability under controlled variability.

The stratified results identify where reliability degrades. Limited symbol visibility was the strongest factor: objects annotated as occluded or too small were identified in 9.2% of cases against 67.9% for normal objects. Viewpoint and blur produced differences of 30.8 and 25.4 percentage points respectively, and identification improved with apparent symbol area, from 37.6% in the smallest area category to 86.0% in the largest. These are conditional comparisons within one dataset, and the factors co-occur, so they describe where the system is weak rather than isolating an independent cause. In particular, the

dissertation does not derive a universal minimum symbol size from the observed area categories.

The evidence also identifies a specific reliability limitation. Errors were confined to a single mechanism. No validated decode produced an incorrect identifier at either evaluation level, whereas dominance-based similarity acceptance produced eight incorrect frame-level identifiers and both sequence-level errors. Five of the nine similarity identifiers accepted for occluded objects were incorrect. Both episode-level errors were admitted by the single-candidate branch of the dominance rule, in episodes contributing one and two observations, where a lone candidate is dominant by construction and no separation is tested. This is a property of the acceptance rule rather than of accumulation, and it indicates that the decision would benefit from a minimum-evidence condition in addition to the relative-separation test.

An important conceptual outcome of RQ2 is that implementation and evaluation are framed as inseparable from system design. The structured dataset, the modular pipeline, and the layered outcome taxonomy together reflect the view that robustness should be investigated through controlled variability rather than inferred from aggregate end metrics alone. This is methodologically relevant because it shifts the emphasis from reporting isolated recognition scores toward understanding how environmental degradation and architectural choices influence identification reliability.

The discussion of RQ2 also highlights several boundaries of the evidence. The sequence-level results rest on 34 episodes from three acquisition videos, which is a small sample: the Wilson 95% interval for the 85.3% identification accuracy runs from 69.9% to 93.6%, and the lighting groups cannot be separated at all at this size. The comparison against alternative aggregation rules is internal, since no external or commercial system was evaluated. The reported latency measures when correct evidence first became available, and although the sampling-stride analysis bounds how much of the frame stream the decision requires, the evaluation hardware and per-frame processing time are not characterized. Claims about real-time feasibility are therefore not supported by the present evidence.

8.3 Discussion of RQ3: operational implications for inventory-related decision support

The third research question asked: *How can temporally validated visual product-identification outcomes be interpreted as evidence for inventory-related decision support in dynamic warehouse operations?* This question extends the scope of the dissertation beyond perception-level system design and addresses the conceptual interpretation of validated visual observations. It is associated with Objectives O7 and O8, namely: *(O7) to formulate how accepted, ambiguous, and unresolved episode-level identification outcomes may be interpreted as uncertainty-aware evidence for inventory-related decision*

support; and (O8) to assess conceptually the integration requirements and boundaries between the proposed identification outputs and warehouse-management processes.

In the present dissertation, RQ3 has been addressed primarily at a conceptual and interpretive level. The architectural formulation developed in Chapter 6 establishes a system in which identification outcomes are produced at the level of temporally aggregated symbol episodes rather than individual frames. This episode-level formulation enables visual observations to be interpreted as structured evidence of product presence over time, thereby extending the meaning of recognition results beyond symbolic decoding. Since the identifier is defined at the level of an individual physical item, an accepted episode-level identification can be interpreted as evidence that a particular inventory item was observed. It does not, by itself, establish the item's warehouse location or justify an automatic update of a warehouse-management record.

The evaluation gives this operational interpretation a concrete output unit. Each episode yields one of three states: an accepted identifier, an ambiguous outcome, or an unresolved outcome. On the test data these occurred for 31, zero, and three of the 34 episodes. An unresolved episode withholds an observation and can be escalated to a human check. An accepted identifier, in contrast, could provide a probabilistic observation to a downstream inventory-state estimator rather than trigger an unconditional overwrite of an inventory record. The measured precision of 93.5% among accepted identifiers characterizes the reliability of this evidence source on the present test data. It does not provide an episode-specific calibrated probability or establish an operational update policy, but it indicates how an uncertainty-aware integration could distinguish accepted evidence from unresolved or ambiguous observations.

In a scanning session in which each reference identifier is expected to occur at most once, a future integration could extend reference validation from an episode-level rule to a session-level assignment constraint. Once a reference-validated direct decode has established an identifier for one tracked object or episode, that identifier could be marked as allocated and excluded from similarity candidate lists for subsequently independent objects. Such a mechanism could reduce candidate ambiguity and prevent the same unique identifier from being assigned to multiple objects during one session. It would require an explicit session definition and would apply only where the identifier denotes a unique physical item; it would not be appropriate for stock keeping units or reusable codes that legitimately occur on multiple items. The present implementation evaluates each episode independently and does not apply such a global constraint.

From an operational perspective, validated identification events may be interpreted as probabilistic observations of product existence within a spatial context. Although the proposed pipeline does not implement explicit geometric localization, the sequence-level aggregation of observations may capture information related to viewing configuration, apparent symbol scale, and temporal persistence. These characteristics suggest that vision-

based identification may serve as a complementary information source for inventory-state estimation in environments where continuous physical tracking is limited or unavailable. Objectives O7 and O8 are therefore addressed through the conceptual translation of identification outcomes into operationally meaningful information units. Rather than proposing a fully implemented warehouse management integration, the dissertation introduces a system-level interpretation in which validated visual observations may contribute incremental updates to a digital representation of product state. In this interpretation, identification is positioned as an interface between visual sensing and higher-level warehouse decision processes.

At the same time, the research does not provide empirical validation of real-time spatial inference or operational deployment within industrial warehouse settings. No warehouse management system was connected, no inventory state was updated, and no spatial position was estimated from the identification outputs. The spatial interpretation of symbol tracks (episodes) is introduced as a conceptual extension of the identification framework rather than as a completed subsystem. Consequently, the contribution related to RQ3 lies in establishing the theoretical and architectural conditions under which vision-based identification may support inventory intelligence, rather than in demonstrating a finalized localization or warehouse integration solution.

Overall, the findings suggest that reliable visual identification in HMLV-inspired environments should be interpreted as an enabling layer for higher-level operational reasoning. By transforming uncertain frame-level perception into validated, temporally grounded identification events, the proposed system establishes a structured foundation upon which spatial inference and inventory-state updating mechanisms may be developed in future work.

8.4 Positioning the results against prior work

Contextualizing these results against prior work is informative, subject to an important caveat. No publicly available benchmark covers detection, decoding, recovery, and sequence-level identification together under the kind of acquisition variability considered here, and the reviewed studies differ in code type, dataset, and acquisition protocol. Table 36 should therefore be read as contextualization rather than as a shared benchmark, and the reported rates are not directly comparable across rows.

Study	Domain	Reported decoding rate	Dataset conditions	Recovery / multi-frame use
Almeida et al., 2021)	DM, self-localization	51.2% (458/895)	Real indoor scenes, cluttered, varying illumination and planes; no blur factor	Recovery: none. Multi-frame: none explicit (tracking proposed only to reduce latency)
Karrach & Pivarciová (Karrach and Pivarciová, 2020)	DM, classical geometric	95.6% (172/180)	Isolated code crops: synthetic, web images, and industrial laser-marked codes; not full-scene detection	Recovery: none. Multi-frame: none
Karrach & Pivarciová (Karrach and Pivarciová, 2021)	DM, classical geometric	99.1%, 100% at multi-scale	Isolated code crops, same three sources; assumes a largely intact finder pattern	Recovery: cascaded geometric fall-back. Multi-frame: none
Liao et al. (Liao et al., 2022)	DM, laser-marked metal	99.5% (against 8.5% for libdmtx)	Isolated codes on aluminium and mobile phone, controlled lab lighting; single degradation type	Recovery: unconditional CNN reconstruction. Multi-frame: none
Matuszczyk & Weichert (Matuszczyk and Weichert, 2023)	DM, polymer	DPM 86.0% (average, corrected)	Close range (3–10 cm), limited angle (0 to ± 30 degrees), low-contrast codes, controlled lighting	Recovery: unconditional encoding network. Multi-frame: none
Liu et al. (Liu et al., 2024)	DM, industrial defects	90.2% overall, 85–88% under defects	Curated single-defect categories (protrusion, interruption, rotation); not compounded degradation	Recovery: unconditional geometric reconstruction. Multi-frame: none
Rybakova et al. (Rybakova et al., 2025)	DM, synthetic	semi- 90.3%	Semi-synthetic; code covers at least 25% of the frame; region of interest already given	Recovery: cascaded geometric fall-back. Multi-frame: none
Edula et al. (Edula et al., 2023)	QR code	33.5% rising to 50.3% with enhancement	Real photographs with varied glare, damage, perspective, and height	Recovery: unconditional super-resolution. Multi-frame: none
Muallim et al. (Muallim et al., 2025)	QR and 1D barcode	15% to 68% (validation), 0% to 90% (real world)	Heavily augmented synthetic damage; manually damaged real codes	Recovery: failure-gated GAN restoration. Multi-frame: Kalman tracker (object continuity only)
Xu et al. (Xu et al., 2017)	1D barcode, video	78% rising to 89% (ever coded)	Real warehouse video, but only 18 distinct barcodes; 1D barcodes only	Recovery: none. Multi-frame: ever-decoded criterion only
Present study, primary decoding only	DM, HMLV	35.5%	Blur, oblique viewpoint, occlusion, and illumination variation combined within the same test set, by design	Recovery: none. Multi-frame: none
Present study, full frame-level pipeline	DM, HMLV	59.1% (98.7% precision)	Same	Recovery: failure-gated cluster decoding and similarity matching. Multi-frame: none
Present study, episode level	DM, HMLV	91.2% resolved, 85.3% correct (93.5% precision)	Same	Recovery: as above. Multi-frame: support accumulation with dominance test

Table 36: Decoding and recognition rates and the recovery and temporal mechanisms reported in prior Data Matrix and barcode studies, contextualized against the present study. DM abbreviates Data Matrix. Reported rates are not directly comparable across studies because datasets, acquisition protocols, and code types differ.

The comparison supports the central argument of the dissertation in one respect that is independent of the differences between datasets. Every reviewed study that evaluates a plain decoder against the same decoder augmented with a failure-conditioned processing

step reports a substantial improvement: Liao et al. (Liao et al., 2022) from 8.5% to 99.5%, Edula et al. (Edula et al., 2023) from 33.5% to 50.3%, and Muallim et al. (Muallim et al., 2025) from 15% to 68% and from 0% to 90%. The pattern recurs across independently developed methods and application domains, and it supports the premise that primary decoding alone is insufficient under degraded acquisition. The present study adds a directly measured instance of the same effect: the recovery branch raised frame-level identification from 35.5% to 59.1% on an identical detector, decoder, and test set, so this particular comparison is not confounded by differences in dataset or hardware.

More distinctive is what the reviewed literature does not do. None of these studies accumulates evidence across frames in order to resolve object identity. Where multiple observations of the same object are used at all, as in Almeida et al. (Almeida et al., 2021) or Muallim et al. (Muallim et al., 2025), the purpose is latency reduction, geometric refinement, or tracking continuity, not identification confidence. Xu et al. (Xu et al., 2017) come closest by reporting an ever-decoded criterion over warehouse video, but that criterion records whether a correct reading occurred at some point rather than committing to an identifier, which is the retrospective proxy this dissertation reports as a ceiling rather than as a decision. Against that same proxy, measured on the present data at 73.5%, the implemented decision rule reached 85.3%, and the ablation in Section 7.7 shows that replacing accumulation with a first-valid rule costs 14.7 percentage points at identical coverage. The gap identified in the literature is therefore not merely unaddressed; closing it is measurably worthwhile.

The comparison also raises a question that warrants explicit consideration rather than omission. Several reviewed studies report decoding rates between 85% and 99% using classical geometric methods, well above the 35.5% obtained here by primary decoding alone. Those results are obtained on narrower material: isolated code crops rather than full scenes, a single defect type, a limited viewing angle, or a fixed close range. The present test set combines blur, oblique viewpoint, occlusion, and illumination variation within the same evaluation by design, and Section 7.5.7 shows that a large part of it lies below the size at which direct decoding becomes productive. The difference in acquisition difficulty is therefore a plausible contributor to the gap. It does not, however, exclude the possibility that part of the gap reflects room for improvement in the primary decoding and geometric preprocessing stages themselves, independent of what temporal reasoning contributes. The best-case result reported in Section 7.5, where primary decoding alone resolved 98.2% of symbols under favourable conditions, is consistent with the first explanation without settling the question.

Several of the geometric reconstruction techniques identified in this comparison address sub-problems closely related to the one the recovery branch solves. Hough-transform line fitting for the finder pattern (Rybakova et al., 2025), dashed timing-pattern edge localization under a two-level screening strategy (Liu et al., 2024), and CNN-based module

classification combined with graph-theoretic edge correction (Liao et al., 2022) each reconstruct a decodable representation from a geometrically degraded symbol by means other than the approach used here. None has been incorporated into the present implementation and none is claimed as part of this contribution; they are noted as candidate directions for extending the recovery branch, whose acceptance stage rather than whose reconstruction stage is the measured bottleneck.

Taken together, this positioning supports the broader argument of the dissertation: that robustness under dynamic, HMLV-representative conditions is better pursued as a system-level evidence-integration problem than as a search for a marginally more accurate single-frame decoder, while also indicating that the primary decoding and recovery components retain potential for independent improvement.

The empirical comparisons in this dissertation are internal architectural ablations: they isolate the contribution of recovery and temporal evidence aggregation while keeping the detector, decoder, reference library, and evaluation data fixed. They do not constitute a benchmark against commercial barcode readers, transformer-based perception architectures, or end-to-end industrial systems. The literature comparison provides context for these alternatives, but no claim is made that the implemented pipeline is superior to systems that were not reproduced and evaluated under the same protocol. A direct comparison would require common annotations, acquisition conditions, operating points, and end-to-end evaluation units, and is outside the present study.

9. CONCLUSION

This dissertation addressed the problem of reliable visual product identification in HMLV-inspired dynamic intralogistics environments. The research was motivated by the observation that existing vision-based recognition approaches predominantly emphasize component-level performance, while real operational contexts require explicit handling of environmental variability, incomplete observations, and uncertainty in the interpretation of visual evidence.

The work therefore formulated visual product identification not as an isolated decoding task, but as a system-level reasoning problem. A modular architecture was proposed in which detection, geometric normalization, decoding, recovery, validation, and temporal reasoning jointly contribute to identification reliability. This perspective shifts the focus from recognition accuracy alone toward the structured integration of perception outputs into decision-relevant information.

The three research questions were addressed at different levels of resolution. The architectural formulation established a conceptual framework for robust identification under uncertainty. The methodological design provided an evaluation-oriented dataset structure and a structured basis for systematic analysis of system behaviour under controlled variability. Finally, the operational interpretation of episode-level identification outcomes established a conceptual basis through which validated visual observations may contribute to maintaining consistency between physical and digital representations of inventory state.

The implemented system was evaluated on a held-out test split of 1,015 annotated Data Matrix objects and 34 episodes. At frame level, primary decoding alone identified 35.5% of annotated objects; adding secondary decoding raised this value to 40.4%, and adding similarity recovery raised it to 59.1%, with 98.7% precision among accepted identifiers. At episode level, the corresponding successive configurations returned an identifier for 73.5%, 76.5%, and 91.2% of episodes. The complete pipeline correctly identified 85.3% of episodes, with 93.5% precision among accepted episode-level identifiers. These results distinguish the incremental contribution of the recovery stages from the additional value obtained by combining evidence across observations.

The novelty of the research does not lie in proposing a new recognition algorithm or achieving state-of-the-art detection performance. Instead, it lies in the formulation of an integrated identification framework that explicitly combines uncertainty handling, temporal reasoning, and operational interpretation within a unified architectural model.

The contributions of the dissertation can be summarized along three complementary dimensions.

From a theoretical perspective, this dissertation formulates a system-oriented interpretation of visual product identification as a reasoning process operating under uncertainty.

By explicitly distinguishing perception from decision and frame-level recognition from sequence-level identification, the work contributes to a conceptual understanding of how reliability may emerge from structured integration of heterogeneous visual evidence.

From a methodological perspective, this dissertation develops an evaluation-oriented dataset design based on controlled environmental variability and macro-scene-level data separation. This approach supports systematic investigation of performance degradation patterns and provides a structured basis for analysing identification reliability beyond aggregate recognition metrics.

From an applied perspective, this dissertation establishes a conceptual bridge between visual recognition outputs and inventory-related decision processes. By framing validated identification events as potential inputs to inventory-state estimation, the proposed framework outlines how vision-based systems may contribute to maintaining data consistency in dynamic intralogistics-inspired environments.

Several limitations must be acknowledged. First, the evaluation covers one dataset: three acquisition videos, 1,015 annotated objects, and 34 episodes. The episode-level sample is small, and no external or commercial system was compared. The contribution of temporal integration was evaluated both against the 73.5% single-frame decoding ceiling and against first-valid and unweighted-majority aggregation rules applied to the same frame-level evidence. The implemented accumulation rule improved accuracy by 14.7 percentage points over first-valid aggregation at identical coverage, whereas unweighted majority produced the same outcomes on this small episode set. This internal comparison does not establish broader superiority of support weighting. Second, the spatial interpretation of visual identification remains a conceptual extension rather than a fully developed localization mechanism. Third, the dataset and acquisition environment, although designed to reflect realistic variability, do not represent full-scale industrial deployment conditions. Fourth, several decision parameters remain hand-set rather than calibrated, including the relative margin of the dominance test, its confidence level, and the viewpoint-category tolerance. The two observed episode-level errors are attributable to this configuration, since both were accepted from a single candidate supported by one or two observations. Fifth, the reported latency describes when correct evidence first became available rather than when the decision rule would commit, and the evaluation hardware and throughput are not characterized, so no claim is made about real-time feasibility. Consequently, the findings should be interpreted as applicable to controlled dynamic environments rather than as direct evidence of operational performance in real warehouse systems.

Future research may extend the proposed framework in several directions. One important avenue concerns large-scale empirical validation of temporal reasoning and recovery mechanisms. A second avenue follows directly from the observed error mechanism: replacing the hand-set acceptance parameters with calibrated ones, and adding a minimum-

evidence condition so that an episode supported by a single low-support observation is returned as unresolved rather than identified. Establishing frames to stable identification, by re-applying the decision rule to successive episode prefixes, would also allow the responsiveness of the system to be characterized. Another direction involves the development of explicit spatial inference models that integrate visual identification with geometric estimation and sensor fusion. Such extensions may enable probabilistic product localization and trajectory modelling in dynamic warehouse settings.

Further directions concern the perception layer itself. Transformer-based detection and end-to-end perception architectures could be compared against the segmentation model used here under the same controlled variability protocol. Self-supervised identifier embeddings offer an alternative to fixed reference matching where the identifier set is large or changes frequently. Multimodal sensing, for example depth, LiDAR, or ultra-wideband positioning, could supply the spatial component that visual identification alone does not provide.

Further work may also explore system integration pathways with warehouse management processes, enabling automated or semi-automated inventory-state updating based on validated visual observations. In addition, the modular structure of the proposed architecture supports investigation of alternative perception models, decoding engines, and data acquisition strategies as part of iterative system refinement.

A further operational extension concerns session-level evidence management. Where a scanning session is known to contain each unique Data Matrix identifier at most once, reference-validated identifications could be retained as allocated observations and removed from the candidate set of later independent objects. This would turn the reference library from a static lookup set into a session-level consistency constraint, while retaining repeated observations of the same tracked object. Its suitability would depend on the semantics of the identifier and on whether repeated physical instances are possible; its operational value should therefore be evaluated only in a deployment setting with an explicit inventory and session model.

Taken together, the results address the research problem at architectural, methodological, and operational interpretation levels. The dissertation establishes a coherent conceptual and methodological foundation for vision-based product identification in visually dynamic industrial environments, and shows, on a controlled evaluation dataset, that combining evidence across observations can resolve identifier decisions that a single observation alone does not yield. By repositioning identification as an uncertainty-aware system process rather than a purely recognition-centric task, the research contributes to a broader understanding of how computer vision may support reliable information integration in cyber-physical decision-support contexts.

9.1 New scientific contributions

The new scientific results of this dissertation can be stated as follows.

1. This dissertation formulates an integrated identification architecture that separates frame-level perception from sequence-level identification, combining detection, decoding, reference validation, failure-activated recovery, and temporal evidence aggregation within a single modular pipeline.
2. This dissertation develops an evaluation-oriented dataset design and a layered evaluation protocol, spanning detection, decoding, frame-level, and episode-level units under macro-scene-level data separation and controlled environmental variability, that decomposes identification reliability into its component sources rather than reporting a single aggregate metric.
3. This dissertation establishes a conceptual framework in which accepted, ambiguous, and unresolved episode-level identification outcomes are interpreted as uncertainty-aware evidence for inventory-related decision support, positioning validated visual identification as a candidate input to inventory-state estimation without assuming warehouse-management integration or spatial localization.

A. LITERATURE SEARCH PROTOCOL

A.1 Conceptual dimensions and keyword formulation

The search logic was structured around four conceptual dimensions that jointly define the scope of this dissertation. These dimensions were derived from the research objectives and reflect the technological, functional, and contextual aspects of the investigated problem.

1. **Technology dimension.** This dimension captures the methodological foundations of the research, focusing on computer vision and deep learning techniques commonly applied to visual perception tasks in industrial environments. Keywords in this group include “*deep learning*”, “*computer vision*”, and “*object detection*”. For shorthand reference, this dimension is denoted as **CV**.
2. **Task dimension.** The second dimension addresses the specific problem of two-dimensional visual code recognition. As terminology varies across application domains and publication venues, multiple keyword variants were included to cover both generic and specific formulations. This group includes “*QR code*”, “*qrcode*”, “*qr-code*”, “*datamatrix*”, “*data matrix*”, “*2D code*”, and “*barcode*”. For shorthand reference, this dimension is denoted as **QR**.
3. **Functional and domain dimension.** This dimension links visual code detection to its operational context within logistics and warehouse systems. Keywords were selected to capture typical intralogistics processes, localization tasks, and smart warehouse applications, including “*intralogistics*”, “*inventory management*”, “*inventory tracking*”, “*smart warehouse*”, “*product localization*”, “*item localization*”, “*object localization*”, “*visual localization*”, and “*self-localization*”. For shorthand reference, this dimension is denoted as **Domain**.
4. **Specific subdomain dimension.** The fourth dimension reflects the specific industrial setting addressed in this dissertation: high-mix low-volume warehousing and manufacturing systems. Keywords include “*high-mix low-volume*”, “*HMLV*”, “*high-mix*”, and “*high mix low volume*”. For shorthand reference, this dimension is denoted as **HMLV**.

To systematically distinguish between survey-oriented publications and original research contributions, an additional keyword, “*review*”, was explicitly introduced into selected query combinations. Based on this distinction, two types of queries were defined:

- **Review-oriented queries**, created by adding the term “review” to the keyword combinations. These queries aimed to identify survey papers, systematic literature reviews, and state-of-the-art overviews. Starting the analysis with review papers

enabled an efficient high-level understanding of existing research directions and helped assess whether any prior work had already integrated the four dimensions of interest.

- **Research-oriented queries**, which focused on original publications presenting empirical results, methodological contributions, algorithms, datasets, or applied systems.

It is important to note that the review-oriented and research-oriented queries were not mutually exclusive. Review-style papers could still appear in the research-oriented results if they matched the remaining keywords. Such overlaps were expected and resolved during the manual screening phase. This two-stage query design supported the parallel identification of synthesized knowledge and underlying technical contributions.

A.2 Query construction

The four conceptual dimensions were combined using Boolean operators (“AND”, “OR”) to construct compound queries. In principle, the most informative scenario would involve combining **all four dimensions simultaneously**, thereby retrieving publications that jointly address computer vision methods, two-dimensional code recognition, logistics-related localization tasks, and HMLV environments.

However, combining all four dimensions resulted in zero publications, both for review-oriented and research-oriented queries (Queries 1 and 2 in Table 37). This outcome indicates that no existing publication simultaneously addresses all four aspects within a single study. Consequently, the dimensionality of the queries was systematically reduced, and combinations of three and two dimensions were explored.

To maintain methodological consistency across all queries, the following filters were applied:

- **Time range:** publications from *2019 onwards*, reflecting recent advances in deep learning based vision systems and their industrial applications.
- **Language:** restricted to *English*.
- **Document type:** included only *journal articles*, *conference papers*, and *review papers*.

Each query was manually verified within Scopus to check for duplicates and to assess basic relevance.

A.3 Query results and interpretation

The full search logic, including dimension combinations and the corresponding number of retrieved publications, is summarized in Table 37.

Query	CV	QR	HMLV	Domain	Review	Number of publications
1	X	X	X	X	X	0
2	X	X	X	X		0
3	X	X		X	X	0
4	X	X	X		X	0
5	X		X	X	X	0
6	X	X		X		24
7	X	X	X			0
8	X		X	X		1
9	X		X		X	1
10	X	X			X	19
11	X			X	X	56
12	X		X			39
13	X			X		2216
14	X	X				633

Table 37: Overview of search queries and result counts across different dimension combinations. Grey highlighting marks the query groups selected for further screening. CV = computer vision, QR = Data Matrix/QR code, HMLV = high-mix low-volume, Domain = logistics and warehouse.

1. **Full (four-dimension) combinations** The search process began with the most restrictive scenario, combining all four conceptual dimensions. Both review-, and research-oriented queries were applied to examine whether any publication simultaneously addressed computer vision techniques, two-dimensional code recognition, logistics-related localization tasks, and HMLV environments. None of these combinations returned any results. Within the applied search strategy, no publication was identified that simultaneously addressed all four dimensions, indicating that this intersection remains weakly represented in the retrieved literature.
2. **Three-dimension combinations (CV + any two of QR, HMLV, Domain)** After the full combinations yielded no results, the search was relaxed to include only three of the four dimensions at a time. Even with this reduced complexity, review-oriented searches consistently returned zero publications, indicating that no survey or systematic review has yet attempted to integrate these topics. Research-oriented queries produced only very limited results:

- CV + QR + Domain: 24 publications
- CV + HMLV + Domain: 1 publication
- CV + QR + HMLV: 0 publications

The combination of computer vision, Data Matrix/QR-related tasks (detection, recognition, identification, etc.), and logistics-related domain terms returned a modest number of publications (24), while only a single study linked computer vision with HMLV environments and logistics terms. No publication was found that combined computer vision, Data Matrix/QR-related tasks and HMLV-specific contexts. These results suggest that the relevant research fields remain only weakly connected in the retrieved literature.

3. **Two-dimension combinations (CV + one other dimension)** Further reducing the combinations to pairs of dimensions revealed significantly broader research landscapes:

- CV + Domain: 56 review papers and 2216 research articles
- CV + QR: 19 review papers and 633 research articles
- CV + HMLV: 1 review article and 39 research papers

Research on computer vision paired with general intralogistics or localization terms resulted in a very large number of publications, reflecting the diversity of CV applications in warehouse automation, robot navigation, and inventory processes. Similarly, the combination of computer vision with Data Matrix/QR codes produced several hundred studies, covering a wide spectrum of tasks from industrial inspection to mobile scanning and robotic perception. In contrast, the intersection of computer vision and HMLV environments was much smaller, amounting to a few dozen publications. This niche set is highly relevant, as HMLV settings have unique operational constraints that differ fundamentally from high-volume or structured warehouse environments.

The results reveal a clear pattern. Full four-dimensional combinations yielded no publications, while three-dimensional combinations produced only a small number of results and no review papers. In contrast, two-dimensional combinations, particularly those linking computer vision with general intralogistics or visual code detection, returned substantially larger result sets.

Specifically, the combinations of computer vision with logistics-related domain terms and with Data Matrix/QR codes yielded thousands of publications. While these large result sets indicate broader and more established search spaces, their size also makes comprehensive manual screening not practical within the scope of this dissertation. Consequently,

within these broad groups, the analysis focused primarily on review papers, which provide condensed overviews of dominant methods, trends, and open challenges.

A.4 Selection and screening process

Based on the query outcomes, publications from Queries 6, 8, 9, 10, 11, and 12 were selected for further analysis, as these represent the most informative intersections with manageable result sizes. The screening process was conducted in two stages.

1. **Title-abstract-keyword screening** Each paper was manually evaluated according to the following criteria:

- clear problem definition,
- publication type (review-oriented or research-oriented),
- industrial application domain,
- explicit mention of HMLV or similar environments,
- consideration of visual or environmental factors,
- whether the computer vision target involved a 2D code,
- overall relevance to the dissertation.

These criteria partially overlap with the four conceptual search dimensions. However, additional manual filtering was necessary because several queries were intentionally reduced (i.e., one dimension omitted) in order to avoid overly restrictive search conditions. The combination of multiple reduced queries therefore ensured comprehensive coverage of the four conceptual dimensions while maintaining manageable result sets.

Only those publications that satisfied the relevance criteria in this stage were advanced to the next phase (Table 38).

2. **Full-text review.** All publications shortlisted during the title–abstract–keyword screening were subsequently reviewed in full. The full-text assessment focused on methodological depth, technological contribution, experimental validation, and applicability to HMLV warehouse environments. Publications that did not demonstrate sufficient methodological or domain relevance were excluded at this stage.

The final selection resulted in six paper groups: three review-oriented groups (Queries 9, 10, and 11) and three research-oriented groups (Queries 6, 8, and 12).

Table 38: Screening results by query

Query No.	Query	Initial results	Relevant
6	CV + QR + Domain	24	12
8	CV + HMLV + Domain	1	0
9	CV + HMLV + review	1	1
10	CV + QR + review	19	10
11	CV + Domain + review	56	33
12	CV + HMLV	39	6
Total		140	62
Total review papers		76	44
Total research papers		64	18
Reclassified from review to research			10
Final review corpus			34
Final research corpus			28

After the initial screening process, a total of 62 publications were identified as relevant. Among these, 44 were classified as review-oriented and 18 as research-oriented contributions. During the first *title–abstract–keyword screening*, 10 publications originally categorized as review papers were reclassified as research articles due to their methodological and experimental focus.

As a result, the final literature corpus consists of 34 review papers and 28 research papers. These publications jointly form the empirical basis of the Related Work chapter and support the structured identification of research gaps.

B. SUPPLEMENTARY QUALITATIVE AND DISTRIBUTION FIGURES



Figure 46: Representative fully visible observations across lighting setups F1–F3, view-point categories S1–S2, and sharp or blurred acquisition. The examples show that scale, illumination, viewing angle, and motion-induced blur can vary independently within the dataset. Source: own elaboration.



Figure 47: Representative occluded or insufficiently visible observations across lighting setups F1–F3, viewpoint categories S1–S2, and sharp or blurred acquisition. Partial visibility co-occurs with variation in scale, contrast, and perspective. Source: own elaboration.

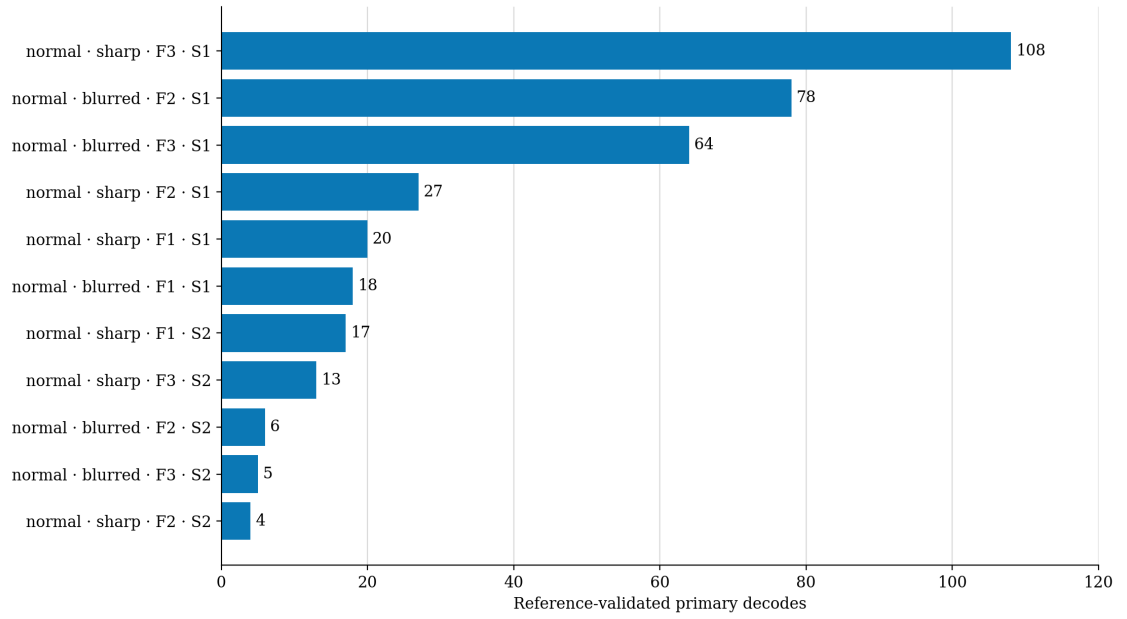


Figure 48: Counts of reference-validated primary decodes by controlled-factor combination, ordered by frequency. Labels identify object condition, blur status, lighting setup, and viewpoint category; most accepted decodes occur for normal S1 observations. Source: own elaboration.

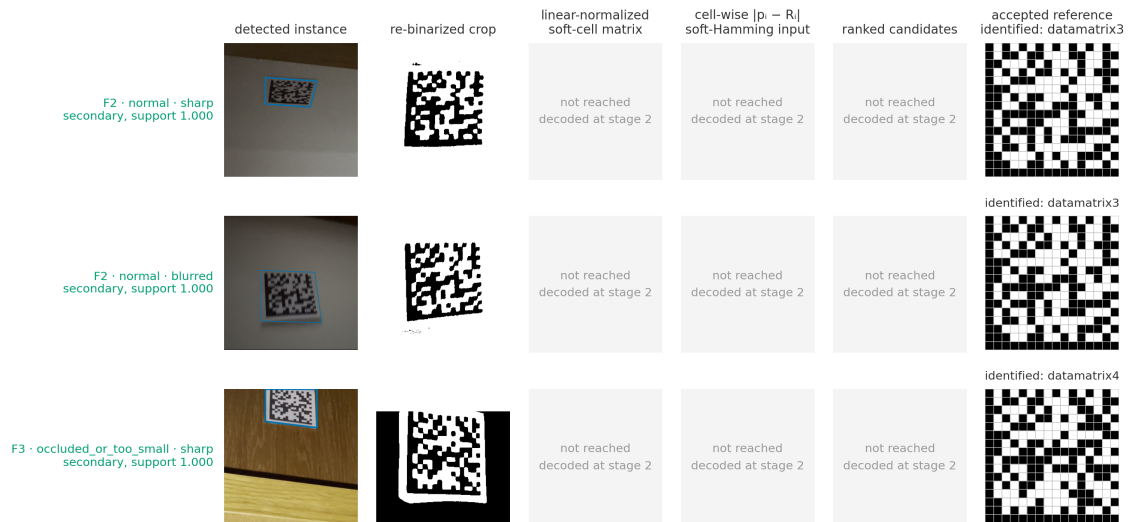


Figure 49: Examples resolved by reference-validated secondary decoding in Stage 2 of the recovery branch. Cluster-based binarization restores sufficient module separation for the dedicated decoder, which returns the expected identifier and terminates recovery before geometric reconstruction. Source: own elaboration.



Figure 50: Examples in which the geometric recovery branch does not produce an admissible symbol quadrilateral. Occlusion, small apparent size, or severe blur prevents a connected finder-pattern geometry from being selected, so no soft-cell matrix reaches reference comparison. Source: own elaboration.

REFERENCES

- Abosinnee, A. S., Bencsik, G., and Abedi, F. (2025). Edges in image with illumination variations scenarios: a review: As abosinnee et al. *The Visual Computer*, 41(14):12277–12305.
- Agostinho, L. R., Ricardo, N. M., Pereira, M. I., Hiolle, A., and Pinto, A. M. (2022). A practical survey on visual odometry for autonomous driving in challenging scenarios and conditions. *IEEE Access*, 10:72182–72205.
- Al-hamadani, S. and Al-Darraji, I. (2025). Warehouse in industry 4.0 based drone, computer vision, and artificial intelligence technologies: A. *Proceedings on Engineering*, 7(1):517–526.
- Al-hamadani, S., Al-Darraji, I., and Jerbi, H. (2025). Improving barcode vision scanning process using a drone-based tracking pid controller for warehouse in industry 4.0. *Al-Khwarizmi Engineering Journal*, 21(2):72–92.
- Alam, E., Sufian, A., Das, A. K., Bhattacharya, A., Ali, M. F., and Rahman, M. H. (2021). Leveraging deep learning for computer vision: A review. In *2021 22nd International Arab Conference on Information Technology (ACIT)*, pages 1–8. IEEE.

- Albayrak Ünal, Ö., Erkeyman, B., and Usanmaz, B. (2023). Applications of artificial intelligence in inventory management: A systematic review of the literature. *Archives of computational methods in engineering*, 30(4):2605–2625.
- Almeida, T., Santos, V., Lourenço, B., and Fonseca, P. (2020). Detection of data matrix encoded landmarks in unstructured environments using deep learning. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 74–80. IEEE.
- Almeida, T., Santos, V., Mozos, O. M., and Lourenco, B. (2021). Comparative analysis of deep neural networks for the detection and decoding of data matrix landmarks in cluttered indoor environments. *Journal of Intelligent & Robotic Systems*, 103(1):13.
- Antoniou, S., Balaska, V., Kansizoglou, I., Symeonidis, S., Karakatsanis, T., and Gasteratos, A. (2025). Humanized tasc: Tag-less and automated stock counting in smart warehouses. In *International Conference on Intelligent Tutoring Systems*, pages 302–312. Springer.
- Asare-Manu, V., Karunathilake, S., and Geng, Z. (2023). Image segmentation with human-in-the-loop in automated de-caking process for powder bed additive manufacturing. page V002T02A029.
- Awan, M., Whangbo, T. K., and Shin, J. (2025). Deep learning methods for autonomous driving scene understanding tasks: A review. *Expert Systems with Applications*, page 128098.
- Azadeh, K., De Koster, R., and Roy, D. (2019). Robotized and automated warehouse systems: Review and recent developments. *Transportation Science*, 53(4):917–945.
- Bauer, J. C. and Daub, R. (2023). Continuous adaptation of deep learning models for optical quality monitoring tasks. In *2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–4. IEEE.
- Bengueddoudj, A., Belhadj, F., and Mostefai, M. (2025). An improved flexible design with efficient detection algorithm for oilu code. In *2025 7th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–6. IEEE.
- Chang, Y.-M., Lin, T.-L., Chi, H.-C., and Lin, W.-K. (2023). Deep learning-based aoi system for detecting component marks. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 243–247. IEEE.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). The crisp-dm user guide. In *4th CRISP-DM SIG Workshop in Brussels in March*, volume 1999. sn.

- Chen, B. and Huang, J. (2024). Drone-assisted qr code recognition method for warehouse management. In *2024 International Conference on Networking, Sensing and Control (ICNSC)*, pages 1–6. IEEE.
- Corpuz, D., Rosario, L., Cempron, J., Lozano, P. L., and Ilao, J. (2025). ‘qr code detection with perspective correction and decoding in real-world conditions using deep learning and enhanced image processing. *Proc. Copyright*, pages 685–690.
- Couturier, A. and Akhloufi, M. A. (2024). A review on deep learning for uav absolute visual localization. *Drones*, 8(11):622.
- Daasan, M. J. A. and Ishak, M. H. I. B. (2023). Enhancing face recognition accuracy through integration of yolo v8 and deep learning: A custom recognition model approach. In *Asia Simulation Conference*, pages 242–253. Springer.
- de Moura Junior, J. L., Frazzon, E. M., and de Lorena Diniz Chaves, G. (2024). Computer vision as a resource in smart warehouses: A systematic review. In *International Joint conference on Industrial Engineering and Operations Management*, pages 73–83. Springer.
- Deniša, M., Ude, A., Simonič, M., Kaarlela, T., Pitkääho, T., Pieskä, S., Arents, J., Judvaitis, J., Ozols, K., Raj, L., et al. (2023). Technology modules providing solutions for agile manufacturing. *Machines*, 11(9):877.
- Dias, N. J. B., Laureano, G. T., and Da Costa, R. M. (2023). Keyframe selection for visual localization and mapping tasks: A systematic literature review. *Robotics*, 12(3):88.
- Diwan, T., Anirudh, G., and Tembhurne, J. V. (2023). Object detection using yolo: challenges, architectural successors, datasets and applications. *multimedia Tools and Applications*, 82(6):9243–9275.
- Du, G., Wang, K., Lian, S., and Zhao, K. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734.
- Edula, V., Ammisetty, K., Kotha, A., Ravipati, D., Davuluri, A., et al. (2023). A novel framework for qr code detection and decoding from obscure images using yolo object detection and real-esrgan image enhancement technique. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.
- Ellithy, K., Salah, M., Fahim, I. S., and Shalaby, R. (2024). Agv and industry 4.0 in warehouses: a comprehensive analysis of existing literature and an innovative frame-

- work for flexible automation. *The International Journal of Advanced Manufacturing Technology*, 134(1):15–38.
- Faqeer, H. A. and Khajavi, S. H. (2025). Digital twin and computer vision combination for manufacturing and operations: A systematic literature review. *Applied Sciences*, 15(18):10157.
- Fonseca, C., Pasaylo, M. N. A., Kawni, N., and Tassew, B. M. (2024). Identifying out-of-stock items at retail stores using computer vision. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 1067–1071.
- Gan, Z. L., Musa, S. N., and Yap, H. J. (2023). A review of the high-mix, low-volume manufacturing industry. *Applied Sciences*, 13(3):1687.
- Gunawan, A. A. S., Andreal, E. J., Budiharto, W., Ngarianto, H., Attamimi, M., and Tolle, H. (2023). Item verification on the smart trolley system using object recognition based on the structural similarity index. *Procedia Computer Science*, 227:147–158.
- Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.
- Heruatmadja, C. H., Prabowo, H., Warnars, H. L. H. S., and Heryadi, Y. (2024). Suitable deep learning based for high accuracy object detection in inventory management: Systematic literature review. In *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, pages 406–412. IEEE.
- Hu, L., Zhong, G., Chen, Z., and Chen, Z. (2025). Datamatrix code recognition method based on coarse positioning of images. *Electronics*, 14(12):2395.
- Ilioudi, A., Dabiri, A., Wolf, B. J., and De Schutter, B. (2022). Deep learning for object detection and segmentation in videos: Toward an integration with domain knowledge. *IEEE Access*, 10:34562–34576.
- Issaoui, H., ElAdel, A., and Zaied, M. (2024). Object detection using convolutional neural networks: A comprehensive review. In *2024 IEEE 27th International Symposium on Real-Time Distributed Computing (ISORC)*, pages 1–6. IEEE.
- Jarahizadeh, S. and Salehi, B. (2026). Tree-net: A novel deep learning tree detection architecture using uav lidar data. *Remote sensing of environment*, 332:115088.
- Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics YOLOv8. <https://github.com/ultralytics/ultralytics>.

- Kalinov, I., Petrovsky, A., Ilin, V., Pristanskiy, E., Kurenkov, M., Ramzhaev, V., Idrisov, I., and Tsetserukou, D. (2020). Warevision: Cnn barcode detection-based uav trajectory optimization for autonomous warehouse stocktaking. *IEEE Robotics and Automation Letters*, 5(4):6647–6653.
- Karrach, L. and Pivarciová, E. (2020). Recognition of data matrix codes in images and their applications in production processes. *Management Systems in Production Engineering*, 28(3):154–161.
- Karrach, L. and Pivarčiová, E. (2021). Comparative study of data matrix codes localization and recognition methods. *Journal of Imaging*, 7(9):163.
- Krishna, P., Kollem, S., and Rao, B. D. (2024). An optimal review of deep learning algorithms for object detection. In *2024 IEEE Region 10 Symposium (TENSYP)*, pages 1–6. IEEE.
- Kumar, D., Singh, R. K., Mishra, R., and Wamba, S. F. (2022). Applications of the internet of things for optimizing warehousing and logistics operations: A systematic literature review and future research directions. *Computers & Industrial Engineering*, 171:108455.
- Kyosuke, T., Koji, T., Shogo, A., Haoming, Y., Tomoe, H., and Kanji, T. (2022). Taylor-made visual burr detection for high-mix low-volume production of non-convex cylindrical metal objects. In *ICPRAM 2022-Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, Volume 1*, pages 395–400.
- Lee, C. K., Lv, Y., Ng, K. K., Ho, W., and Choy, K. L. (2018). Design and application of internet of things-based warehouse management system for smart logistics. *International Journal of Production Research*, 56(8):2753–2768.
- Liao, L., Li, J., and Lu, C. (2022). Data extraction method for industrial data matrix codes based on local adjacent modules structure. *Applied Sciences*, 12(5).
- Liu, Y., Song, Y., Gu, G., Luo, J., Wang, T., and Jiang, Q. (2024). A data matrix code recognition method based on l-shaped dashed edge localization using central prior. *Sensors*, 24(13):4042.
- Losada del Olmo, J. J., de Teruel, P. E. L., Ruiz, A., and García-Clemente, F. J. (2025). Computer vision on the edge: A scalable auto-id solution for industrial logistics. *Procedia Computer Science*, 265:276–284. 20th International Conference on Future Networks and Communications/ 22nd International Conference on Mobile Systems and Pervasive Computing/15th International Conference on Sustainable Energy Information Technology (FNC/MobiSPC/SEIT 2025).

- Manyar, O. M., Cheng, J., Levine, R., Krishnan, V., Barbič, J., and Gupta, S. K. (2023). Physics informed synthetic image generation for deep learning-based detection of wrinkles and folds. *Journal of Computing and Information Science in Engineering*, 23(3):030903.
- Matuszczyk, D. and Weichert, F. (2023). Reading direct-part marking data matrix code in the context of polymer-based additive manufacturing. *Sensors*, 23(3):1619.
- Morar, A., Moldoveanu, A., Mocanu, I., Moldoveanu, F., Radoi, I. E., Asavei, V., Gradi-naru, A., and Butean, A. (2020). A comprehensive survey of indoor localization meth-ods based on computer vision. *Sensors*, 20(9):2641.
- Muallim, T., Kucuk, H., Bareket, M., and Kahraman, M. (2025). Lightweight deep learn-ing model and novel dataset for restoring damaged barcodes and qr codes in logistics applications. *Computer Modeling in Engineering & Sciences*, 143(3):3557.
- Nani, P., Das, S., and Dey, S. (2025). Enhancing object recognition: a comprehensive analysis of cnn based deep learning models considering lighting conditions and per-spectives. *Evolutionary Intelligence*, 18(4):72.
- Nwobodo, O. J., Wereszczyński, K., and Cyran, K. (2023). Slam methods for augmented reality systems for flight simulators. In *International Conference on Computational Science*, pages 653–667. Springer.
- Papadopoulos, S., Patsiouras, E., Ioannidis, K., Vrochidis, S., Kompatsiaris, I., and Pa-tras, I. (2025). Unsupervised object localization driven by self-supervised foundation models: A comprehensive review. *Image and Vision Computing*, page 105807.
- Płaczek, E. and Osieczko-Potoczna, K. (2024). Current state of knowledge and research needs of intralogistics.
- Pohudina, O., Tauseef, M. M., Sivo, A., Cordeschi, N., and Grieco, L. A. (2024). Local-ization of things simulation in smart warehouses applications based on uwb technol-ogy. In *Conference on Integrated Computer Technologies in Mechanical Engineering–Synergetic Engineering*, pages 450–460. Springer.
- Prasad, K. A., Kumar, H., and Tiwari, U. (2025). A systematic review of slam techniques for amrs in warehouse environments. In *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Infor-matics (IC3ECSBHI)*, pages 1270–1275. IEEE.
- Rekavandi, A. M., Xu, L., Boussaid, F., Seghouane, A.-K., Hoefs, S., and Bennamoun, M. (2025). A guide to image-and video-based small object detection using deep learning:

- Case study of maritime surveillance. *IEEE Transactions on Intelligent Transportation Systems*.
- Rodríguez-Martínez, E. A., Flores-Fuentes, W., Achakir, F., Sergiyenko, O., and Murrieta-Rico, F. N. (2025). Vision-based navigation and perception for autonomous robots: Sensors, slam, control strategies, and cross-domain applications—a review. *Eng*, 6(7):153.
- Rybakova, E. O., Limonova, E. E., and Bezmaternykh, P. V. (2025). Improving data matrix mobile recognition via fast hough transform and adaptive grid extractors. *Computer Optics*, 49(6):1182–1190.
- Simeth, A., Kumar, A. A., and Plapper, P. (2025). Flexible and robust detection for assembly automation with yolov5: a case study on hmlv manufacturing line. *Journal of Intelligent Manufacturing*, 36(5):3447–3463.
- Szeszák, B. M., Keréjártó, I. G., Soltész, L., and Galambos, P. (2025). Industrial revolutions and automation: Tracing economic and social transformations of manufacturing. *Societies*, 15(4):88.
- Tadjine, C., Ouafi, A., Benlamoudi, A., and Taleb-Ahmed, A. (2025). Computer vision in warehouse management automation: A survey on implemented methods with prototyping hardware. *Engineering Applications of Artificial Intelligence*, 160:111886.
- Töper, F., Araya-Martinez, J. M., Reig, A. S., Tom, T., Sardari, S., and Ohlhausen, P. (2025). Leveraging synthetic training data for object detection to enhance autonomous depalletizing systems. In *European Robotics Forum*, pages 229–235. Springer.
- Tubis, A. A. and Rohman, J. (2023). Intelligent warehouse in industry 4.0—systematic literature review. *Sensors*, 23(8):4105.
- Vaccari, L., Coruzzolo, A. M., Lolli, F., and Sellitto, M. A. (2024). Indoor positioning systems in logistics: A review. *Logistics*, 8(4):126.
- Varghese, R. and Sambath, M. (2023). A comprehensive review on two-stage object detection algorithms. In *2023 International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security (iQ-CCHES)*, pages 1–7. IEEE.
- Vijayaraj, J., Rithikravivarma, V., Siddharth, S., et al. (2025). Smart barcode scanning using x-ray technology with computer vision and machine learning. In *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pages 1495–1498. IEEE.

- Vivek, N., Suriya, A., Kamala, V., NV, C., Akshya, J., Sundarrajan, M., et al. (2024). Integrating deep q-networks and yolo with autonomous robots for efficient warehouse management. In *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, pages 419–424. IEEE.
- Winkelhaus, S., Grosse, E. H., and Morana, S. (2021). Towards a conceptualisation of order picking 4.0. *Computers & Industrial Engineering*, 159:107511.
- Winkler, H. and Zinsmeister, L. (2019). Trends in digitalization of intralogistics and the critical success factors of its implementation. *Brazilian Journal of Operations & Production Management*, 16(3):537–549.
- Wudhikarn, R., Charoenkwan, P., and Malang, K. (2022). Deep learning in barcode recognition: A systematic literature review. *IEEE Access*, 10:8049–8072.
- Xie, Z., Chen, J. Y. S., Lim, G. W., and Bai, F. (2023). Data-driven robotic tactile grasping for hyper-personalization line pick-and-place. In *Actuators*, volume 12, page 192. MDPI.
- Xu, L., Kamat, V. R., and Menassa, C. C. (2017). Automatic barcode extraction for efficient large-scale inventory management. In *Computing in Civil Engineering 2017*, pages 340–348. ASCE.
- Yadav, A. and Kumar, E. (2023). Object detection datasets and product detection datasets based on deep learning methodology. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–9. IEEE.
- Yoon, B., Kim, H., Youn, G., and Rhee, J. (2023). 3d position estimation of objects for inventory management automation using drones. *Applied Sciences*, 13(19):10830.
- Yu, X., Wang, Y., and Sun, H. (2025). Open-vocabulary object detection survey. *Procedia Computer Science*, 266:62–71.
- Yu, Z. (2023). Exploring deep learning-based visual localization techniques. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, pages 1476–1479. IEEE.
- Zhang, D., Han, J., Cheng, G., and Yang, M.-H. (2021). Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885.
- Zhen, L. and Li, H. (2022). A literature review of smart warehouse operations management. *Frontiers of Engineering Management*, 9(1):31–55.

Zilpe, S., Sani, S., Mitra, D., Ghosh, A., Kumar, P., and Singh, A. (2023). A review of object-based detection using convolutional neural networks. *Artificial Intelligence, Blockchain, Computing and Security Volume 2*, pages 440–447.

LIST OF FIGURES

1	Conceptual distinction between object detection, which localizes an object with a bounding box, and instance segmentation, which estimates an object-specific pixel mask. Source: own elaboration.	43
2	Schematic example of a 16×16 Data Matrix symbol. The solid L-shaped finder pattern fixes orientation, while the alternating borders support estimation of the module grid. Source: own elaboration.	44
3	Structural elements of a Data Matrix symbol: the solid finder pattern, alternating timing pattern, data region, individual modules, and surrounding quiet zone. Adapted from Liao et al. (2022).	45
4	Proposed two-level Data Matrix identification architecture. At frame level, segmentation and geometric normalization feed primary decoding and, if needed, recovery; each observation yields at most one accepted identifier-support pair. At episode level, evidence is accumulated by candidate and classified as identified, ambiguous, or unresolved. Source: own elaboration.	53
5	Segmentation outputs for fully visible Data Matrix symbols across lighting setups F1–F3, sharpness, and viewpoint. Blue contours show the predicted instance boundaries; localization remains stable under moderate blur and oblique viewing, although boundary uncertainty increases. Source: own elaboration.	65
6	Segmentation outputs for occluded or insufficiently visible Data Matrix symbols across lighting setups F1–F3, sharpness, and viewpoint. The predicted contours follow the visible symbol extent, while severe occlusion and small apparent size reduce boundary completeness. Source: own elaboration.	66
7	Successful primary decoding under F1, normal object condition, S1 viewpoint, and sharp acquisition. Instance segmentation is converted into an ordered quadrilateral, rectified and cropped with a quiet zone, after which the dedicated decoder returns a reference-validated identifier. Source: own elaboration.	68

8	Primary-decoding failure under F1, normal object condition, S1 view-point, and blurred acquisition. Detection and geometric normalization produce a usable crop, but blur merges module boundaries and the dedicated decoder returns no symbol, so processing continues to the recovery branch. Source: own elaboration.	69
9	Recovery branch of the evaluated frame-level pipeline. A successful reference-validated secondary decode terminates the branch after Step 2; Steps 3–7 are evaluated only when decoding remains unresolved. Source: own elaboration.	70
10	Cluster-based binarization for an observation resolved by secondary decoding. The crop is converted to grayscale, split using two intensity centroids, and binarized before the decoder returns the reference-validated identifier <code>datamatrix4</code> . Source: own elaboration.	70
11	Cluster-based binarization for an observation that remains unresolved after primary and secondary decoding. Although the intensity separation produces a binary image, blur merges module boundaries and the secondary decoder returns no symbol. Source: own elaboration.	71
12	Successful geometric reconstruction after decoding failure. Edge extraction and line-segment filtering produce one admissible L-pair; RANSAC fits the solid finder-pattern arms, estimates the opposite boundaries, and yields a selected quadrilateral for subsequent rectification. Source: own elaboration.	72
13	Unsuccessful geometric reconstruction under F3, occlusion, S2 view-point, and blur. Edge extraction yields many segments, but none satisfy the L-pair constraints; consequently, no quadrilateral is selected and no soft-cell evidence is produced. Source: own elaboration.	73
14	Soft-cell reconstruction from the selected symbol quadrilateral. The clustered binary crop is oriented and rectified, partitioned into a nominal 16×16 grid, summarized by mean intensity per cell, and linearly normalized to form the soft-cell matrix used for similarity comparison. Source: own elaboration.	74
15	Soft-Hamming reference comparison for a reconstructed symbol. The linearly normalized soft-cell matrix is compared cellwise with each reference; the smallest mean absolute difference gives a similarity score of 0.7873 for <code>datamatrix4</code> , which ranks first and is forwarded to the dominance rule. Source: own elaboration.	74

16	Episode-level evidence accumulation across associated observations. Each frame contributes at most one accepted identifier–support pair; supports are summed by candidate and normalized before the dominance rule returns identified, ambiguous, or unresolved. In the illustrated episode, conflicting frame-level candidates are resolved in favour of ID_12345. Source: own elaboration.	80
17	Detection average precision as a function of the IoU matching threshold. Source: own elaboration.	92
18	Distribution of the mask IoU achieved between each matched prediction and its annotated polygon. The dashed line marks the median and the dotted line the acceptance threshold used for matching. Source: own elaboration.	92
19	Primary decode rate within each controlled factor, over the 1,015 annotated objects. The dotted line marks the overall rate of 35.6%. Every value returned and accepted by reference validation was correct, so decode accuracy was 100% in every group shown. Source: own elaboration.	93
20	Rank of the ground-truth identifier among the reference candidates for the 574 successful soft-cell reconstructions. Source: own elaboration.	96
21	Correctness of the leading similarity candidate before and after the dominance test. Source: own elaboration.	96
22	Support of the leading similarity candidate among reconstructed objects, separated into candidates accepted and rejected by the dominance test. Dotted lines mark the group medians. Source: own elaboration.	97
23	Terminal recovery outcomes within each controlled factor. Each bar covers the objects of one subgroup that entered the recovery branch and sums to 100%. Source: own elaboration.	98
24	Examples of correct similarity-based identification after successful geometric reconstruction. Each row follows the observation from the detector-derived crop through reconstruction and reference comparison; the expected identifier ranks first and passes the dominance rule. Source: own elaboration.	100
25	The eight incorrect frame-level identifications accepted by similarity search. Geometric reconstruction succeeds, but the expected reference does not rank first and an incorrect leader passes the dominance rule; these cases are concentrated in degraded, low-support observations. Source: own elaboration.	100

26	Examples that remain unresolved after successful geometric reconstruction. The candidate similarities are retained for analysis, but the leading reference does not satisfy the configured dominance condition, so no identifier is forwarded. Source: own elaboration.	101
27	Composition of frame-level outcomes over the 1,015 annotated Data Matrix objects. Source: own elaboration.	103
28	Flow of the 1,015 annotated objects through detection, primary decoding, secondary decoding, reconstruction, and similarity acceptance. Recovery raises identified coverage from 35.5% to 59.1%; eight accepted similarity candidates are incorrect, while 40.9% of annotated objects remain unresolved. Source: own elaboration.	104
29	Composition of frame-level outcomes for the most favourable condition combination: lighting setup F3, normal object condition, viewpoint category S1, and no blur ($n = 110$). Source: own elaboration.	105
30	Composition of frame-level outcomes by lighting setup. Each bar represents the annotated objects acquired under the corresponding setup and sums to 100%. Source: own elaboration.	108
31	Composition of frame-level outcomes for sharp and blurred observations. Blur status was assigned through manual review of the annotation suggestion. Source: own elaboration.	108
32	Composition of frame-level outcomes by object condition. The occluded-or-too-small category combines observations with insufficient visible symbol information. Source: own elaboration.	109
33	Composition of frame-level outcomes by viewpoint category. S1 and S2 denote the manually reviewed viewpoint groups derived from annotated quadrilateral geometry. Source: own elaboration.	110
34	Frame-level identification outcome by relative symbol area, overall and within each lighting setup. Each bar presents the outcome percentages within one area bin; empty bins contain no annotated objects. Source: own elaboration.	111
35	Composition of the sequence-level decision outcomes across the 34 test episodes. Precision conditional on an identifier being returned was 93.5%. Source: own elaboration.	115
36	Timeline of the 34 test episodes grouped by lighting setup. Horizontal bars show the frame span of each code-3 or code-4 episode, green circles mark the first correct frame-level identification, and symbols at right show the final episode decision. Most episodes are identified correctly, while two F2 episodes are incorrect and three episodes remain unresolved. Source: own elaboration.	116

37	Sequence-level decision outcomes within each lighting setup. Each bar contains the episodes acquired under the corresponding setup and sums to 100%. Source: own elaboration.	117
38	Correctness of the identified episodes by the evidence basis of the accepted identifier. Decode-supported episodes contain at least one validated primary or secondary decode; similarity-only episodes rest on accepted recovery candidates alone. Source: own elaboration.	117
39	Episode-level agreement between the operational sequence decision and the retrospective frame-level indicator of at least one correct primary decode. Source: own elaboration.	119
40	Frames to the first correct frame-level identification, by lighting setup. Boxes span the interquartile range, and each point is one episode. Only the 29 episodes containing a correct observation are shown. Source: own elaboration.	120
41	Share of episodes identified and correctly identified as a function of the number of frames observed, obtained by re-applying the episode decision rule to each prefix. The dotted line marks the accuracy reached on the complete episodes. Source: own elaboration.	121
42	Episode-level accuracy against frame-sampling stride. Each bar keeps every n -th observed frame before the episode decision is re-applied; the second annotation gives the resulting share of the full observation cost. Source: own elaboration.	122
43	Episode-level accuracy against frame-sampling stride within each lighting setup. Source: own elaboration.	123
44	Share of episodes correctly identified, separated by whether the annotation confirms at least one fully visible frame within the episode. Source: own elaboration.	125
45	Cumulative internal ablation of the implemented recovery stages at frame and episode level. Each successive configuration retains the evidence sources of the preceding configuration and adds one further stage. Source: own elaboration.	127
46	Representative fully visible observations across lighting setups F1–F3, viewpoint categories S1–S2, and sharp or blurred acquisition. The examples show that scale, illumination, viewing angle, and motion-induced blur can vary independently within the dataset. Source: own elaboration. .	150
47	Representative occluded or insufficiently visible observations across lighting setups F1–F3, viewpoint categories S1–S2, and sharp or blurred acquisition. Partial visibility co-occurs with variation in scale, contrast, and perspective. Source: own elaboration.	151

48	Counts of reference-validated primary decodes by controlled-factor combination, ordered by frequency. Labels identify object condition, blur status, lighting setup, and viewpoint category; most accepted decodes occur for normal S1 observations. Source: own elaboration.	152
49	Examples resolved by reference-validated secondary decoding in Stage 2 of the recovery branch. Cluster-based binarization restores sufficient module separation for the dedicated decoder, which returns the expected identifier and terminates recovery before geometric reconstruction. Source: own elaboration.	152
50	Examples in which the geometric recovery branch does not produce an admissible symbol quadrilateral. Occlusion, small apparent size, or severe blur prevents a connected finder-pattern geometry from being selected, so no soft-cell matrix reaches reference comparison. Source: own elaboration.	153

LIST OF TABLES

1	Thematic grouping of the reviewed survey papers according to their primary research focus	16
2	Search Results and Selection Process for Research Papers	22
3	Deep Learning Model Families and Representative Papers in the Reviewed Corpus	24
4	Mapping from limitations identified in the reviewed literature to the requirements they imply and the framework components that address them.	30
5	Mapping between research questions and operational objectives	37
6	Mapping of research questions, objectives, and dissertation chapters	38
7	Alignment of the CRISP-DM methodology with the dissertation structure	39
8	Positioning of the principal pipeline components within the dissertation contribution.	54
9	Structured environmental variability dimensions used in dataset design . .	57
10	Train-validation-test split sizes (frame count)	61
11	Macro-scene manifest and illumination-specific videos. Each split contains one distinct macro-scene and a disjoint identifier set.	61
12	Composition of the frame-level evaluation set. Counts are annotated ground-truth objects; the detection column is the share of those objects matched by the detector at $\text{IoU} \geq 0.50$	83

13	Composition of the sequence-level evaluation set. An episode is one product instance followed through consecutive frames. The observation rate is the share of an episode's frames on which the detector produced a quadrilateral, and the penultimate column counts episodes containing at least one fully visible annotated frame.	83
14	Composition of the frame-level record. True positives and false negatives together form the annotated-object denominator used by every identification rate; false positives and empty-frame rows carry no reference identifier and are excluded from those rates.	84
15	Object-detection performance overall and under each controlled acquisition factor. Precision, recall and F1 are evaluated at $\text{IoU} \geq 0.50$; mAP50-95 averages 101-point interpolated average precision over IoU 0.50 to 0.95 in steps of 0.05.	91
16	Stage-wise performance of the operational recovery branch. The denominator is the set of detected objects without an accepted primary identifier. The reconstruction rate is conditional on secondary decoding having failed, while the contribution column expresses each output as a share of the recovery input. Precision is reported only for stages that output an identifier.	95
17	Terminal outcomes of the operational recovery branch, as a share of the 640 objects entering it.	95
18	Terminal recovery outcomes under controlled acquisition variability, as percentages within each subgroup of objects entering the operational recovery branch.	97
19	Marginal distribution of the incorrect dominance-accepted similarity identifications. Rates are conditional on the accepted similarity identifications within each factor level.	99
20	The eight incorrect frame-level identifications. Support is that of the accepted candidate, margin is its separation from the runner-up, and true rank is the position of the correct reference in the similarity ranking. . . .	102
21	Effect of an absolute support floor applied in addition to the dominance test. Accepted, correct, and incorrect counts refer to similarity-based identifications; the two rightmost columns are computed over all 1,015 annotated objects.	102
22	Overall frame-level identification performance. Detection denotes entry into the downstream identification pipeline, and the recovery contribution combines secondary decoding with accepted similarity-based identification. Precision is conditional on an accepted identifier.	104

23	Decoding performance under controlled acquisition conditions. P denotes the primary decoding rate and the bold value in parentheses (R) denotes the final resolved rate after application of the recovery branch. All rates are reported as percentages.	106
24	Frame-level identification by object condition and blur. Rates are percentages of annotated objects within each subgroup, except precision, which is conditional on an accepted frame-level identifier.	107
25	Composition of frame-level outcomes by relative Data Matrix area. . . .	110
26	Frame-level outcome composition by the shorter side of the annotated symbol in pixels. All values are percentages within the row. The right-hand block restricts the same analysis to objects annotated as normal, separating apparent size from occlusion.	111
27	Frame-level identification for objects observed alone and for objects observed alongside a second Data Matrix symbol.	113
28	Comparison of downstream identification results by polygon source. . . .	113
29	Composition of accumulated sequence-level evidence by source.	118
30	The two incorrect sequence-level identifications and the evidence on which they were decided.	118
31	Frames elapsed from the start of an episode to its first correct frame-level identification.	120
32	Episode-level decision performance against frame-sampling stride. Every n -th observed frame is retained before the episode decision is re-applied. .	122
33	Stage-wise processing-time profile of the serial implementation across 268 evaluated instances.	124
34	Retrieval performance of the three candidate similarity measures over the reconstructed matrices of the reported run. Top-1 is the share of objects for which the correct reference receives the highest score; the mean true rank is its average position among the eight references.	128
35	Episode-level performance of three aggregation rules applied to the same frame-level evidence stream. The last row is the implemented rule. . . .	129
36	Decoding and recognition rates and the recovery and temporal mechanisms reported in prior Data Matrix and barcode studies, contextualized against the present study. DM abbreviates Data Matrix. Reported rates are not directly comparable across studies because datasets, acquisition protocols, and code types differ.	137
37	Overview of search queries and result counts across different dimension combinations. Grey highlighting marks the query groups selected for further screening. CV = computer vision, QR = Data Matrix/QR code, HMLV = high-mix low-volume, Domain = logistics and warehouse. . . .	146

38	Screening results by query	149
----	--------------------------------------	-----