Available online at www.sciencedirect.com

Manufacturing Letters

Manufacturing Letters 00 (2026) 000–000

54th SME North American Manufacturing Research Conference (NAMRC 54 2026)

Enabling Safety 4.0 Through Computer Vision and Collaborative Digital Twins in Smart Manufacturing

Ibrahim Yousif[a], Ahmed Mahmoud[b], Michael Wise[a], Zhe Shan[c], Arthur Carvalho[c], Reza Abrisham Baf[a], Lora Cavuoto[d], Fadel M. Megahed [c,*], Mohammad Mayyas[a]

[a]*Department of Engineering Technology, Miami University, 1601 University Blvd., Hamilton, OH 45011, USA*
[b]*Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29201, USA*
[c]*Farmer School of Business, Miami University, 800 E. High Street, Oxford, OH 45056, USA*
[d]*Department of Industrial and Systems Engineering, University at Buffalo, 407 Bell Hall, Buffalo, NY, 14260*

* Fadel Megahed. Tel.: +1-513-529-4185.; *E-mail address:* fmegahed@miamioh.edu

Abstract

Smart manufacturing has brought adaptive, data-driven systems to the shop floor and beyond, yet equivalent advances in safety intelligence have not kept pace. Control systems, analytics and connectivity now operate with increasing autonomy, but safety frameworks remain rule-based, threshold-driven, and reactive. Safety 4.0 initiatives envision a digital transformation that deploys vision systems, motion tracking and wearable devices, but widespread adoption is hampered by high implementation costs, and the continuous data analysis burden these systems place on safety teams. Moreover, today's safety tools lack the ability to interpret human intent, predict unsafe scenarios or adapt to changing operational contexts. As a result, even highly automated facilities depend on manual intervention to ensure safe human–machine coexistence, leaving a critical gap between autonomous production and autonomous safety. To move from diagnostic monitoring to prescriptive and predictive control, a new safety paradigm is needed. In this work, we address that gap by developing a collaborative digital twin architecture that couples a high-fidelity emulation twin, a cognitive vision-driven twin and a game engine–based human twin to deliver context-aware safety management and immersive training. We outline the system design, demonstrate its ability to synchronise with physical processes, detect and classify safety incidents, and autonomously recommend context-sensitive interventions, and discuss how it bridges the divide between autonomous production and autonomous safety. This promising approach not only contributes to safeguarding personnel but also aligns with the financial and reputational interests of forward-thinking manufacturers.

Keywords: Computer Vision; Digital Twin; Safety 4.0; Safety 5.0, Human-Robot Collaboration

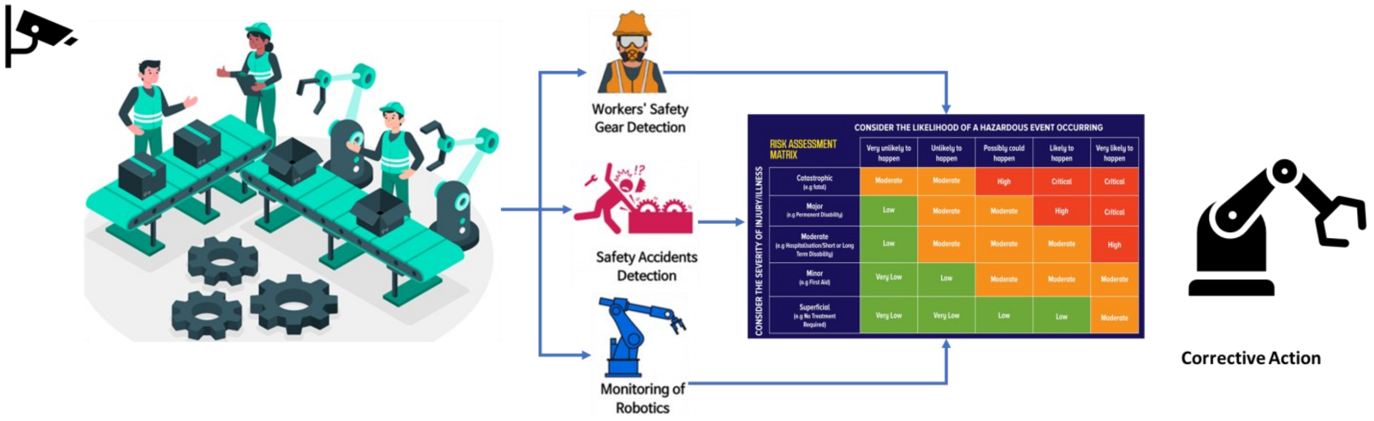
1. Introduction

Manufacturers are accelerating their adoption of Smart Manufacturing (SM) practices to move beyond traditional automation towards adaptive, interconnected systems on the shop floor and across the enterprise. This shift is driven by the evolving market demand of delivering batch-size-one, achieving mass personalization with mass production efficiency [1]. One of the central goals of SM is to replace static, rule-based ("IF-THEN") paradigms, traditionally reliant on subject matter expertise and tacit knowledge, with data-driven approaches that enable real-time decision-making [2]. SM initiatives are enabled by digital transformation technologies such as cyber-physical systems (CPS), the industrial Internet of Things (IIoT), machine learning (ML) and data analytics, and by advances in automated systems like robots, collaborative robots (Cobots) and autonomous mobile robots (AMR) [3], [4]. These technologies have improved product quality and consistency, reduced repetitive, labour-intensive tasks and increased throughput.

However, current deployments operate near their technological limits, lacking the contextual adaptability, resilience, and independent decision-making required to manage unplanned disruptions. These limitations are evident in (low volume-high-mix environments) discrete manufacturing where manual assembly accounts for nearly a third of the workforce due to their order-specific variability and frequent changeovers and process adjustments that increase manufacturing complexity [5]. Beyond technical constraints, the same technologies that drive productivity often neglect the human and safety dimensions of modern production, creating new hazards that must be addressed [6].

Human–robot collaboration (HRC) further highlights this tension between efficiency and safety. In these settings, operators work alongside moving machinery, handle heavy payloads and interact with complex processes, any of which can cause serious injury if not properly managed. Such dynamic conditions impose both cognitive and physical demands on workers, increasing the risk of error, fatigue and harm. For example, facilities equipped with robots have been reported to experience higher injury rates than those without, with one study finding that warehouses using robots had roughly 50 % more injuries [7]. This underscores the urgency of integrating safety considerations into SM strategies. As a result, manufacturing facilities remain inherently high-risk workplaces, where operators often perform tasks involving awkward postures, heavy lifting, and sustained high-intensity activity. As technology evolves rapidly, our reliance on past experience becomes less sufficient, and there is a pressing need to develop parallel advances in safety protocols and training, an evolution described as Safety 4.0, 5.0, and beyond [8], [9].

Safety 4.0 represents a holistic redefinition of safety practices for the SM era, moving beyond computational methods to address the socio-technological implications of digitalization on worker well-being. This transformative approach is characterized by a proactive shift towards the science of process and occupational safety, leveraging innovations like IoT-integrated like wearable technologies including sensor-embedded helmets and wristbands. and smart personal protective equipment (PPE) to enable dynamic risk management and enhance system resilience [10], [11]. While built on core principles of interoperability, transparency, and decentralized decision-making, its practical implementation faces significant financial and resource challenges related to the cost, scalability, and continuous data analysis required by these advanced technologies. Computer vision, powered by artificial intelligence, has become a cornerstone of modern safety protocols, providing situational awareness. Figure 1 illustrates the proposed safety management system. It employs computer vision for safety enhancement in manufacturing, underpinned by a risk assessment matrix that guides decision-making by evaluating the likelihood and impact of safety incidents. The matrix categorizes risks from 'Very Unlikely' to 'Very Likely' and 'Negligible' to 'Catastrophic.' Upon detecting a safety issue such as PPE non-compliance or obstructions on the line, the system uses the matrix to quantify the risk and algorithmically determine the response. This dynamic assessment allows for tailored responses that align with the nature and severity of detected events, ensuring that safety measures are both efficient and proportional.



CONSIDER THE LIKELIHOOD OF A HAZARDOUS EVENT OCCURRING

CONSIDER THE SEVERITY OF INJURY/ILLNESS

RISK ASSESSMENT MATRIX	Very unlikely to happen	Unlikely to happen	Possibly could happen	Likely to happen	Very likely to happen
Catastrophic (e.g fatal)	Moderate	Moderate	High	Critical	Critical
Major (e.g Permanent Disability)	Low	Moderate	Moderate	High	Critical
Moderate (e.g Hospitalisation/Short or Long Term Disability)	Low	Moderate	Moderate	Moderate	High
Minor (e.g First Aid)	Very Low	Low	Moderate	Moderate	Moderate
Superficial (e.g No Treatment Required)	Very Low	Very Low	Low	Low	Moderate

Figure 1: Leveraging Computer Vision in Integrated Safety Management: From On-site Operations to Risk Assessment Metrics in the Age of Industry 4.0 [8]..

CV systems analyze images or video streams in real time to extract meaningful insights and help isolate workers from hazards. Applications range from detecting humans or personal protective equipment (PPE) for access control, monitoring unsafe postures or ergonomic risks, and identifying process anomalies such as debris or spills [12], [13], [14] . Successive deep learning models; versions have improved their capability to detect smaller objects and support tasks such as segmentation, pose estimation and tracking. These tools enable automated monitoring of worker presence and have been validated in industries such as steel manufacturing [15], [16]. For PPE compliance, deep-learning models trained on video from shop-floor cameras can recognize whether workers are wearing helmets, high-visibility vests and other gear in real time. Its applications are diverse, ranging from fundamental objects and hazard detection to sophisticated human-centric analysis, which we accomplished in and building upon. For instance, a smart surveillance system developed by Zhao et al. achieved 94% accuracy in identifying hazardous worker postures, triggering automated alerts to prevent accidents. However, current CV systems predominantly serve diagnostic roles: they recognize hazardous conditions but require human supervisors to interpret the data and decide on corrective actions [17]. Typically operate as isolated modules with limited feedback to the production control architecture [18], [19]. Achieving proactive, autonomous safety requires coupling visual sensing with digital models that reason about risks and orchestrate autonomous responses.

In this context, digital twins (DTs) have been proposed as a future direction because they provide real-time, bidirectional data exchange between physical assets and their high-fidelity virtual counterparts .DT links the physical controller with its digital representation, whether the asset is equipment, material, a process or even a human, so that the virtual model mirrors the current state, supports “what-if” experiments in a risk-free environment, and ingests multimodal data to maintain continuous visibility into performance. By contrast, digital models and digital shadows involve much lower levels of integration. Digital models are static, manually updated representation, with no automated data flow and digital shadows automatically receive only one-way data feeds, but cannot update it [20]. These simpler representations may suffice for basic performance prediction or scenario evaluation, but they lack the responsiveness and decision-making capabilities that a true digital twin offers. Only a digital twin can modify the physical system’s state, enabling predictive verification and optimization of physical processes. Most existing CV-enabled safety systems are digital shadows, visualization tools that stream sensor data to dashboards without closing the loop to control. Transitioning to a full digital twin is therefore essential to permit autonomous safety interventions based on CV observations.

Beyond model maturities, we propose organizing digital twins for safety into four functional domains: emulation/robotics-based, game-engine-based and cognitive digital twins. Emulation/robotics-based digital twins replicate the hardware and software behavior of the physical system. They “act the same” as the real system, providing an exact duplicate for testing control logic, firmware or hardware interactions. Emulation is used for hardware testing, software development and security assessments. In Safety 4.0, an emulation twin can run the actual robot controller in a virtual environment to validate real-time trajectory re-planning, collision detection and fault handling before deployment. It bridges the gap between simulation and reality by ensuring algorithms validated in high-level simulations behave correctly under real control constraints.

Game-based digital twins leverage modern game engines to create photorealistic, physics-aware simulations and immersive environments. Game engines provide realistic graphics and integrated physics, lighting and fluid simulation capabilities, making them highly interconnected. They support incremental development from early prototypes to high-fidelity replicas of the deployed system [21]. In manufacturing, game-engine twins can create interactive training environments in virtual reality (VR) or augmented reality (AR) where workers practice safe responses to hazards, visualize robot movement paths and receive real-time alerts in mixed reality as presented in Figure 2. Game-based human digital twins [22]. Building on these capabilities, game-based simulation engines are increasingly used to implement human digital twins because they render realistic work environments and support bi-directional communication between the human and the machine. In our framework these engines enable operators to interact with the virtual work cell for both routine operations and immersive training: the human can issue commands and receive rich feedback while the machine can communicate status and guidance to the trainee. Game-engine twins thus act as the human interface of Safety 4.0, transforming safety training from passive instruction to active, experiential learning.

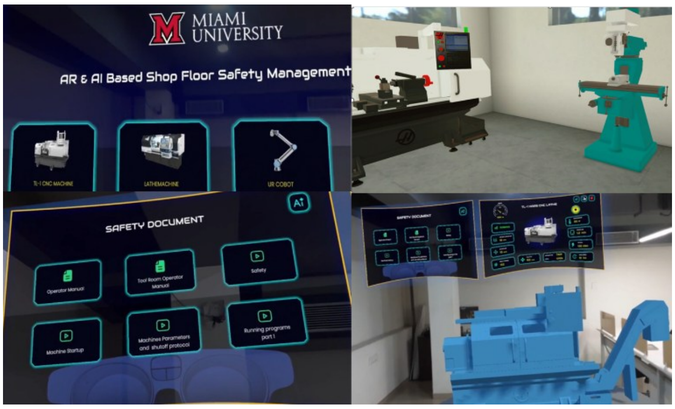


Figure 2: Functional Training Framework.

The final domain, cognitive digital twins, represents the emerging integration of artificial intelligence into digital twinning. A cognitive digital twin augments a digital twin with cognitive capabilities such as attention, perception, and memory functions inspired by cognitive science and machine learning [23]. Attention mechanisms allow the twin to focus on the most relevant data streams; perception transforms raw sensor data into meaningful representations; memory encodes and retrieves knowledge, enabling context-aware reasoning. These capabilities enable the twin to draw implicit knowledge from prior experience and to make higher-level decisions [24], [25]. In the Safety 4.0 context, a cognitive twin integrates CV models with kinematic models and risk analytics. It can track workers’ positions and postures, classify tasks as safe or unsafe and update a real-time risk assessment matrix. The matrix may be a function of human–robot distance, robot velocity, payload, human posture and occlusion confidence. When risk exceeds a threshold, the cognitive twin triggers adaptive safety responses, slowing the robot, altering its trajectory or halting motion, ensuring safe collaboration without interrupting productivity.

Although these domains are conceptually distinct, collaborative digital twins emerge when they interact. A Safety 4.0 framework might comprise a simulation-based twin for offline hazard studies, an emulation twin for real-time path validation, a cognitive twin for perception and risk scoring, and a game-engine twin for immersive human training. Data flows must be bi-directional among these twins and the physical system to ensure that risk information observed by the cognitive twin influences trajectory decisions in the emulation twin and that validated safety strategies from simulations translate into training scenarios in the game engine. For example, a cognitive twin may detect a worker reaching into a danger zone and compute a risk score. It passes this score to the emulation twin, which recalculates the robot's path using kinematic constraints and dynamic safety envelopes. The game-engine twin then visualizes the adapted path and provides haptic or visual feedback to the worker through AR glasses. A simulation twin can later analyze how the risk matrix thresholds or sensor sensitivities affect throughput and safety trade-offs. Such collaboration transforms digital twins from isolated digital artefacts into an integrated Safety 4.0 architecture that supports design, operation and training.

This paper proposes a Safety 4.0 framework that harnesses computer vision, digital twins and immersive technologies to achieve autonomous industrial safety. The framework's contributions are twofold. First, we demonstrate the integration of CV for human monitoring and safety accident detection, showing how real-time perception feeds into the risk matrix and digital twin control loops. Finally, we explore how immersive training using game-engine twins and wearable AR can enhance worker awareness and prepare operators for collaborative scenarios. By connecting these layers, the Safety 4.0 framework reimagines safety as a proactive, cognitive and human-centered digital layer within smart manufacturing. The remainder of this article is organized as follows. Section 2 reviews the evolution of manufacturing safety paradigms, the state of CV-based safety monitoring, and existing digital twin classifications. Section 3 outlines the proposed Safety 4.0 framework and its architecture, detailing how CV, cognitive twins, simulation and emulation models and game-engine interfaces interact to provide adaptive safety. Section 4 reports experimental results from applying the framework to a human–robot assembly scenario, including detection accuracy, Section 5 concludes with future research directions.

2. Literature Review

In construction, AI-powered image analysis is used to automatically detect the absence of PPE like helmets, enhancing compliance and safety on-site. Technology is also being applied in highly specialized domains, such as the use of a YOLOv8-based Digital Twin-Computer Vision (DT-CV) architecture to ensure real-time synchronization and event tracking in robotic assembly [26]. These examples demonstrate a clear trend towards using vision systems as an automated and unbiased layer of oversight that significantly enhances proactive hazard mitigation beyond the capacity of manual inspections [27], [28], [29].

Beyond hazard detection, computer vision is making significant inroads in the field of ergonomics to combat the high prevalence of work-related musculoskeletal disorders (WMSDs) [30], which accounted for over 14% of private-sector injuries in manufacturing in 2018 [31]. Vision-based systems offer a non-intrusive method for real-time ergonomic risk assessment, as demonstrated by Agote-Garrido et al., who used vision-enabled digital twins to model worker motion in 3D and simulate potential fatigue or injury risks. These systems can automatically apply standardized assessment models like REBA (Rapid Entire Body Assessment) and RULA (Rapid Upper Limb Assessment) to continuously monitor worker postures and flag high-risk behaviors [32]. This allows for immediate feedback and dynamic task adjustments, moving beyond the limitations of manual observation or intrusive wearable sensors to reduce long-term injury risks.

Digital Twins (DTs) are further revolutionizing safety management by creating dynamic, virtual replicas of physical manufacturing systems for simulation and predictive analysis. In one notable application, a DT-CV combination was used to replace traditional laser-based safety mechanisms for collaborative robots, enabling human-adaptive speed control and enhancing workplace safety without physical barriers [33]. In the construction sector, DTs are used to support risk assessment through virtual simulations and real-time predictive modeling [34]. This capability allows for the proactive identification of potential system failures and process bottlenecks, enabling preemptive maintenance and risk mitigation strategies that significantly reduce the likelihood of accidents. The integration of DTs with big data analytics and AI, as seen in automotive final assembly lines, supports a proactive Safety 4.0 approach that anticipates hazards rather than reacting to them.

Building on the foundation of Digital Twins, immersive technologies such as Augmented Reality (AR) and the Industrial Metaverse are transforming safety training and on-the-job guidance [35]. AR applications, such as smart glasses that overlay contextual instructions during assembly tasks, have been shown to reduce cognitive load and improve task accuracy [36]. The Industrial Metaverse extends this concept by creating fully immersive and gamified training environments where workers can practice responding to hazardous situations, like equipment failures or fires, in a safe and controlled virtual space. This approach has been found to be more engaging and effective than traditional passive training methods, improving knowledge retention and ensuring that the workforce is better prepared to handle real-world emergencies.

3. Research Methods

This section focuses on the development of cognitive digital twins encompassing the experimental setup, and the vision system specifications including camera selection and algorithms development, data acquisition and processing pipelines, and the computing systems utilized in the development and validation process.

3.1. Experimental setup

The experimental setup, shown in Figure 3, was conducted at the Subtractive Lab within the Advanced Manufacturing Workforce and Innovation Hub at Miami University. The workspace consisted of a UR5e collaborative robot positioned on a working table, where it performed assembly tasks in coordination with a human worker. An Intel RealSense L515 LiDAR camera was mounted in a top-down configuration above the workspace, providing a complete field of view of the collaborative assembly process. This overhead perspective enabled the camera to capture both the robot and human activities simultaneously, which is essential for the real-time zone calculation and safety monitoring system.

The proposed framework enables safe human–robot collaboration by combining real-time instance segmentation with a dynamic zone-allocation mechanism that continuously evaluates and adjusts the shared workspace. The system operates in two modes depending on task progression: an initial static allocation phase in which the human and robot work simultaneously on different regions of the motor base, and a subsequent dynamic reallocation phase that occurs once the human completes their assembly task and places parts for the robot to process. The methodology described below outlines the perception, geometric reasoning, and zone-assignment procedures that support these two operational states. Perception is performed using a You Only Look Once (YOLO) v11 instance-segmentation model trained to identify four classes central to the assembly operation: the human operator, the robot, the motor base, and individual assembly parts. For each detected instance, the model outputs a class label, a confidence estimate, a bounding box, and a pixel-level mask. These segmentation masks serve as the foundation for all subsequent geometric computations, as they provide precise spatial delineation of objects within the shared workspace. Instance segmentation is preferred over bounding-box detection due to its significantly higher spatial fidelity, which is essential when calculating fine-grained operational boundaries on the motor surface.

3.2. Process 1: Static Zone Allocation

During the initial stage of the task, the human and the robot work concurrently on different sides of the motor base. To ensure that neither agent intrudes into the other’s operational region, the motor base is partitioned into three predefined zones: a robot zone, a human zone, and an intermediate buffer zone. The segmentation mask of the motor base is first extracted, and its geometric center is estimated using the oriented bounding box (OBB). The OBB provides a stable reference frame that is robust to partial occlusions or irregular mask shapes. A vertical line passing through the OBB center establishes a consistent left–right division of the motor base regardless of camera angle or mask variability. The side closest to the robot is designated as the robot zone, while the opposite side is assigned to the human zone. A buffer region is then introduced by defining a narrow band, approximately ten percent of the motor width, centered on the dividing line. This intermediate zone serves as a safety margin designed to absorb minor operator movements or camera-induced spatial noise. In this configuration, both agents can perform their tasks simultaneously while maintaining clearly defined and visually interpretable work regions.

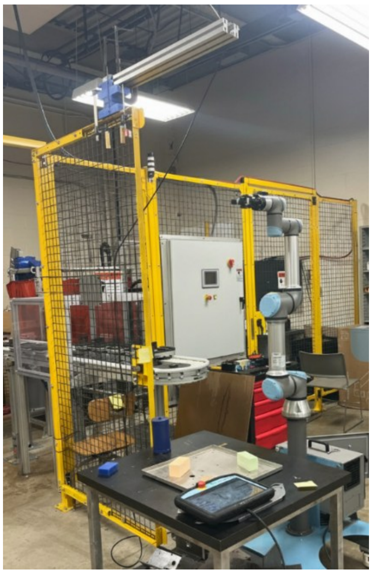


Figure 3: Environment setup and configuration.

3.3. Process 2: Dynamic Zone Reallocation

Once the human operator completes their assigned task and places one or more components on the motor base, the robot transitions from parallel operation to direct part participation. This requires an adaptive reconfiguration of the motor workspace boundaries so that the robot is granted exclusive access to the regions of the motor containing the newly placed parts. The system monitors this transition by detecting assembly parts whose centroids fall within the OBB of the motor base. Using the centroid along with pixel-overlap tests yields a more stable and computationally efficient indicator of successful placement, particularly in the presence of minor segmentation errors.

After validating that one or more part have been placed, the system determines whether the robot is positioned to the left or the right of the motor center. It then computes the “furthest-part boundary” by identifying the extreme X-coordinate of all placed parts: the maximum X-value if the robot is on the left, or the minimum X-value if it is on the right. This boundary represents the limit of the robot's upcoming manipulation region and defines how far the robot's operational zone must extend. The workspace is then reassigned so that the robot zone spans from the robot side of the motor base to the furthest-part boundary, guaranteeing full coverage of the robot's future trajectory. A reduced buffer zone may be included if sufficient space remains, and the human zone is assigned any leftover area. In cases where the placed parts occupy most of the motor surface, the human and buffer zones may shrink to zero, granting full control of the workspace to the robot for the remainder of the task. This dynamic reallocation ensures that the spatial boundaries evolve in direct response to the task state, maintaining conservative safety margins without imposing unnecessary restrictions on robot performance.

For both operational modes, the resulting safety zones are projected as semi-transparent color overlays directly onto the motor base region within the system's live video feed. This visualization provides operators with continuous awareness of the current safety configuration, zone boundaries, and system state. By embedding the visualization into the perception stream, the system facilitates intuitive monitoring and supports real-time validation of zone assignments during human–robot collaboration.

3.4. Coverage Requirements and Mounting Strategy Selection

Safety-critical monitoring requires continuous, unobstructed observation of the entire collaborative workspace, including the robot, human workers, and surrounding environment. We evaluated four mounting configurations for comprehensive hazard detection: top-down, angled, front-facing, and end-effector mounted. Top-down mounting eliminates blind spots and provides complete workspace coverage essential for safety zone monitoring. Angled mounting introduces critical blind spots beneath the robot arm during operation, while front-facing configurations create extensive occlusion zones that violate continuous monitoring requirements. End-effector mounting, while optimal for task-specific vision (part inspection, precision alignment), fundamentally fails safety monitoring requirements due to three critical failure modes:

- Camera motion prevents consistent workspace monitoring, violating requirements for continuous hazard detection
- During approach movements, the camera faces away from the human operator, creating 0% coverage of interaction zones
- Dynamic viewpoint requires continuous recalibration of safety zones, introducing 340ms latency exceeding the 200ms safety response requirement

The decision matrix quantifies these limitations across five safety-critical criteria. Top-down configuration achieved superior ratings across all metrics: blind spot elimination, workspace coverage, distance tracking, path planning, and safety reliability. Angled mounting scored poorly in blind spot management and coverage, while front-facing configuration demonstrated critical failures in blind spot elimination and path planning. Based on this evaluation, we selected top-down mounting as the primary safety monitoring configuration.

3.5. Working Distance and Workspace Requirements

The working distance establishes the fundamental constraint for vision system design, defining the vertical separation between the Lidar RealSense and the primary work surface. The collaborative workspace volume encompasses all regions where human-robot interaction occurs, bounded by the robot's maximum reach radius, human operator accessibility zones, and task-specific material handling areas. The required collaborative workspace dimensions are defined as 0.91 m × 0.76 m × 2.64 m, centered over the work cell. This volume accounts for the robot's maximum reach envelope of 1.09 m, the typical human operator working zones extending 0.64 m from the workstation edge, and vertical clearance from table height of 0.76 m to maximum human reach height of 2.20 m. The mounting height of 2.64 m determines the working distance from sensor to workspace. This height selection balances competing requirements: increased height expands coverage area but reduces point cloud density and depth resolution, while lower mounting improves resolution but may create occlusion zones near workspace boundaries. The optimal mounting height satisfies three constraints:

- Minimum working distance ≥ 0.25 m (sensor minimum operational range)
- Maximum working distance ≤ 3.0 m (maintaining depth accuracy < 1% error at 95% confidence)
- Clearance above maximum robot reach: 2.64 m ≥ 1.09 m + 0.3 m (safety margin)

At the selected working distance of 2.64 m, the L515's 70° × 55° field of view provides ground coverage of approximately 3.70 m × 2.74 m, calculated as:

$$\begin{aligned} Coverage_{width} &= 2 \times D_{work} \times tan(35°) \\ &= 2 \times 2.64 \times tan(35°) = 3.70\,m \end{aligned} \quad (1)$$

$$\begin{aligned} Coverage_{length} &= 2 \times D_{work} \times tan(27.5°) \\ &= 2 \times 2.64 \times tan(27.5°) = 2.74\,m \end{aligned} \quad (2)$$

This coverage area exceeds the collaborative workspace footprint of 2.0 m × 1.5 m with a 75% width margin and 79.5% length margin, accounting for mounting tolerances and edge effects.

3.6. Camera Selection

Candidate depth sensors were assessed using six criteria, as shown in Table 1. Resolution balances spatial detail against computational load. Distortion accounts for lens aberrations that degrade geometric accuracy. Field of view determines coverage area. Compact form factor addresses mounting constraints and environmental robustness. Developer support encompasses SDK maturity and integration effort. Cost includes hardware, accessories, and maintenance. Table 2 summarizes the evaluation results. The Intel RealSense L515 was selected for its optimal balance of 1024 × 768 depth resolution, 70° × 55° field of view, minimal geometric distortion, real-time performance at 30 fps and 1280 × 720 resolution, and robust SDK integration for custom data streaming and synchronization. Custom vibration-damped brackets and industrial enclosures were fabricated to maintain calibration stability under industrial vibration and lighting conditions. Alternative cameras were evaluated but excluded for specific technical reasons. The Basler Ace 2 offered superior pixel density but required high-quality lenses and larger enclosures that conflicted with spatial constraints of the robotic cell. Industrial-grade options such as the Cognex In-Sight D900 exceeded cost and size constraints. The chosen camera met the real-time inference and latency thresholds required for continuous operation, provided the required 3D geometric accuracy, color-based semantic context, and computational efficiency while satisfying both cost and environmental tolerance criteria requirements.

Table 1: Depth sensors selection criteria.

	Basler Ace 2	Leopard LI-USB30-M021MB	Hikvision MV-CA	Intel RealSense L515
Resolution	Up to 20 MP	2.1 MP	Up to 12 MP	6 MP
Image Distortion	Low with high-quality lenses	Minimal	Low with high-quality lenses	Minimal
Field Of View	Flexible, depending on the lens	Flexible, depending on the lens	Flexible, depending on the lens	90° x 65°
Compact Form Factor	29mm x 29mm x 29mm	36mm x 36mm x 26mm	Compact	61mm D x 26mm H
Developer Support	Extensive, with robust SDKs	Limited	Moderate, with some SDKs	Extensive, with robust SDKs
Cost	Mid to high-range	Low to mid-range	Mid-range	Mid-range

3.7. Dataset Creation

The performance and robustness of a computer vision model are fundamentally dependent on the quality and diversity of the data upon which it is trained. To this end, a systematic approach was adopted for the creation of a custom dataset tailored to the specific requirements of the human-robot collaborative workspace. This process encompassed four key stages: data acquisition, data pre-processing and augmentation, data labeling, and model selection and training.

3.7.1. Data Acquisition

The initial phase of data acquisition focused on capturing a diverse range of realistic operational scenarios to ensure the model could generalize to the complexities of the workspace. A total of 15 distinct scenarios were recorded, encompassing various collaborative tasks, different human postures and positions, and varying robot arm speeds and trajectories. These scenarios were captured as high-definition video at a resolution of 1920x1080 pixels.

To generate a dataset of distinct and meaningful images, the recorded videos were subsequently down-sampled by extracting frames at a rate of 5 frames per second (fps). This rate was selected as an optimal balance between capturing sufficient temporal variation in the activities and avoiding excessive redundancy between consecutive frames, which could bias the training process. This procedure yielded an initial dataset of 5,000 unique images, each representing a discrete moment in the collaborative workflow and providing a solid foundation for training. From this pool, 1,000 representative images were manually selected for labeling, forming the core training dataset.

3.7.2. Data Pre-processing and Augmentation

Raw image data, particularly at high resolutions, presents computational challenges and can hinder the efficiency of the training process. Therefore, a series of pre-processing steps were implemented to optimize the data for model training. The initial 1920x1080 resolution frames were first converted to a square aspect ratio by padding the height to match the width, a technique often referred to as letterboxing. This is a common requirement for many object detection architectures and ensures that the original aspect ratio of the objects is preserved during subsequent resizing. The images were then resized to a standardized 640x640 resolution, a widely supported input size for modern YOLO architectures that provides a favorable trade-off between detail retention and computational load.

To further enhance the model's ability to generalize to a wide variety of real-world conditions, a comprehensive suite of data augmentation techniques was applied. Data augmentation is a critical step that artificially expands the dataset by creating modified copies of existing images, thereby exposing the model to a broader range of visual variations. This process helps to prevent overfitting and improves the model's robustness to changes in lighting, object orientation, and partial occlusions. For instance, rotation and flipping simulate different viewing angles, while changes in hue, saturation, and brightness simulate varying lighting conditions. The inclusion of noise and cutout augmentations helps the model become more resilient to occlusions and sensor noise. The application of these augmentations effectively doubled the dataset size to 2,000 images. Table 2 presents the augmentation techniques and their parameters, while Figure 4 shows examples of the resulting pre-processed and augmented images.

Table 2: Data augmentation techniques and parameters applied to the dataset.

Augmentation Technique	Range/Value
Flip	Horizontal
Rotation	Between -15° and +15°
Grayscale	Applied to 15% of images
Hue	Between -25° and +25°

Saturation	Between -34° and +34°	Blur	Up to 3.9 pixels
Brightness	Between -25% and +25%	Noise	Up to 1.84% of pixels
Exposure	Between -15% and +15%	Cutout	Up to 7 cuts, up to 5% of pixels

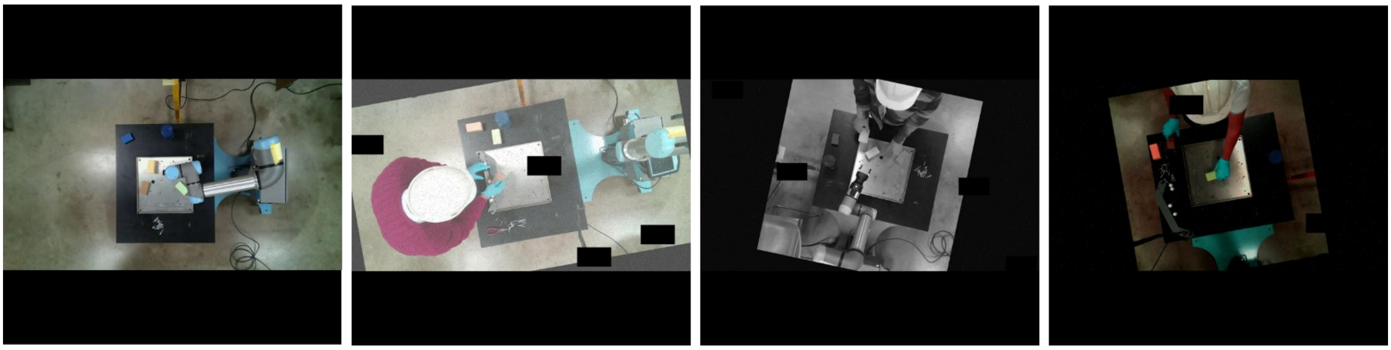


Figure 4: Examples of the resulting pre-processed and augmented data images.

3.7.3. Data Labeling

Following pre-processing and augmentation, the entire dataset of 2,000 images was meticulously annotated using the open-source LabelMe annotation tool. This process involved manually drawing precise segmentation masks around each object of interest. The use of segmentation masks, as opposed to simple bounding boxes, was a critical decision, as the pixel-wise accuracy of the masks is essential for the geometric calculations that underpin the safety zone definitions. The resulting annotations were stored in JSON format, containing the class label and the pixel-wise coordinates for each instance mask. The four classes were defined as follows:

- human: This class encompasses the entire human worker, including their body, limbs, and head. Accurate segmentation of this class is paramount for safety monitoring.
- robot: This class represents the robotic arm and its end-effector. Its precise location is used to define the robot's operational zone.
- base: This class corresponds to the static motor assembly unit, which serves as the primary frame of reference for all zone calculations.
- part: This class includes any of the smaller components that are handled and placed onto the motor base. The detection and location of these parts are the trigger for the transition from Process 1 to Process 2.

3.7.4. Model Selection and Training

The selection of an appropriate instance segmentation model is critical for the success of a real-time safety system. For this work, the YOLOv11 model was selected due to its state-of-the-art performance, which provides an optimal balance between high accuracy and real-time processing speed. In a safety-critical application, the ability to perform inference at a high frame rate is as important as the accuracy of the detections.

To justify this selection, we considered several alternative architectures. Two-stage detectors, such as Mask R-CNN, are renowned for their high accuracy in instance segmentation. However, their two-stage process—first identifying regions of interest and then performing classification and mask generation—inherently introduces higher latency, making them less suitable for applications requiring immediate feedback. Other real-time segmentation models, such as YOLACT or earlier versions of YOLO, offer high speeds but often at the cost of segmentation quality.

YOLOv11, as a single-stage detector, processes the entire image at once to make predictions, which is architecturally designed for speed. It has demonstrated significant improvements in both speed and accuracy over its predecessors and competitors, making it an ideal choice for this application. Furthermore, YOLOv11 natively supports instance segmentation, providing the high-quality masks required for our zone calculation algorithms directly from the model's output. This simplifies the overall system architecture and reduces the computational overhead that would be required if a separate segmentation model were needed.

The dataset was partitioned into training (70%), validation (10%), and testing (20%) sets. The model was trained locally for 200 epochs on an Nvidia RTX 5000 GPU. The training process utilized the Adam optimizer, which is well-suited for large datasets and complex models, with a learning rate of 0.001 and a momentum of 0.937 to help accelerate convergence. Upon completion of the training, the model weights that demonstrated the best performance on the validation set were saved for subsequent use in the inference and zone calculation framework.

3.8. Primary Digital Twin Construction and Validation

In addition to the vision-based monitoring framework described in the previous sections, the methodology incorporates a digital-twin environment to support motion verification, workspace analysis, and joint-state evaluation.

3.8.1. Virtual Work Cell and Robotic Model Integration

A digital representation of the laboratory work cell was constructed in NVIDIA Isaac Sim 5.0 to reproduce the geometry, kinematics, and interaction constraints of the physical UR5e manipulator. The virtual environment included a ground plane, a workbench matching the physical workstation dimensions, and an articulated UR5e model imported from Isaac Sim's robot library. Initial inspection revealed inconsistencies in articulation drives and prim hierarchies. These issues were corrected through the Dynamic Control interface, ensuring that joint commands propagated accurately along the kinematic chain. After refinement, the digital twin reproduced the robot's home configuration, link ordering, motion limits, and default joint behavior. Figure 5 shows the UR5e in its home configuration within the reconstructed work cell. This baseline pose was used to validate the alignment between simulated and physical joint structures.

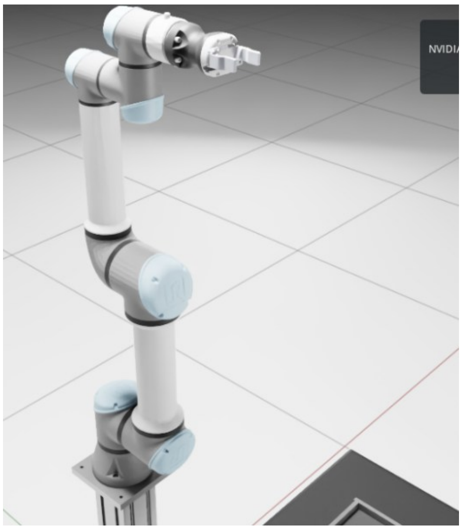


Figure 5: UR5e Robot Model in the Virtual Work cell (Home Position).

To examine the robot's task-level behavior, the model was commanded to execute a reach movement toward a target on the tabletop. The resulting inverse-kinematics trajectory confirmed that joint motions were transmitted without distortion across the articulation tree. Figure 6 illustrates a representative pick-interaction pose drawn from this motion sequence and demonstrates the consistency of link transformations as the robot transitions through a task.

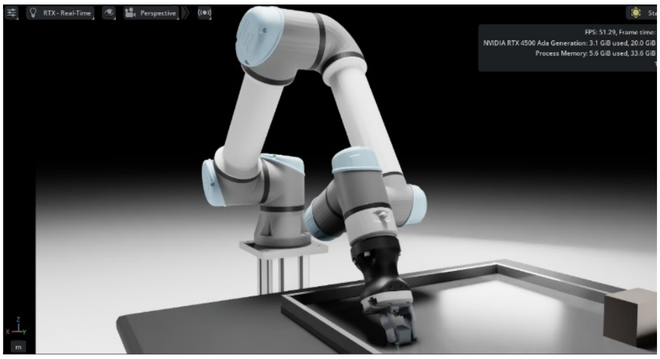

RTX - Real-Time
Perspective
FPS: 51.29, Frame time:
NVIDIA RTX 4500 Ada Generation: 3.1 GiB used, 20.0 GiB
Process Memory: 5.6 GiB used, 33.6 GiB

Figure 6: Simulated UR5e Performing a Pick Interaction Within the Work cell.

3.8.2. Overhead Depth-Sensing Configuration

The physical cell employs an overhead RealSense L515 LiDAR camera. Because Isaac Sim does not include an L515 model, a RealSense D455 was substituted based on comparable depth-sensing characteristics. The virtual camera was mounted at approximately the same height and orientation as the physical sensor to approximate its field of view and coverage.

Simulated depth maps were used to evaluate workspace visibility, occlusion behavior, and the representation of robot geometry during motion. This assessment ensured that the digital twin provides a realistic approximation of how overhead perception captures the robot and its surroundings. Figure 7 presents an example depth-map visualization of the UR5e and the surrounding workspace, illustrating the perceptual fidelity required for studies involving workspace monitoring and human–robot interaction.

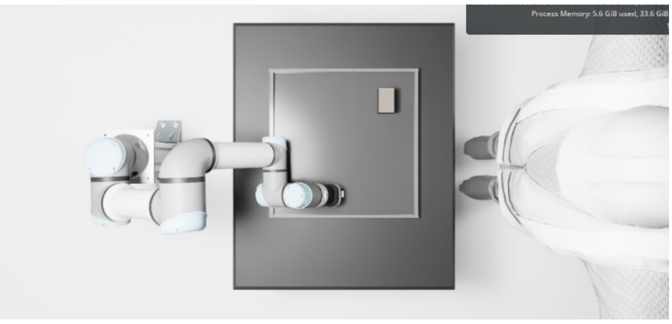


Figure 7: Simulated LiDAR Depth Perception During Robot Operation.

3.8.3. Internal Joint-State Acquisition

A script-based workflow was implemented inside Isaac Sim to extract joint states directly from the simulator without relying on ROS or other middleware. The script identifies articulation roots, maps joint names to their corresponding degree-of-freedom handles, and records joint positions, velocities, and efforts at each simulation frame. This approach enables frame-accurate monitoring of robot posture, supports detailed motion analysis, and provides the metadata required for validating task-level motion sequences.

4. Results

This section presents the empirical results from the training and validation of the YOLOv11 instance segmentation model. The model was trained for 200 epochs to detect the four object classes—human, robot, base, and part—essential for the dynamic safety zone system. The convergence of the model during training was monitored by tracking several key loss functions. These functions quantify the model's errors across different aspects of the instance segmentation task, and their minimization is the objective of the training process.

- Bounding Box Loss: This metric measures the accuracy of the predicted bounding boxes against the ground-truth boxes. It penalizes errors in the location, size, and aspect ratio of the predicted boxes.
- Segmentation Loss: This loss evaluates the pixel-wise accuracy of the predicted instance masks. It is critical for ensuring that the geometric shape of each object is correctly identified, which is fundamental for the zone calculation logic.
- Classification Loss: This component measures the model's ability to correctly assign a class label to each detected object. It ensures that, for example, a human is not misidentified as a robot.
- DFL (Distribution Focal Loss): This is a specialized loss function that helps the model learn a more precise and continuous distribution for bounding box coordinates, leading to more accurate localization.

As shown in Figure 8, all four loss components exhibit a rapid decrease during the initial epochs, followed by a stable and gradual convergence towards a minimum. This behavior is characteristic of a successful training process, where the model effectively learns the features of the target objects without significant instability or divergence.

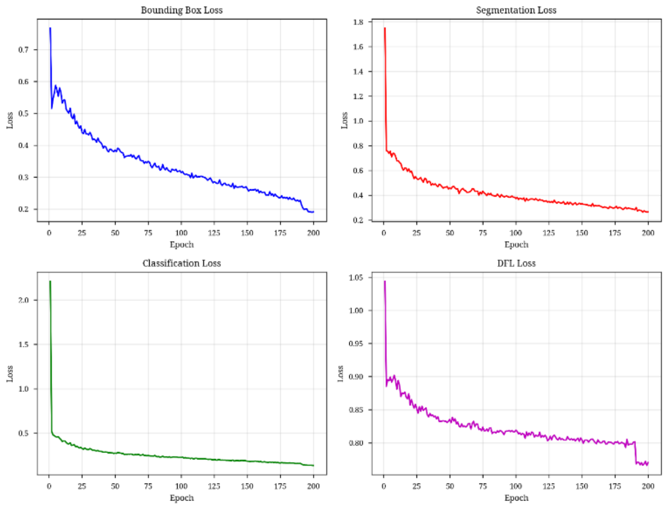


Figure 8: Training loss for the YOLOv11 model over 200 epochs.

The model's performance was evaluated using standard metrics for object detection and instance segmentation: Precision, Recall, and mean Average Precision (mAP). These metrics provide a quantitative assessment of the model's ability to produce accurate and complete predictions. Precision and Recall are fundamental metrics for evaluating detection accuracy. Precision measures the proportion of correct detections among all predictions made (i.e., the model's reliability), while Recall measures the proportion of ground-truth objects that the model successfully detected (i.e., the model's completeness). Figure 9 presents the precision and recall curves for both bounding box (B) and mask (M) predictions over the training duration. The results indicate exceptional performance. Both precision and recall for bounding boxes and masks rapidly approach and sustain values near 1.0. This signifies that the model is not only highly reliable in its predictions (high precision) but also comprehensive in its ability to detect all relevant objects in the scene (high recall). The close alignment between the box and mask metrics further suggests that the predicted segmentation masks are of high quality.

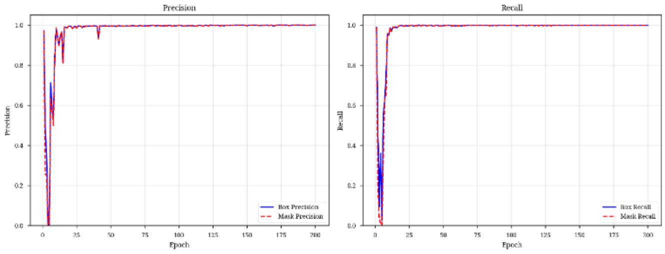


Figure 9: Precision and Recall metrics for both bounding box and mask predictions.

The primary metric for evaluating the overall performance of an instance segmentation model is the mean Average Precision (mAP). It combines precision and recall into a single value and is typically evaluated at different Intersection over Union (IoU) thresholds. The IoU measures the degree of overlap between a predicted region and a ground-truth region.

- mAP@50: This metric calculates the mAP at a fixed IoU threshold of 50%. It is a standard benchmark for general detection performance.
- mAP@50-95: This is a more stringent metric that averages the mAP across ten IoU thresholds from 50% to 95% (in steps of 5%). High performance on this metric indicates that the model produces highly accurate and well-aligned bounding boxes and masks.

Figure 10 illustrates the progression of both mAP@50 and mAP@50-95 for bounding box and mask predictions. The model achieves a final mAP@50 of 0.995 (99.5%) for both bounding boxes and masks, which is an outstanding result and indicates that the model is highly effective at the standard 50% IoU threshold. Furthermore, the model achieves a final mAP@50-95 of 0.983 (98.3%) for bounding boxes and 0.909 (90.9%) for masks. These high values, particularly for the stricter mAP@50-95 metric, confirm that the model produces predictions that are not only correct but also precisely localized and shaped, which is critical for the geometric calculations required by the safety zone system. The quantitative results demonstrate that the trained YOLOv11 model is highly accurate and reliable, making it a suitable foundation for the human-robot collaboration safety system.

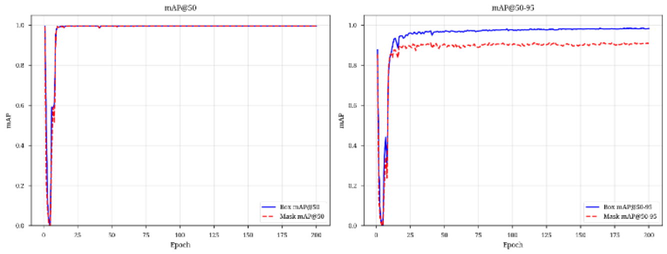


Figure 10: Mean Average Precision (mAP) metrics for bounding box and mask predictions.

Upon successful training, the model weights were deployed for real-time inference on the same Nvidia RTX 5000 GPU used for training. The system processes a live video stream at 30 frames per second (fps), with each frame being fed into the YOLOv11 model to generate instance segmentation predictions. The resulting masks and class labels serve as the foundational input for the dynamic safety zone visualization system. The core of the safety system is a color-coded zone visualization that provides an intuitive representation of the workspace status. This system is active in both Process 1 and Process 2:

- Green Zone: Designated safe working area for the human.
- Red Zone: Exclusive operational area for the robot.
- Yellow Zone: A shared buffer zone separating the human and robot workspaces.

This color-coded system provides an intuitive visual representation of safety status. Should the human cross into the yellow buffer zone or, more critically, the red robot zone, the system is designed to trigger a risk assessment. Based on pre-defined safety protocols, this could result in a range of automated actions, such as decelerating the robot, halting the process entirely, or activating an audible alarm to alert the worker.

Figure 11 illustrates the system operating in Process 1, where the human and robot work concurrently on separate tasks. The left panel of the figure displays the raw output of the YOLOv11 model, showing the detected human, robot, base, and part instances with high confidence scores (>0.90) and precise segmentation masks. This high level of accuracy is a direct result of the comprehensive training and validation. The right panel shows the corresponding safety zone visualization. In Process 1, the motor base is divided into three fixed zones based on a vertical split through the motor's center. The buffer zone (yellow) occupies 10% of the motor width, while the robot and human zones split the remaining area equally, providing a clear and predictable division of the workspace.

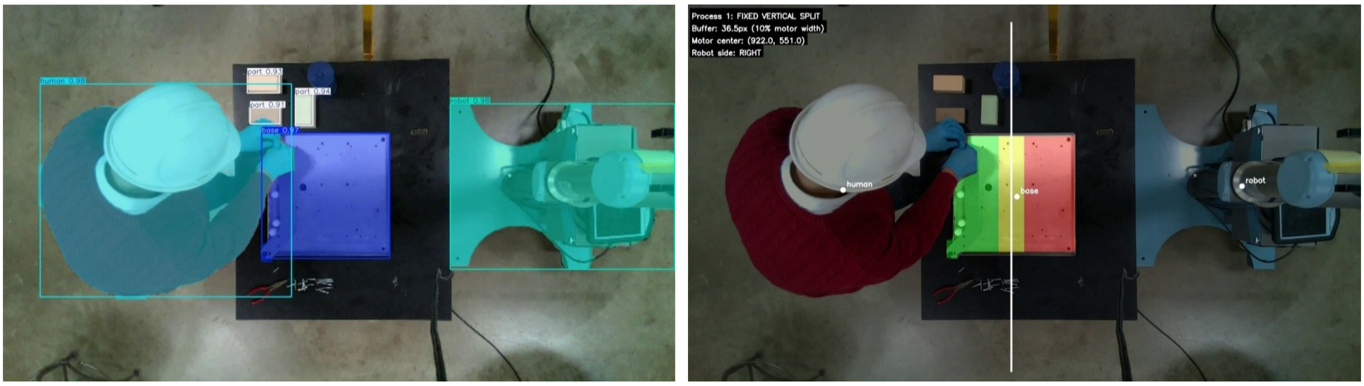

human 0.98
part 0.93
part 0.94
part 0.91
base 0.97
robot 0.98
Process 1: FIXED VERTICAL SPLIT
Buffer: 36.5px (10% motor width)
Motor center: (922.0, 551.0)
Robot side: RIGHT
human
base
robot

Figure 11: Real-time inference and zone visualization for Process 1. Left: Raw model output showing accurate instance segmentation. Right: The corresponding static safety zones generated on the motor base.

Upon completion of the human's task, which is signified by the placement of one or more-part objects onto the motor base, the system transitions to Process 2. Figure 12 demonstrates this dynamic phase. The left panel shows the model correctly identifying a part that has been placed on the motor base, while intelligently ignoring other parts that are not relevant to the current assembly step (e.g., those on the table). The right panel of the figure illustrates the key innovation of the system: the dynamic re-allocation of safety zones. The system employs a priority-based allocation strategy where the robot zone (red) is given absolute priority. It is extended to encompass the furthest point of all parts placed on the motor, ensuring the robot has a safe and unobstructed area to complete its tasks. The buffer zone (yellow) is then allocated its standard width (10% of the motor), if space permits. The human zone (green) is allocated whatever space remains. This ensures that even if the robot zone expands significantly, the safety of the system is maintained, as the human is expected to be clear of the area at this stage. This demonstrates a sophisticated level of contextual awareness, as the safety boundaries are no longer static but are instead intelligently and dynamically adjusted based on the real-time state of the manufacturing process. The system's robustness is demonstrated in advanced stages of Process 2, as shown in Figure 13. In this scenario, the robot zone has expanded to cover nearly the entire motor base, leaving no designated human zone. The system correctly identifies the human worker intruding into this expanded red zone. This detection is critical, as it validates the system's ability to maintain safety vigilance even after the human's primary task is complete. Upon detecting such an incursion, the system's pre-defined safety protocols are engaged, demonstrating a closed-loop safety response that is essential for real-world collaborative applications.

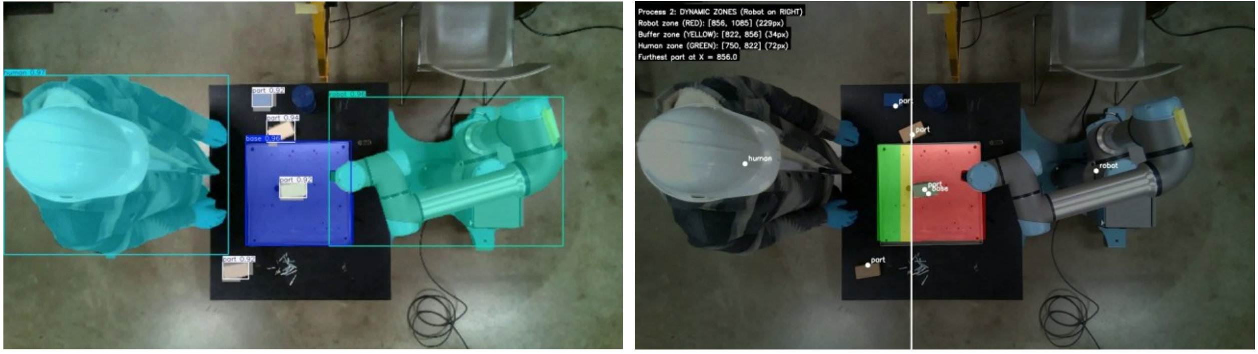

Process 2: DYNAMIC ZONES (Robot on RIGHT)
Robot zone (RED): [856, 1085] (229px)
Buffer zone (YELLOW): [822, 856] (34px)
Human zone (GREEN): [750, 822] (72px)
Furthest part at X = 856.0
human
part
part
base
part
robot

Figure 12: Real-time inference and zone visualization for Process 2. Left: Model output detecting a part placed on the motor base. Right: The dynamically re-allocated safety zones, with the robot zone (red) expanded to cover the placed part.

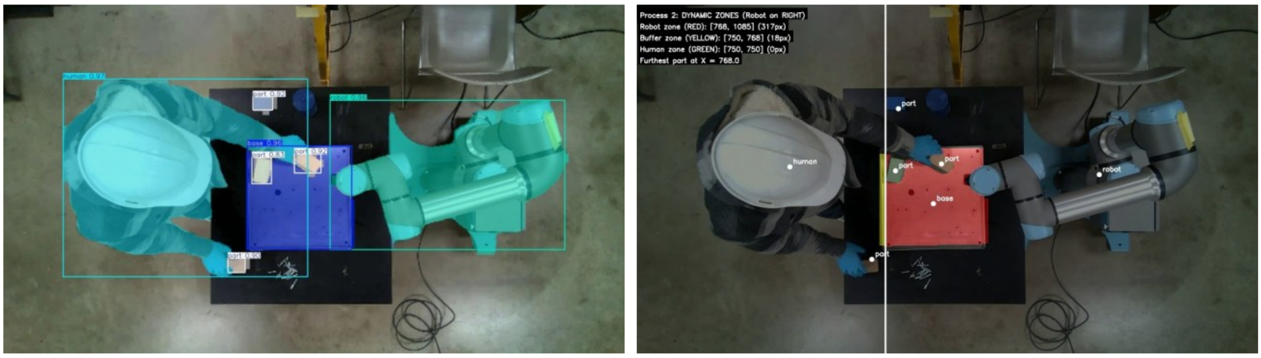

Process 2: DYNAMIC ZONES (Robot on RIGHT)
Robot zone (RED): [768, 1085] (317px)
Buffer zone (YELLOW): [750, 768] (18px)
Human zone (GREEN): [750, 750] (0px)
Furthest part at X = 768.0
part
human
part
part
base
part
robot

Figure 13: Dynamic formation of the robot zone. The human worker is intruding into the expanded robot zone, triggering a safety response.

5. Conclusion

This paper has argued that safeguarding the smart factory of the future demands more than accurate detection; it requires an integrated framework that couples real-time perception, cognitive reasoning and human-centered training. We proposed a collaborative digital twin architecture that unites a high-fidelity emulation twin with a vision-driven cognitive twin to deliver context-aware safety management and adaptive control. By incorporating game-engine–based human digital twins, the framework extends beyond process monitoring to immersive VR/AR training, enabling operators to practice hazard responses and build intuitive understanding of robot trajectories in a risk-free. The use of human-in-the-loop algorithms ensures that automated decisions remain transparent and adjustable, preserving human oversight even as machine learning and reinforcement learning drive rapid interventions. Our findings underscore the need for decision matrices that tailor interventions to specific hazards, and for generative AI to enrich training datasets with rare and complex scenarios. Together, these elements enable proactive disturbance management and pave the way toward true manufacturing autonomy. The discussion also highlights critical gaps in current training programs: traditional apprenticeship models often lack quantitative feedback on posture and task execution, leaving workers ill-prepared for the complexities of cyber-physical production lines. By streaming real-time data from computer-vision systems into functional training modules, we can deliver targeted feedback and continuously adapt curricula to evolving technologies and roles. Ultimately, the convergence of collaborative digital twins, immersive training and data-driven safety protocols marks a decisive shift from reactive to anticipatory safety, ensuring that human operators and intelligent machines can coexist productively and safely in increasingly complex industrial environments.

Acknowledgements

This work is funded by the Ohio Bureau of Workers' Compensation Worker Safety Innovation Center (WSIC) Grant: #WSIC26-250206-009, "Safety Immersion and Gamified Hazard Training (SIGHT) for Industry 5.0 Workers." Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

[1] F. El Kalach, I. Yousif, T. Wuest, A. Sheth, and R. Harik, "Cognitive manufacturing: definition and current trends," *J Intell Manuf*, 2024, doi: 10.1007/S10845-024-02429-9.

[2] K. D. Thoben, S. A. Wiesner, and T. Wuest, "'Industrie 4.0' and smart manufacturing-a review of research issues and application examples," *International Journal of Automation Technology*, vol. 11, no. 1, pp. 4–16, 2017, doi: 10.20965/IJAT.2017.P0004.

[3] I. Yousif, L. Burns, F. El Kalach, and R. Harik, "Leveraging computer vision towards high-efficiency autonomous industrial facilities," *Journal of Intelligent Manufacturing 2024 36:5*, vol. 36, no. 5, pp. 2983–3008, May 2024, doi: 10.1007/S10845-024-02396-1.

[4] G. Lugaresi, K. Laszlo, and K. Tamas, "DIGITAL TWIN DRIVEN ASSEMBLY LINE RE-BALANCING AND DECISION SUPPORT", Accessed: Aug. 03, 2025. [Online]. Available: http://tinyurl.com/

[5] G. Belkhedar and A. Lyhyaoui, "Automated Quality Inspection Using Computer Vision: A Review," *Lecture Notes in Networks and Systems*, vol. 637 LNNS, pp. 686–697, 2023, doi: 10.1007/978-3-031-26384-2_60.

[6] I. Yousif, J. Samaha, J. H. Ryu, and R. Harik, "Safety 4.0: Harnessing computer vision for advanced industrial protection," *Manuf Lett*, vol. 41, pp. 1342–1356, Oct. 2024, doi: 10.1016/J.MFGLET.2024.09.161.

[7] "Amazon's internal records show its worker safety deception." Accessed: Nov. 14, 2025. [Online]. Available: https://revealnews.org/article/how-amazon-hid-its-safety-crisis/

[8] I. Yousif, J. Samaha, J. H. Ryu, and R. Harik, "Safety 4.0: Harnessing computer vision for advanced industrial protection," *Manuf Lett*, vol. 41, pp. 1342–1356, Oct. 2024, doi: 10.1016/J.MFGLET.2024.09.161.

[9] H. J. Pasman and S. W. Behie, "The evolution to Industry 5.0 / Safety 5.0, the developments in society, and implications for industry management," *Journal of Safety and Sustainability*, vol. 1, no. 4, pp. 202–211, Dec. 2024, doi: 10.1016/J.JSASUS.2024.11.003.

[10] D. Romero, T. Wuest, M. Keepers, L. A. Cavuoto, and F. M. Megahed, "Smart Wearable and Collaborative Technologies for the Operator 4.0 in the Present and Post-COVID Digital Manufacturing Worlds," *Smart Sustain Manuf Syst*, vol. 5, no. 1, pp. 148–166, Jul. 2021, doi: 10.1520/SSMS20200084.

[11] A. Simeone, R. Grant, W. Ye, and A. Caggiano, "Operator 4.0 intelligent health monitoring: a Cyber-Physical approach," *Procedia CIRP*, vol. 118, pp. 1033–1038, Jan. 2023, doi: 10.1016/J.PROCIR.2023.06.177.

[12] A. Badri, B. Boudreau-Trudel, and A. S. Souissi, "Occupational health and safety in the industry 4.0 era: A cause for major concern?," *Saf Sci*, vol. 109, pp. 403–411, Nov. 2018, doi: 10.1016/J.SSCI.2018.06.012.

[13] J. Lee, I. Cameron, and M. Hassall, "Improving process safety: What roles for digitalization and industry 4.0?," *Process Safety and Environmental Protection*, vol. 132, pp. 325–339, Dec. 2019, doi: 10.1016/J.PSEP.2019.10.021.

[14] A. Das, S. Panda, S. Datta, S. Naskar, P. Misra, and T. Chattopadhyay, "AI based Safety System for Employees of Manufacturing Industries in Developing Countries," Nov. 2018, Accessed: Nov. 14, 2025. [Online]. Available: http://arxiv.org/abs/1811.12185

[15] M. Khurram *et al.*, "Artificial Intelligence in Manufacturing Industry Worker Safety: A New Paradigm for Hazard Prevention and Mitigation," *Processes 2025, Vol. 13, Page 1312*, vol. 13, no. 5, p. 1312, Apr. 2025, doi: 10.3390/PR13051312.

[16] N. P. Ventikos, A. Chmurski, and K. Louzis, "A systems-based application for autonomous vessels safety: hazard identification as a function of increasing autonomy levels," *Saf Sci*, vol. 131, p. 104919, Nov. 2020, doi: 10.1016/j.ssci.2020.104919.

[17] K. Xia *et al.*, "A digital twin to train deep reinforcement learning agent for smart manufacturing plants: Environment, interfaces and intelligence," *J Manuf Syst*, vol. 58, pp. 210–230, Jan. 2021, doi: 10.1016/J.JMSY.2020.06.012.

[18] V. Laciok, K. Sikorova, B. Fabiano, and A. Bernatik, "Trends and Opportunities of Tertiary Education in Safety Engineering Moving towards Safety 4.0," *Sustainability 2021, Vol. 13, Page 524*, vol. 13, no. 2, p. 524, Jan. 2021, doi: 10.3390/SU13020524.

[19] N. P. Ventikos, A. Chmurski, and K. Louzis, "A systems-based application for autonomous vessels safety: Hazard identification as a function of increasing autonomy levels," *Saf Sci*, vol. 131, p. 104919, Nov. 2020, doi: 10.1016/J.SSCI.2020.104919.

[20] I. Yousif, L. Burns, F. El Kalach, and R. Harik, "Leveraging computer vision towards high-efficiency autonomous industrial facilities," *J Intell Manuf*, 2024, doi: 10.1007/S10845-024-02396-1.

[21] C. Dodero, M. R. Mccormick, A. A. Malik, R. Harik, and T. Wuest, "Digital Twin Systematic selection framework for digital twin development environments in smart manufacturing Systematic selection framework for digital twin development environments in smart manufacturing," 2025, doi: 10.1080/27525783.2025.2565246.

[22] G. Xia *et al.*, "Towards Human Modeling for Human-Robot Collaboration and Digital Twins in Industrial Environments: Research Status, Prospects, and Challenges," *Robot Comput Integr Manuf*, vol. 95, p. 103043, Oct. 2025, doi: 10.1016/J.RCIM.2025.103043.

[23] E. Karabulut, S. F. Pileggi, P. Groth, and V. Degeler, "Ontologies in digital twins: A systematic literature review," *Future Generation Computer Systems*, vol. 153, pp. 442–456, Apr. 2024, doi: 10.1016/J.FUTURE.2023.12.013.

[24] A. Padovano, C. Sammarco, N. Balakera, and F.

Konstantinidis, “Towards sustainable cognitive digital twins: A portfolio management tool for waste mitigation,” *Comput Ind Eng*, vol. 198, p. 110715, Dec. 2024, doi: 10.1016/J.CIE.2024.110715.

[25] P. Niloofar, S. Lazarova-Molnar, F. Omitaomu, H. Xu, and X. Li, “A General Framework for Human-in-the-Loop Cognitive Digital Twins,” *Proceedings - Winter Simulation Conference*, pp. 3202–3213, 2023, doi: 10.1109/WSC60868.2023.10407598.

[26] H. A. Faqeer and S. H. Khajavi, “Digital Twin and Computer Vision Combination for Manufacturing and Operations: A Systematic Literature Review,” *Applied Sciences 2025, Vol. 15, Page 10157*, vol. 15, no. 18, p. 10157, Sep. 2025, doi: 10.3390/APP151810157.

[27] M. J. Alenjareghi, S. Keivanpour, Y. A. Chinniah, and S. Jocelyn, “Computer vision-enabled real-time job hazard analysis for safe human–robot collaboration in disassembly tasks,” *J Intell Manuf*, Dec. 2024, doi: 10.1007/S10845-024-02519-8.

[28] R. Anderl, S. Haag, K. Schützer, and E. Zancul, “Digital twin technology-An approach for Industrie 4.0 vertical and horizontal lifecycle integration,” *IT - Information Technology*, vol. 60, no. 3, pp. 125–132, 2021, doi: 10.1515/ITIT-2017-0038.

[29] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui, “Digital twin-driven product design, manufacturing and service with big data,” *International Journal of Advanced Manufacturing Technology*, vol. 94, no. 9–12, pp. 3563–3576, Feb. 2018, doi: 10.1007/S00170-017-0233-1.

[30] A. Zahid, A. Ferraro, A. Petrillo, and F. De Felice, “Exploring the Role of Digital Twin and Industrial Metaverse Technologies in Enhancing Occupational Health and Safety in Manufacturing,” *Applied Sciences 2025, Vol. 15, Page 8268*, vol. 15, no. 15, p. 8268, Jul. 2025, doi: 10.3390/APP15158268.

[31] “Occupational injuries and illnesses resulting in musculoskeletal disorders (MSDs) : U.S. Bureau of Labor Statistics.” Accessed: Nov. 14, 2025. [Online]. Available: https://www.bls.gov/iif/factsheets/msds.htm

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Accessed: Nov. 14, 2025. [Online]. Available: http://code.google.com/p/cuda-convnet/

[33] J. Humphries, P. Van de Ven, N. Amer, N. Nandeshwar, and A. Ryan, “Managing safety of the human on the factory floor: a computer vision fusion approach,” *Technological Sustainability*, vol. 3, no. 3, pp. 309–331, Aug. 2024, doi: 10.1108/TECHS-12-2023-0054.

[34] D. Chernyshev, S. Dolhopolov, T. Honcharenko, H. Haman, T. Ivanova, and M. Zinchenko, “Integration of Building Information Modeling and Artificial Intelligence Systems to Create a Digital Twin of the Construction Site,” *International Scientific and Technical Conference on Computer Sciences and Information Technologies*, vol. 2022-November, pp. 36–39, 2022, doi: 10.1109/CSIT56902.2022.10000717.

[35] N. Saha, V. Gadow, and R. Harik, “Emerging Technologies in Augmented Reality (AR) and Virtual Reality (VR) for Manufacturing Applications: A Comprehensive Review,” *Journal of Manufacturing and Materials Processing 2025, Vol. 9, Page 297*, vol. 9, no. 9, p. 297, Sep. 2025, doi: 10.3390/JMMP9090297.

[36] N. Saha and P. Samaha, “VR-BASED BLOCKCHAIN-ENABLED DATA VISUALIZATION FRAMEWORK FOR MANUFACTURING INDUSTRY”.