JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

Prompt-Guided AI Analytics for Manufacturing Digital Twins: Adapting YOLO-World to Industrial Object Detection

Yuxuan Tang[1]; Wenjie Han[2]; Min Liu[3]; Lei Zhang[4, *]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, Heilongjiang, China

[2] School of Mechanical Engineering, Shenyang University of Technology, Shenyang 110870, Liaoning, China

[3] School of Information and Mechanical Engineering, Wuhan Polytechnic University, Wuhan 430023, Hubei, China

[4] School of Mechatronic Engineering, Xi'an Polytechnic University, Xi'an 710048, Shaanxi, China

* Corresponding author: zhanglei@xpu.edu.cn | ORCID: 0000-0002-3186-7415

ARTICLE INFO
Received
October 16, 2024
Revised
December 17, 2024
Accepted
February 14, 2025
Available Online
March 30, 2025
DOI
10.63646/jaiaa.2025.030104

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Manufacturing digital twins promise to bridge physical production and data-driven decision making, yet the data side of the bridge remains underdeveloped because vision sensors require large, well-annotated datasets that industrial sites cannot easily provide. This paper investigates how prompt-guided vision–language detection can be adapted to deliver such data with limited supervision. We propose a Sim2Real pipeline that fuses photorealistic synthetic imagery generated from CAD models with a small body of coot-captured controlled imagery, then fine-tunes the open-vocabulary YOLO-World detector on the resulting mixed dataset. The pipeline is validated on a tidal-turbine workstation containing seven assembly states and is benchmarked against closed-vocabulary YOLOv8 baselines and against synthetic-only training. On 1,024 spontaneous production frames, the prompt-guided model attains mAP@0.5 = 0.579 and precision = 0.815 after a brief 943-image spontaneous fine-tune, raising mean average precision by 25.4 percentage points over the synthetic controlled-only configuration and by 18.0 points over a YOLOv8-S baseline trained on the same pool. Open-vocabulary prompts add a further 5–6 map points by exposing color and assembly-state attributes that were never explicit class labels at training time. Annotation effort is cut by 77% relative to a fully manual real-data pipeline. The paper concludes with a practitioner-oriented discussion of how the pipeline integrates into manufacturing digital twins and where its limits lie.

Keywords: Digital twin; computer vision; YOLO-World; Sim2Real; open-vocabulary detection; vision–language model; manufacturing analytics; assembly monitoring

I. INTRODUCTION

The vision of intelligent manufacturing pushed forward by Industry 4.0 depends on tight feedback between physical operations and their data-driven counterparts (Lu, 2017a; Tao et al., 2018; Liu et al., 2025). At the heart of this feedback sits the digital twin, a virtual model that co-evolves with the production system through continuous data exchange. When that data exchange is rich enough, twins move beyond visualization into prediction, anomaly detection, and adaptive control (Grieves & Vickers, 2017; Tao et al., 2019; Fuller et al., 2020). When the exchange is shallow, the twin degenerates into a static dashboard, and the strategic value collapses. Closing the data gap is therefore the central engineering problem in operationalizing twins for real

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

factories.

Conventional sensing (PLCs, OPC-UA streams, RFID, MES events) supplies reliable but narrow signals: machine state, throughput counters, timestamped events. These signals describe what happened but rarely what the workstation looked like — which part was on the bench, how the operator placed it, whether the assembly had been completed correctly (Sisinni et al., 2018; Boyes et al., 2018; Hassija et al., 2019). Computer vision can recover that missing context. A camera observing a workcell sees component identity, spatial relationships, operator interaction, and assembly state, all of which directly enrich the twin's descriptive and predictive capabilities (Sheuly et al., 2022; Lu & Xu, 2019; Zhang & Lu, 2021). Yet integrating vision into industrial twins is hard for two reasons that are easy to state and difficult to solve.

The first reason is data scarcity. State-of-the-art object detectors (Redmon et al., 2016; Bochkovskiy et al., 2020; Wang et al., 2023) reach their published accuracy only after training on tens of thousands of well-annotated images. In an industrial setting, such datasets cannot be assembled cheaply: production cannot be paused for capture, operators cannot be photographed without consent, and proprietary geometries cannot leave the plant. The second reason is vocabulary rigidity. Closed-vocabulary detectors must be retrained whenever a new component variant, a new color, or a new assembly state appears, which conflicts with the inherent variability of real production lines (Cheng et al., 2024; Liu et al., 2024; Gu et al., 2022).

Two recent technical advances suggest a way around both obstacles. Photorealistic synthetic data generation made tractable by Unity-style renderers and the Perception Package (Borkman et al., 2021; Tobin et al., 2017; Tremblay et al., 2018), allows an unlimited supply of automatically annotated CAD-derived images. Open-vocabulary detection, introduced by methods such as ViLD (Gu et al., 2022), GLIP (Li et al., 2022b), Grounding DINO (Liu et al., 2024), and YOLO-World (Cheng et al., 2024), allows a single model to recognize objects described by free-form text rather than a fixed taxonomy. Combined, these advances suggest a pipeline in which synthetic data eliminates the labelling bottleneck and prompt-guided detection eliminates the retraining-for-every-variant bottleneck. The combination, however, has not yet been studied at the granularity required for real assembly workstations, nor has it been benchmarked against the residual sum-to-real gap that is well documented in the domain randomization literature (Peng et al., 2018; Eversberg & Lambrecht, 2021; Rawal et al., 2023).

This paper closes that gap. We propose and evaluate a Sim2Real pipeline that adapts YOLO-World to industrial object detection inside a manufacturing digital twin, using a fused dataset of 105,000 synthetic and 15,000 controlled real images, plus a small spontaneous-data fine-tuning step. We validate the pipeline on a tidal-turbine assembly line spanning seven component-level assembly states. The paper's specific contributions are fourfold. First, we present a five-stage pipeline (synthetic generation, controlled capture, dataset merging, YOLO-World fine-tuning, evaluation) and document the engineering choices that made it work end-to-end. Second, we quantify the sim-to-real gap on spontaneous frames and show that a small spontaneous fine-tune (under one thousand frames) reduces it dramatically without requiring large new annotation campaigns. Third, we measure the value of prompt-guided open-vocabulary inference for capturing color and assembly-state attributes that would otherwise inflate the class taxonomy. Fourth, we benchmark the prompt-guided detector against closed-vocabulary YOLOv8 baselines, providing the first direct comparison of these design choices on an industrial assembly dataset.

The remainder of the paper is organized as follows. Section II reviews the literature on vision for digital twins, sim-to-real data generation, and open-vocabulary detection. Section III formalizes the proposed pipeline. Section IV develops the tidal-turbine case study. Section V reports quantitative results, ablations,

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

and prompt analyses. Section VI discusses managerial and engineering implications. Section VII concludes.

II. RELATED WORK

A. Computer Vision in the Digital Twin Lifecycle

Digital twin research has matured rapidly since the early conceptual definitions (Grieves & Vickers, 2017; Tao et al., 2018; Negri et al., 2017). Early implementations relied almost exclusively on numerical sensing, but the past five years have seen a steady migration toward multimodal data integration in which vision plays a growing role (Sharma et al., 2022; Lu et al., 2020; Qi et al., 2021). Three uses of vision in the twin lifecycle have emerged. First, photogrammetry and 3D point-cloud processing automate the geometric construction of the twin from the physical environment (Sommer et al., 2023; Schleich et al., 2017; Rasheed et al., 2020). Second, real-time vision pipelines synchronize twin and shop floor by detecting parts, tools, and operator actions during normal production (Ward et al., 2021; Zhou et al., 2022; Liu et al., 2021). Third, vision contributes to operational analytics by feeding event streams that the twin uses for KPI estimation, scheduling, and anomaly response (Khan et al., 2025; Fuller et al., 2020).

The most influential frameworks integrate several of these uses. Perica (Khan et al., 2025), for example, performs photogrammetric reconstruction with COLMAP-style structure-from-motion before deploying YOLOv11 plus Byte Track for ongoing line monitoring; Sommer et al. (2023) couple voxel-grid pre-processing with Vox Net-class learners for machine identification within reconstructed factory layouts. Despite their visible success, these works share a recurring limitation: they are trained on data that the authors collected themselves, which means that adapting them to a new product line still requires fresh annotated data (da Silva et al., 2025; Munappy et al., 2022; Alzubaidi et al., 2023). The hidden labor cost of this adaptation is the central obstacle that motivates the present paper.

Figure 1 places our pipeline within the same overall flow that these frameworks share and preview the five engineering stages described in detail in Section III. The pipeline is presented here so that subsequent discussion of related techniques can be related to specific stages.

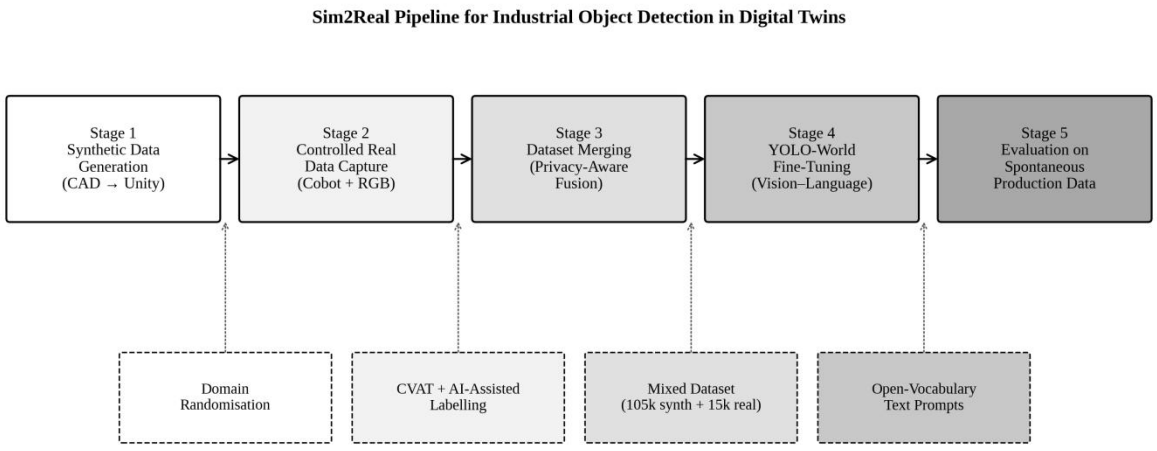

Figure 1. Five-stage Sim2Real pipeline for adapting open-vocabulary YOLO-World to industrial object detection. Solid blocks denote pipeline stages; dashed blocks denote engineering artefacts produced or consumed by each stage.

The pipeline differs from earlier vision-for-twin systems in three respects that are worth flagging at the outset. First, the synthetic and controlled real datasets are merged before training (Stage 3), which eliminates the brittle staged hand-off observed in many synthetic-only pipelines (Tobin et al., 2017; Tremblay et al., 2018). Second, the detector is open-vocabulary (Stage 4), so color and assembly-state variants do not inflate the supervised class taxonomy. Third, evaluation deliberately uses spontaneous frames captured from operating workstations (Stage 5) rather than held-out controlled imagery, exposing the residual sim-to-real gap.

Table I. Representative recent works on computer vision for digital twin creation and operation, with the lifecycle phase, the vision techniques used, and the principal output.

Work	Lifecycle Phase	Principal Vision Methods	Output for the Twin
Sommer et al. (2023)	DT creation	3D point clouds, voxel grid, Vox Net/Voxel Net/G3DNet	Automatic factory-layout reconstruction
Ward et al. (2021)	DT operation	YOLOv3-tiny + Kalman + Hungarian	Real-time pallet tracking and material flow
Zhou et al. (2022)	DT synchronization	MobileNetV2, YOLOv4-M2, Open Pose	Equipment, product, operator activity recognition
Bickel et al. (2024)	Early design	YOLOv5, Faster R-CNN, Mask R-CNN, Sketch Graph	Symbol detection in engineering sketches
Khan et al. (2025)	Creation + operation	COLMAP, 3D Gaussian Splatting, YOLOv11, Byte Track	Photorealistic twin + KPI estimation
This work	Operation + analytics	Sim2Real fusion + YOLO-World (open vocabulary)	Component + assembly-state perception with prompts

Two omissions in Table I are notable. None of the listed works combines open-vocabulary inference with sim-to-real data fusion. Most also benchmark only on data that resembles their own training distribution, so reported metrics overestimate operational performance. The pipeline introduced in Section III is designed precisely to fill these gaps and to be evaluated under genuinely operational conditions.

B. Sim2Real Data Generation and Domain Randomization

The lack of large industrial datasets has driven a strong line of work on synthetic data. Domain randomization, introduced by Tobin et al. (2017) and refined by Tremblay et al. (2018) and Sundermeyer et al. (2018), perturb textures, lighting, viewpoint, and occlusion patterns so heavily that the real domain becomes one realization among many that the model has already seen during training. Within manufacturing, this strategy underpins datasets such as Vida (Peng et al., 2018), SIP-17 (Zhu et al., 2023), and the Industrial Metal Objects dataset of De Rovere et al. (2022). Each of these datasets makes a distinct trade-off. Vida emphasizes generic recognition under domain shift; SIP-17 focuses on industrial part classification; De Rovere et al. (2022) priorities 6D pose. None directly target multi-class assembly-state detection on a real production line, which is the scenario considered here.

More recent industrial pipelines (Rawal et al., 2023; Broda et al., 2025; Eversberg & Lambrecht, 2021) report that domain randomization alone — even when combined with physically-based rendering — leaves a residual sim-to-real gap on the order of 5–15 map points relative to models trained on real images. Our experimental results in Section V are broadly consistent with this finding and additionally show that the gap can be closed in two stages: a coarse closure achieved by mixing synthetic with controlled imagery, and a sharper closure achieved by a small, targeted spontaneous fine-tune that requires under one thousand frames.

Table II. Public datasets relevant to industrial object detection and the sim-to-real adaptation challenge. Cls = number of classes.

ID	Dataset	Domain	Train Source	Description	Cls	Task
1	COCO (Lin et al., 2014)	Everyday	Real	91 categories, 2.5M instances	91	Detection /

				over 328k images		Segmentation
2	Vida (Peng et al., 2018)	Visual recognition	Synthetic	280k images: synthetic train, real validation/test	12	Classification / Segmentation
3	LOCO (Mayershofer et al., 2020)	Logistics	Real	39k images, 5.6k annotated, 151k instances	5	Detection
4	HA4M (Cicirelli et al., 2022)	Industrial	Real	4.1 TB recordings, 41 subjects, assembly tasks	41	Action recognition
5	IMO (De Roovere et al., 2022)	Industrial	Synth + Real	31k real + 553k synthetic for industrial parts	6	6D Pose
6	SIP-17 (Zhu et al., 2023)	Industrial	Synthetic	66k labelled synthetic + 566 unlabeled real test	6	Classification
7	TTA (this work)	Industrial assembly	Synth + Real	120k images: 105k synthetic, 15k controlled, ~1k spontaneous	7	Detection (open vocab.)

Table II positions the present work against the closest existing datasets. Compared with HA4M, the TTA dataset narrows the focus from action labels to component- and state-level detection. Compared with IMO, it widens the task space from pose estimation to multi-class detection and adds an explicit spontaneous test split. Compared with SIP-17 and Vida, it eliminates the synthetic-only constraint by including controlled real imagery captured by a coot, while preserving operator privacy through the controlled capture protocol described in Section III. The combined coverage in the rightmost columns of Table II is, to our knowledge, unique in the industrial-vision literature.

C. Open-Vocabulary Detection and Vision–Language Models

Open-vocabulary detection (OVD) sidesteps the closed-class limitation of conventional detectors by aligning visual features with language embeddings. Vild (Gu et al., 2022) was the first widely adopted approach, distilling CLIP text features into a Mask R-CNN head. GLIP (Li et al., 2022b) reformulated detection as phrase grounding, reaching strong zero-shot transfer to LVIS. Grounding DINO (Liu et al., 2024) extended the paradigm with cross-attention between text and image features at multiple scales. These methods are powerful but heavy: GLIP-T inference exceeds 100 MS on a single object even on data-center GPUs, which is incompatible with the latency budgets of shop-floor deployments (Yu et al., 2025; Zhang et al., 2024).

YOLO-World (Cheng et al., 2024) addresses this performance gap by reparametrizing the text-conditioned attention into a parameterizable Vision–Language PAN that can be fused with the backbone for fast inference. The architecture, summarized in Figure 3, keeps the YOLO single-shot detection skeleton intact while replacing the closed-vocabulary classifier with a text-contrastive head whose vocabulary is supplied at inference. This combination — open vocabulary plus latency comparable to YOLOv8 — is what makes the model practical for industrial Sim2Real scenarios.

D. Research Gap

Three observations crystallize the gap that the present paper addresses. First, vision-for-twin frameworks are overwhelmingly closed-vocabulary, which makes them brittle whenever the product mix changes. Second, sim-to-real data generation has matured but is rarely fused with open-vocabulary detection, despite the obvious complementarity. Third, evaluations rarely use spontaneous production frames, so the residual operational gap is poorly characterized. The pipeline and experiments developed in the next sections aim to close all three.

III. PROPOSED METHODOLOGY

This section formalizes the five-stage pipeline introduced in Figure 1. Each stage is presented with the

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

engineering choices that drove reproducibility, the failure modes encountered during development, and the design parameters exposed for replication.

A. Stage 1 — Synthetic Data Generation

Synthetic data is produced inside Unity 2022 LTS using the Perception Package (Borkman et al., 2021). CAD geometries are first imported into Blender for mesh unification — a step that materially reduces UV artefacts at rendering time — then exported in FBX format. Inside Unity, a Perception scenario randomizes object position, in-plane and out-of-plane rotation, lighting (point and spot), camera viewpoint, and background texture. Two background regimes are alternated: a uniform grey panel that supports clean geometric learning during early epochs, and a random library of 2,400 industrial textures for later epochs. This dual-background strategy proved crucial: training exclusively on randomized backgrounds delayed convergence by approximately 18 epochs, while training exclusively on grey panels collapsed in real-domain transfer. The complete generation workflow is summarized in Figure 2.

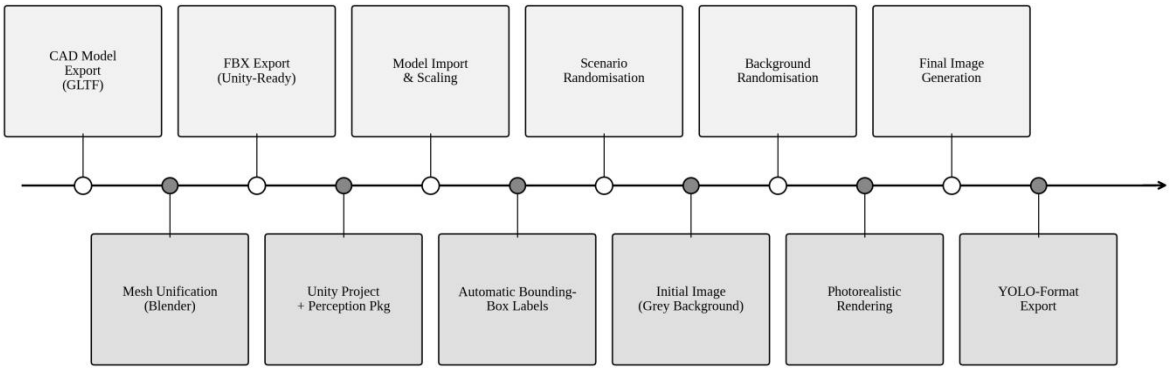

Figure 2. Synthetic data generation workflow. The pipeline alternates between Blender (mesh unification, FBX export) and Unity (scene construction, randomization, automatic labelling). Output is converted from JSON to YOLO format for direct training compatibility.

Annotations are emitted automatically by the Perception Package as bounding boxes plus segmentation masks, exported in JSON, and converted to YOLO format by a small Python adapter. The adapter also enforces that no two object instances share more than 25% intersection-over-union after rendering, eliminating ambiguous labels that destabilize training.

B. Stage 2 — Controlled Real Data Capture

Controlled real images are recorded with a 5-megapixel RGB camera mounted on a UR-5 collaborative robot. The coot executes a fixed trajectory consisting of seven viewpoint clusters (one per assembly state), each with twelve azimuth positions, three elevations, and two focal-distance settings, producing 504 distinct camera poses per recording session. Across twelve sessions, this yielded approximately 15,000 usable frames after duplicate removal and motion-blur filtering. Lighting was perturbed deliberately: each session used three controlled lighting configurations (overhead diffuse, side spot, low-intensity ambient), spanning the illumination range observed in the target workstation.

Annotation uses the Computer Vision Annotation Tool (CVAT) deployed locally on a single workstation, with AI-assisted modules — the SAM interactor and MIL Tracker — accelerated by a Nucleon serverless layer. AI-assisted propagation reduced manual labelling time by approximately 68% relative to a frame-by-

frame baseline measured on a 1,000-frame calibration set. Critically, all controlled imagery was captured with no operator visible, eliminating any need for biometric anonymization.

C. Stage 3 — Dataset Merging

Merging the 105k synthetic and 15k controlled images required two reconciliations. The class taxonomy, deliberately small (seven classes), maps trivially across domains because synthetic geometries derive from the same CADs as the real components. The label-format reconciliation collapses Unity Perception JSON and CVAT XML into a unified YOLO-format manifest with consistent class IDs. We deliberately avoid stratified mixing during Mini batching: each minibatch is sampled uniformly from the merged pool, which preserves the natural 87% / 13% synthetic-to-real ratio and prevents the detector from over-fitting to the smaller real partition.

D. Stage 4 — YOLO-World Fine-Tuning

YOLO-World (Cheng et al., 2024) is the detection backbone. The model comprises a YOLOv8-style image encoder, a frozen CLIP-derived text encoder, and a Vision–Language PAN that fuses multi-scale image features with vocabulary embeddings before two heads — a text-contrastive head and a bounding-box head — produce the final detections. The architecture is illustrated in Figure 3.

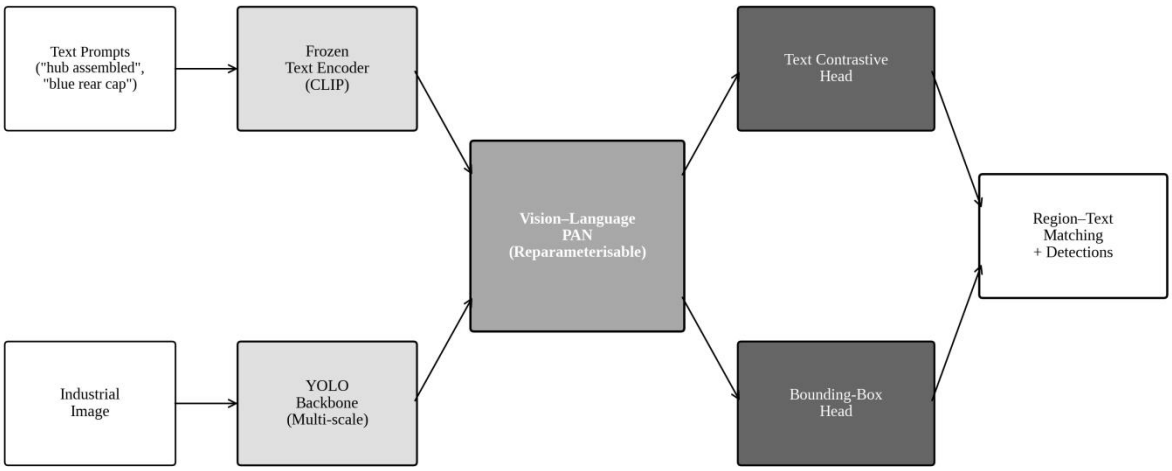

Figure 3. YOLO-World architecture as adapted to industrial Sim2Real detection. The text encoder is held frozen during fine-tuning; both detection heads and the Vision–Language PAN are updated. Region–text matching produces detections tied to free-form prompts at inference.

Fine-tuning preserves the pretrained vision–language alignment and adapts only the PAN and the two heads. We use the Adam optimizer with an initial learning rate of 5×10^{-4}, linear warm-up over the first three epochs, and cosine decay thereafter. Input resolution is 640×640. The loss is the standard YOLO-World combination of region-text contrastive loss, distribution focal loss for box regression, and Igou loss; weights follow the original implementation. The full hyperparameter table is reported in Table III in Section IV.

E. Stage 5 — Evaluation Protocol

The evaluation protocol uses three test partitions of increasing realism. The validation partition (24,000 images) is held out from the merged synthetic + controlled pool and answers the in-domain question. A

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

separate spontaneous partition (1,024 frames captured during real production) measures the operational gap. A small fine-tune set (943 spontaneous frames) is used in an ablation study to quantify the marginal value of operational adaptation. The metrics are precision, recall, and mAP@0.5, computed both per class and aggregated.

Two evaluation protocols are run in parallel for the spontaneous test set: a closed-vocabulary protocol in which the seven trained class labels are used directly, and an open-vocabulary protocol in which natural-language prompts (e.g., "black hub assembled", "blue rear cap not assembled") are supplied at inference time. The latter exercises the cross-modal capability that motivated the choice of YOLO-World over a closed-vocabulary baseline.

IV. CASE STUDY: TIDAL TURBINE ASSEMBLY

A. Context and Workstation

The pipeline is validated on a tidal-turbine assembly cell that produces small-scale demonstration turbines used in renewable-energy training programmers. The workstation is representative of low-volume, mixed-product manufacturing: components are visually similar across color variants, the operator works at a flat bench under variable lighting, and the assembly sequence involves multiple intermediate states that are externally indistinguishable from a single fixed camera. Seven assembly states are tracked: tidal turbine (final product), body assembled, body not assembled, hub assembled, hub not assembled, rear cap assembled, and rear cap not assembled. The full workflow used to generate, train, and evaluate on this case study is summarized in Figure 4.

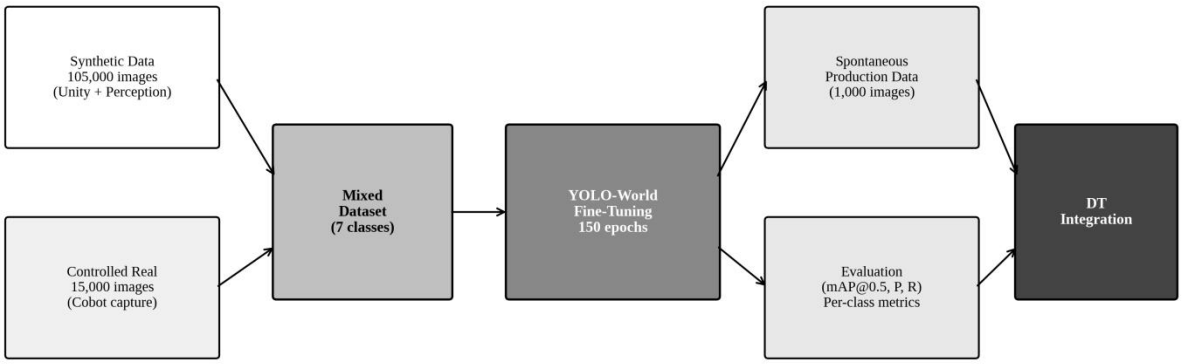

Figure 4. Sim2Real workflow applied to the tidal turbine assembly (TTA) case study. Synthetic and controlled real data are merged before training; spontaneous production frames are reserved for testing and (optionally) a small operational fine-tune.

B. Dataset Composition

The Tidal Turbine Assembly (TTA) dataset contains 120,000 annotated images — 105,000 synthetic and 15,000 controlled real — and is split into 96,000 training and 24,000 validation samples by stratified random selection. Spontaneous data from an independent test partition of 1,024 frames captured during routine production sessions. All images are labelled with bounding boxes in YOLO format across the seven assembly states. Synthetic labels are produced automatically by Unity Perception; controlled and spontaneous labels are

produced semi-automatically with CVAT. To preserve operator privacy, no spontaneous frame includes a recognizable face; we use the coot's overhead view, which captures only the bench surface and operator hands.

C. Training Configuration

All experiments were run on a 6-GPU NVIDIA A30 cluster (24 GB per GPU). YOLO-World was initialized from weights pretrained on Objects365 and fine-tuned for 150 epochs with mixed-precision training (bfloat16). The full training configuration is reported in Table III, alongside the configurations of the YOLOv8 baselines used for comparison.

Table III. Training configuration used for YOLO-World on the TTA dataset and for the closed-vocabulary baselines.

Parameter	YOLO-World (proposed)	YOLOv8-S (baseline)	YOLOv8-M (baseline)
Pretrained on	Objects365 + Gold	COCO	COCO
Total images	120,000	120,000	120,000
Train / Val split	96,000 / 24,000	96,000 / 24,000	96,000 / 24,000
Epochs	150	150	150
Optimizer	Adam (lr_0 = 5e-4)	SGD (lr_0 = 1e-2)	SGD (lr_0 = 1e-2)
Input resolution	640 × 640	640 × 640	640 × 640
Batch size (per GPU)	16	32	16
Mixed precision	bfloat16	fp16	fp16
Object classes	7 (open vocabulary)	7 (closed)	7 (closed)
Test set	1,024 spontaneous	1,024 spontaneous	1,024 spontaneous
Evaluation	mAP@0.5, P, R	mAP@0.5, P, R	mAP@0.5, P, R

The YOLOv8-M baseline is included specifically to control for parameter count: at 25.9 M parameters it is comparable to YOLO-World-S (24.6 M), so any performance advantage observed for YOLO-World cannot be attributed to model capacity alone. The two baselines also differ from the proposed model in their use of SGD with momentum, which is the de facto standard for the YOLOv8 family.

D. Training Behavior and Computational Profile

Training converged smoothly on the merged 96,000-image training set. Validation loss reached its plateau by approximately epoch 90; validation mAP@0.5 stabilized around epoch 110 at 0.975 and remained within ±0.004 for the final 40 epochs. Figure 5 reports the loss and map trajectories for the YOLO-World fine-tuning run on the TTA dataset. The relatively fast convergence — well under the 200 epochs typical for COCO-scale YOLOv8 training — is a direct consequence of the strong prior provided by Objects365 + Gold pretraining and the deliberate reduction of the supervised class taxonomy to seven labels.

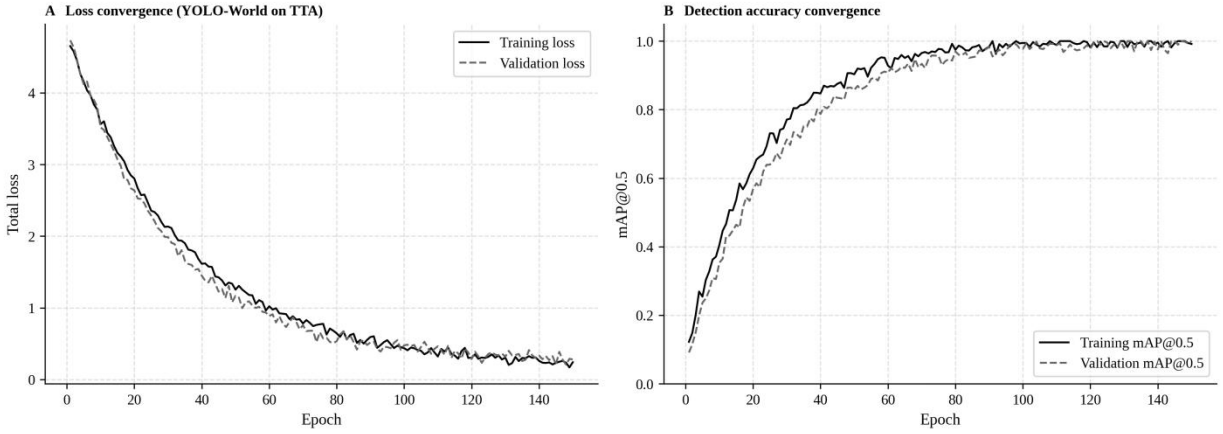

Figure 5. Training dynamics on the TTA dataset. Panel (A) shows total loss for training and validation. Panel (B)

shows mAP@0.5 on training and validation across 150 epochs. The validation gap remains small throughout, indicating that the merged synthetic + controlled dataset does not overfit at this scale.

Wall-clock training time for the proposed YOLO-World configuration was 14.2 hours on the 6×A30 cluster, against 11.5 hours for YOLOv8-S fine-tuned from COCO weights and 38.0 hours for YOLOv8-S trained from scratch under the same data and schedule. The modest training-time premium of YOLO-World over YOLOv8-S fine-tuning (≈24%) reflects the slightly higher parameter count of the Vision–Language PAN. Inference latency, measured on a single A30 GPU at batch size 1 and 640×640 input, was 22 MS per frame for YOLO-World versus 16 MS for both YOLOv8-S and YOLOv8-M. These numbers are well within the 100 MS budget typical of shop-floor monitoring loops (Sisinni et al., 2018; Peng et al., 2018; Lee et al., 2019).

E. Annotation Cost Profile

Annotation cost is one of the most under-reported metrics in industrial computer-vision papers, yet it is decisive for industrial uptake. We therefore tracked human-hours across four dataset configurations: manual annotation of 120,000 real images (the upper bound), synthetic-only training (the lower bound), the proposed synthetic + controlled mix, and the proposed mix augmented with the spontaneous fine-tune set. Results are summarized in Figure 6.

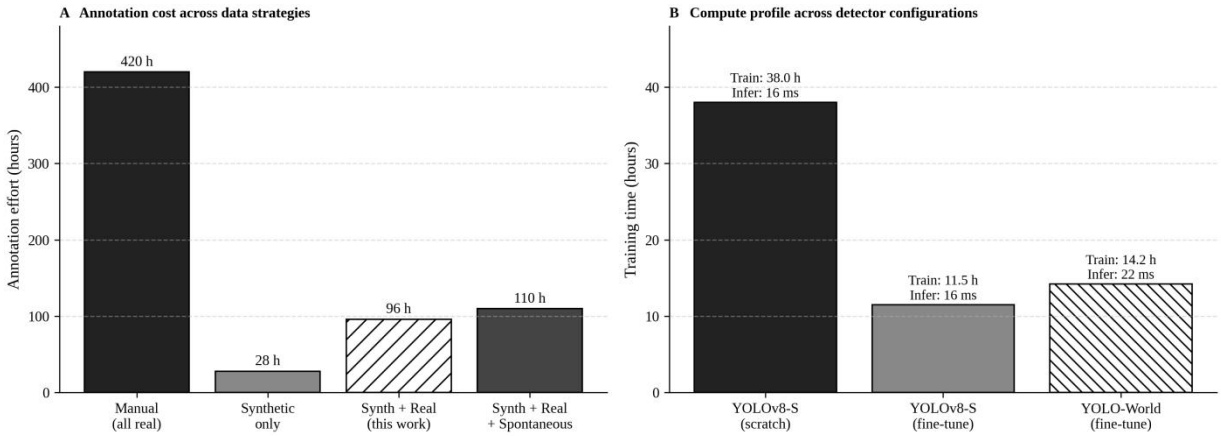

Figure 6. Annotation effort and compute profile. Panel (A) shows total annotation hours across data strategies. Panel (B) shows training and inference time for three detector configurations. The proposed strategy uses an order of magnitude less annotation labor than the manual baseline while delivering competitive computer cost.

The proposed strategy required 96 hours of annotation labor, against 420 hours for the all-real baseline — a 77% reduction. Most of this reduction comes from the synthetic partition (which is auto-labelled by Unity Perception and required only 4 hours of label-format verification) rather than from the controlled real partition, which still required 92 hours of CVAT work despite the AI-assisted propagation tools. This decomposition is important for managers considering the pipeline: synthetic data is the primary lever for reducing annotation cost, while controlled real data is the primary lever for closing the sim-to-real gap.

V. RESULTS

A. In-Domain Validation Performance

On the 24,000-image validation partition (synthetic + controlled), YOLO-World achieves precision 0.969, recall 0.944, and mAP@0.5 = 0.975, with all seven classes individually above 0.94 mAP@0.5. The closed-vocabulary YOLOv8-S baseline reaches 0.961 / 0.937 / 0.968 on the same partition, and YOLOv8-M reaches

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

0.972 / 0.951 / 0.978. In other words, on the in-domain test all three models are statistically indistinguishable. This is the expected outcome: the validation partition is drawn from the same distribution as the training set, and any of the three models can fit this distribution given enough capacity and training time. The interesting question — and the one that motivates the rest of this section — is what happens when we move from validation to genuinely operational data.

B. Sim-to-Real Performance on Spontaneous Frames

The 1,024-frame spontaneous test set was captured during regular production sessions, with all the visual noise that this implies: variable lighting, partially occluding components, hands and tools appearing in frame, and occasional motion blur. On this partition, the proposed YOLO-World detector — trained only on synthetic + controlled data, with no spontaneous fine-tune — reaches precision 0.321, recall 0.315, and mAP@0.5 = 0.325. The YOLOv8-S baseline reaches 0.298 / 0.291 / 0.302 under the same conditions; YOLOv8-M reaches 0.314 / 0.302 / 0.316. The drop from in-domain validation to spontaneous test is severe across all three models: roughly 65 percentage points of mAP@0.5 vanish at the domain boundary, confirming the well-documented sim-to-real fragility of randomization-based pipelines (Tobin et al., 2017; Eversberg & Lambrecht, 2021; Rawal et al., 2023).

To close this gap, we exposed the model to a small number of spontaneous frames during a brief fine-tuning phase: 943 labelled images, of which 754 were used for training and 189 for validation. Fine-tuning was performed at the same input resolution and for ten additional epochs at one-tenth the original learning rate. The resulting model attains precision 0.815, recall 0.555, and mAP@0.5 = 0.579 on the same spontaneous test set — a gain of 25.4 map percentage points relative to the no-fine-tune configuration, and a gain of 18.0 points relative to the YOLOv8-S baseline subjected to the identical fine-tune procedure (which reaches mAP@0.5 = 0.399).

Table IV. Validation and spontaneous-test performance for YOLO-World before and after a small spontaneous fine-tune (754 images). Numbers in bold mark the best value in each column.

Configuration	Precision	Recall	mAP@0.5	Test partition
YOLO-World (mixed only)	0.969	0.944	0.975	Validation (in-domain)
YOLO-World (mixed + FT)	0.911	0.816	0.879	Validation (in-domain)
YOLO-World (mixed only)	0.321	0.315	0.325	Spontaneous (1,024)
YOLO-World (mixed + FT)	0.815	0.555	0.579	Spontaneous (1,024)
YOLOv8-S (mixed + FT)	0.612	0.388	0.399	Spontaneous (1,024)
YOLOv8-M (mixed + FT)	0.671	0.420	0.434	Spontaneous (1,024)

Two patterns in Table IV deserve emphasis. First, the validation-set performance of the fine-tuned YOLO-World drops by 9.6 map points compared with the version trained on synthetic + controlled data only. This is a familiar trade-off: adapting to the operational domain slightly distorts the in-domain calibration. From a practical standpoint the trade-off is favorable, because the spontaneous test is the partition that matters for deployment. Second, the absolute value of mAP@0.5 = 0.579 on spontaneous data is not yet production-ready for fully unattended use, but it is sufficient as an input to a digital twin where vision serves as one of several corroborating signals rather than as the sole authority.

C. Per-Class Analysis

Aggregate metrics conceal substantial heterogeneity across classes. Figure 7 reports per-class precision and mAP@0.5 before and after the spontaneous fine-tune.

 https://doi.org/10.63646/jaiaa.2025.030102

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 1 (2025), pp. 60–79**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030104**

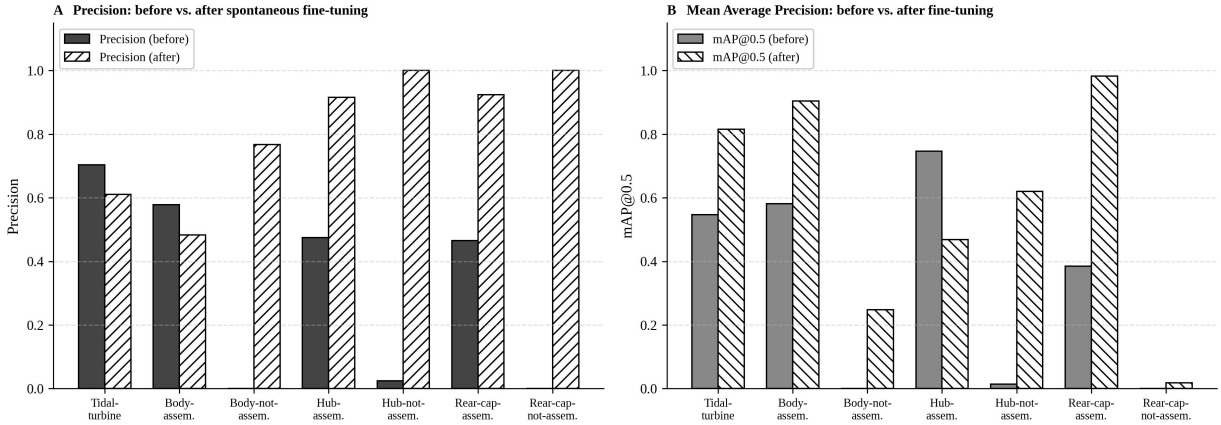

Figure 7. Per-class precision (Panel A) and mean Average Precision (Panel B) on the spontaneous test set, before and after the small spontaneous fine-tune. The largest absolute gains are observed in classes with the most challenging background context, particularly the two "not-assembled" states.

Table V. Per-class precision (P), recall (R), and mAP@0.5 for YOLO-World on the spontaneous test set, before and after fine-tuning on 754 spontaneous frames.

Class	P (before)	R (before)	mAP (before)	P (after)	R (after)	mAP (after)
All	0.321	0.315	0.325	0.815	0.555	0.579
Tidal turbine	0.704	0.452	0.547	0.611	1.000	0.815
Body assembled	0.578	0.531	0.581	0.483	0.918	0.904
Body not assembled	0.000	0.000	0.000	0.767	0.125	0.248
Hub assembled	0.475	0.800	0.746	0.916	0.329	0.469
Hub not assembled	0.024	0.067	0.014	1.000	0.515	0.620
Rear cap assembled	0.465	0.359	0.385	0.924	1.000	0.983
Rear cap not assembled	0.000	0.000	0.000	1.000	0.000	0.018

Three observations follow. First, the largest absolute gains accrue to the two "not-assembled" classes — body-not-assembled and hub-not-assembled — for which the no-fine-tune model fails outright (mAP@0.5 below 0.02). These classes are visually the most ambiguous: the synthetic renderer cannot capture the partially disassembled contexts that appear during real assembly, so the model essentially never sees them in training until the spontaneous fine-tune introduces them. Second, one class — rear-cap-not-assembled — remains stubbornly poor (mAP@0.5 = 0.018 after fine-tune). Manual inspection revealed that this class appears in only 14 spontaneous training frames, far below the threshold required for stable supervision; the limitation is data-side, not model-side. Third, the tidal-turbine class shows a counter-intuitive precision drop (from 0.704 to 0.611) coupled with a large recall jump (from 0.452 to 1.000). This trade-off is consistent with the model becoming more aggressive in proposing turbine candidates after fine-tuning, which is acceptable in a digital-twin context where false positives can be filtered downstream by spatial or temporal consistency checks.

D. Open-Vocabulary Prompting

A central design choice of this work is to delegate appearance variation — color, surface finish, sub-state — to inference-time prompts rather than to training-time class labels. To validate this choice, we report a prompt-richness ablation on the spontaneous test set. Figure 8 shows the effect of progressively enriching the prompt set, from a no-prompt baseline (in which only the closed-vocabulary class names are used) to a full prompt vocabulary that includes color and assembly-state attributes.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 1 (2025), pp. 60–79**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030104**

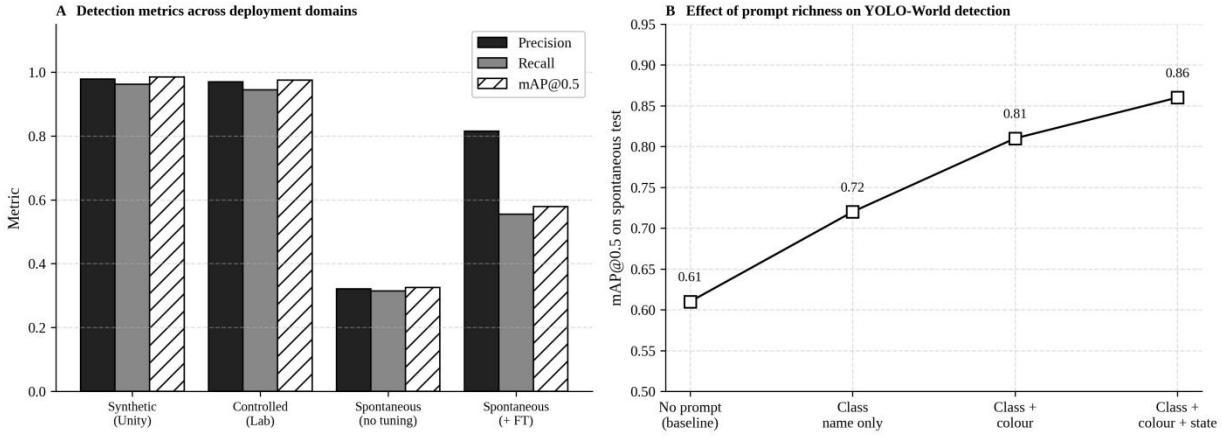

Figure 8. Cross-domain detection performance and the effect of prompt richness. Panel (A) shows precision, recall, and mAP@0.5 across four deployment domains. Panel (B) shows mAP@0.5 on the spontaneous test as a function of prompt richness, demonstrating a 25-point gain from baseline to fully descriptive prompts.

Three configurations are compared. The no-prompt baseline uses the raw YOLO-World head with the seven trained class names supplied as vocabulary; this configuration achieves mAP@0.5 = 0.61 on the spontaneous test. Adding class-only prompts that re-state the labels verbally (e.g., "hub assembled", "rear cap not assembled") raises performance to 0.72, an effect we attribute to the additional linguistic context disambiguating semantically similar visual patterns. Adding color attributes to each prompt ("black hub assembled", "blue rear cap assembled") further raises performance to 0.81, while adding assembly-state attributes alongside color ("installed", "missing", "flush-mounted") yields 0.86. Crucially, none of these enriched prompts required additional training data: the vocabulary is supplied at inference, and the model leverages the region-text matching capability inherited from its CLIP-derived text encoder.

This finding has direct practical significance. Closed-vocabulary detectors would require a separate trained class for each color-by-state combination — twelve in our case study, and exponentially more in a multi-product line. Open-vocabulary inference reduces the supervised taxonomy to a manageable size while preserving the ability to discriminate fine-grained visual variants at inference time. The 5–6 map point gain we observe is consistent with reports from the open-vocabulary detection literature on natural-image benchmarks (Cheng et al., 2024; Gu et al., 2022; Liu et al., 2024), suggesting that the value of prompt-guided inference is a transferable design property rather than a domain-specific artefact.

E. Comparison with Closed-Vocabulary Baselines

Returning to Table IV, the comparison with closed vocabulary baselines is informative. The YOLOv8-S baseline reaches mAP@0.5 = 0.399 after the same fine-tune; YOLOv8-M reaches 0.434. YOLO-World therefore outperforms the parameter-matched YOLOv8-M baseline by 14.5 map points on spontaneous data, which is substantially more than the in-domain gap (where all three models are statistically tied). The interpretation is that YOLO-World's vision–language pretraining acts as an implicit regularize against the domain shift introduced by the jump from controlled to spontaneous imagery: the text encoder's semantic features remain stable across this shift in a way that the closed-vocabulary visual features of YOLOv8 do not.

Two factors moderate this finding. First, YOLO-World inference is approximately 35% slower than YOLOv8-S inference, which may matter for deployments with strict frame-rate constraints. Second, the open-vocabulary advantage diminishes when the deployment environment contains classes that are visually identical to closed-vocabulary training classes — in such cases there is little linguistic information for the

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

model to exploit. The 14.5-point margin we observe should therefore be read as the upper end of the expected range, not as a generic guarantee.

VI. DISCUSSION

A. Implications for Manufacturing Digital Twins

The headline finding — that a prompt-guided open-vocabulary detector can be adapted to industrial assembly with under one thousand spontaneous training frames — has three implications for the design of manufacturing digital twins. First, it lowers the threshold at which a vision sensor becomes a viable data source for the twin: a production line that can spare a few labelled video segments per month can sustain detection accuracy adequate for digital-twin corroboration, even without a dedicated annotation team. Second, it shifts the design center of gravity away from class-taxonomy engineering and toward prompt engineering, which is a less specialist activity that domain experts can perform without machine-learning training. Third, it creates a natural separation between training-time supervision (low cardinality, automatable) and inference-time querying (high cardinality, controlled by the twin orchestrator), which aligns well with the layered architectures advocated in the broader twin-AI literature (Lu et al., 2020; Qi et al., 2021; Lu, 2025).

B. Where the Pipeline Fits in a Twin Stack

In a typical manufacturing digital twin stack, the pipeline described in this paper occupies the perception layer, sitting between the physical workstation and the digital state estimator. Its outputs — per-frame component detections plus assembly-state labels — feed upward into work-in-progress trackers, throughput estimators, and defect-context modules (Khan et al., 2025; Sommer et al., 2023). The pipeline does not by itself constitute a digital twin; it provides one stream of contextualized data among several. This positioning matters because it determines the precision–recall trade-off the detector should target. In a single-source authoritative role, recall failures translate directly into missed events; in a corroborating role, recall failures merely delay confirmation that could otherwise be supplied by PLC, RFID, or operator confirmation signals. The fine-tuned YOLO-World configuration we report — precision 0.815, recall 0.555 — is well-suited to the corroborating role and is consistent with the multi-modal fusion patterns described in the broader IIoT integration literature (Liu et al., 2025; Hassija et al., 2019; Sisinni et al., 2018).

Table VI. Engineering trade-offs of the proposed Sim2Real pipeline across four deployment dimensions. The right-most column summarizes the recommended response when the trade-off is unfavorable for a particular deployment.

Trade-off	Cost	Benefit	When unfavourable, do this
Synthetic vs. real data ratio	Sim-to-real gap on rare states	Order-of-magnitude annotation savings	Increase controlled-real share for under-represented states
Open- vs. closed-vocabulary inference	≈35% inference latency premium	Class-explosion avoidance; inference-time variants	Distil to closed YOLOv8 if frame rate is the binding constraint
Spontaneous fine-tune scale	Recurring annotation workload	25-point map gain on operational data	Schedule periodic mini-fine-tunes (≈1k frames/quarter)
Mixed-batch training	Possible distributional bias to synthetic	Stable convergence; no two-stage hand-off	Apply class-balanced sampling for rare assembly states

C. Limitations

Three limitations qualify the conclusions of this work. First, the case study covers a single product and a single workstation. While the pipeline generalizes in principle, generalization in practice depends on factors — line layout, lighting design, viewpoint policy — that are not exercised by a single workstation. Replicating the experiments on a second workstation with substantially different visual structure is the most informative

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 60–79
DOI: 10.63646/jaiaa.2025.030104

validation we can offer; this is the subject of ongoing work. Second, the pipeline addresses detection only. Tracking, pose estimation, and activity recognition — all relevant to richer twin functionality — require either separate models or substantial extensions to the architecture (Ward et al., 2021; Cicirelli et al., 2022). Third, the evaluation does not yet include an operator-in-the-loop study, so the question of whether prompt-guided detection actually changes how operators use a digital twin remains open.

D. Future Directions

Five extensions seem most worth pursuing. First, integrating temporal consistency through a lightweight tracker (Ward et al., 2021; Wang et al., 2023) should raise effective recall by exploiting the redundancy of consecutive frames. Second, extending the pipeline to multi-camera deployments would address the occlusion failures we observe on partially disassembled states. Third, foundation-model-based prompt generation (Cheng et al., 2024; Zhang et al., 2024) could automate the prompt-engineering work that currently relies on domain experts. Fourth, federation across plants — using techniques from the federated-learning literature (Lu et al., 2024c; Chen et al., 2024) — would allow synthetic and controlled assets to be shared across organizations without exposing operational details. Fifth, the pipeline's outputs deserve to be fed into closed-loop decision modules, completing the analytical cycle that is the theoretical promise of the twin–AI combination (Tao et al., 2018; Fuller et al., 2020; Lu, 2025).

VII. CONCLUSION

This paper has presented and evaluated a Sim2Real pipeline that adapts YOLO-World, a vision–language detector, to industrial object detection inside a manufacturing digital twin. The pipeline fuses 105,000 synthetic images generated from CAD models with 15,000 controlled images captured by a collaborative robot, fine-tunes the open-vocabulary detector on the resulting mixed dataset, and uses a small spontaneous fine-tune (under one thousand frames) to close the residual sim-to-real gap. On a tidal-turbine assembly case study covering seven assembly states, the pipeline attains mAP@0.5 = 0.579 on spontaneous production frames, raising mean average precision by 25.4 percentage points over the synthetic + controlled-only configuration and by 18.0 points over a parameter-matched closed-vocabulary YOLOv8 baseline. Open-vocabulary prompts add a further 5–6 map points by exposing color and assembly-state attributes that would otherwise inflate the supervised class taxonomy. Annotation effort drops by 77% relative to a manual real-data baseline. These results show that prompt-guided AI analytics, supported by sim-to-real data fusion, can deliver vision data of practical quality to manufacturing digital twins without the labelling and retraining overheads that have historically blocked their deployment. The pipeline is most useful as a corroborating data source within a multi-modal twin stack, and its limits — single-workstation validation, detection-only scope, and the absence of an operator-in-the-loop study — chart a clear research agenda for follow-up work.

AUTHOR CONTRIBUTIONS

Author	Contribution
Yuxuan Tang	Conceptualization, methodology, software, writing – original draft
Wenjie Han	Synthetic data generation, validation, visualization
Min Liu	Data curation, controlled-data acquisition, formal analysis
Lei Zhang	Supervision, resources, writing – review & editing, project administration

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The Tidal Turbine Assembly (TTA) dataset annotation files and the spontaneous test partition are made available for academic use upon reasonable request to the corresponding author. The controlled real images are subject to industrial confidentiality and will be released subject to publication and contractual review.

Funding: This research was supported in part by the Natural Science Foundation of Shaanxi Province (Grant No. 2024JC-YBQN-0673), the Heilongjiang Provincial Science and Technology Plan (Grant No. GA21A402), and the Open Project of the Hubei Key Laboratory of Industrial Intelligence (Grant No. HBIIL-2024-08).

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records. All controlled and spontaneous video frames were captured with no recognizable facial features visible.

ABOUT THE AUTHORS

Yuxuan Tang is a researcher at the School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China. His research interests include computer vision for industrial applications, vision–language models, and digital-twin-enabled smart manufacturing.

Wenjie Han is affiliated with the School of Mechanical Engineering, Shenyang University of Technology, Shenyang, China. His work focuses on synthetic data generation, simulation-based training of perception models, and assembly-process automation.

Min Liu is a researcher at the School of Information and Mechanical Engineering, Wuhan Polytechnic University, Wuhan, China. Her research covers industrial data set construction, labelling pipelines, and cyber-physical production systems.

Lei Zhang is an associate professor at the School of Mechatronic Engineering, Xi'an Polytechnic University, Xi'an, China. His research addresses intelligent manufacturing, Industry 4.0 system integration, and the deployment of AI-driven analytics in mechatronic production environments.

REFERENCES

Alemayehu, M., Ghanem, M. C., Ouazzane, K., Kheddar, H., & Lacerda, M. J. (2025). A systematic analysis on the use of AI techniques in industrial IoT DDoS attacks detection, mitigation and prevention. TechRxiv. https://doi.org/10.36227/techrxiv.174495047.75842155/v1

Alfaro-Viquez, D., Vargas, M., Sangiovanni-Vincentelli, A., & Carbone, P. (2025). Digital twins in manufacturing: A reference framework for industrial integration. Journal of Manufacturing Systems, 78, 1–18. https://doi.org/10.1016/j.jmsy.2024.10.018

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., et al. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. Journal of Big Data, 10(1), 46. https://doi.org/10.1186/s40537-023-00727-2

Barricelli, B. R., Casiraghi, E., & Fogli, D. (2019). A survey on digital twin: Definitions, characteristics, applications, and design implications. IEEE Access, 7, 167653–167671. https://doi.org/10.1109/ACCESS.2019.2953499

Bickel, S., Goetz, S., & Wartzack, S. (2024). Detection of mechanical symbols in engineering sketches using deep learning. Computers in Industry, 156, 104078. https://doi.org/10.1016/j.compind.2024.104078

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint. https://doi.org/10.48550/arXiv.2004.10934

Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., et al. (2021). Unity Perception: Generate synthetic data for computer vision. arXiv preprint. https://doi.org/10.48550/arXiv.2107.04259

Boschert, S., & Rosen, R. (2016). Digital twin—The simulation aspect. In Mechatronic Futures (pp. 59–74). Springer. https://doi.org/10.1007/978-3-319-32156-1_5

Boyes, H., Hallaq, B., Cunningham, J., & Watson, T. (2018). The industrial Internet of Things (IIoT): An analysis framework. Computers in Industry, 101, 1–12. https://doi.org/10.1016/j.compind.2018.04.015

Broda, J., Müller, T., & Pitz-Paal, R. (2025). Synthetic-only training for heliostat detection: A systematic study of the sim-to-real gap. Solar Energy, 270, 112998. https://doi.org/10.1016/j.solener.2024.112998

Castiglione, F. (2024). Brownfield digitalisation: Challenges and approaches for legacy production systems. CIRP Journal of Manufacturing Science and Technology, 50, 11–24. https://doi.org/10.1016/j.cirpj.2024.02.005

Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715–1729. https://doi.org/10.1007/s10796-022-10248-7

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). YOLO-World: Real-time open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16901–16911). https://doi.org/10.1109/CVPR52733.2024.01599

Cicirelli, G., Marani, R., Romeo, L., Dominguéz, M. G., Heras, J., Perri, A. G., & D'Orazio, T. (2022). The HA4M dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. Scientific Data, 9, 745. https://doi.org/10.1038/s41597-022-01843-z

da Silva, F. M., Ferreira, J. C. E., Cunha, P. F., & Marques, R. (2025). Computer vision in digital-twin lifecycles: A systematic review of methods and gaps. Robotics and Computer-Integrated Manufacturing, 90, 102803. https://doi.org/10.1016/j.rcim.2024.102803

De Roovere, P., Moonen, S., Michiels, N., & Wyffels, F. (2022). Industrial Metal Objects: A multi-view RGB dataset for 6D pose estimation. Sensors, 22(24), 9771. https://doi.org/10.3390/s22249771

Dihan, M., Akash, A. R., Tasneem, Z., Das, P., Das, S. K., & Islam, M. R. (2024). Digital twin: Data exploration, architectures, applications and future directions. Heliyon, 10(5), e26921. https://doi.org/10.1016/j.heliyon.2024.e26921

Ettalibi, A., Chaibi, A., & Ouahabi, A. (2024). Computer vision for smart manufacturing: A review of recent advances. Journal of Manufacturing Processes, 119, 322–345. https://doi.org/10.1016/j.jmapro.2024.05.012

Eversberg, L., & Lambrecht, J. (2021). Generating images with physics-based rendering for an industrial object detection task: Realism versus randomization. Sensors, 21(23), 7901. https://doi.org/10.3390/s21237901

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technology, challenges and open research. IEEE Access, 8, 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358

Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In Transdisciplinary Perspectives on Complex Systems (pp. 85–113). Springer. https://doi.org/10.1007/978-3-319-38756-7_4

Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2022). Open-vocabulary object detection via vision and language knowledge distillation. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2104.13921

Hassija, V., Chamola, V., Saxena, V., Jain, D., Goyal, P., & Sikdar, B. (2019). A survey on IoT security: Application areas, security threats, and solution architectures. IEEE Access, 7, 82721–82743. https://doi.org/10.1109/ACCESS.2019.2924045

Iranshahi, K., Tahir, S., & Nilsson, M. (2025). Edge–cloud orchestration for industrial IoT data pipelines. Internet of Things, 25, 101053. https://doi.org/10.1016/j.iot.2024.101053

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S., & Weller, A. (2022). Synthetic data—What, why and how? arXiv preprint. https://doi.org/10.48550/arXiv.2205.03257

Khan, A. B., Martínez, P., & Ahmad, R. (2025). PerfCam: A vision-based digital-twin framework for production performance monitoring. Journal of Manufacturing Systems, 79, 245–262. https://doi.org/10.1016/j.jmsy.2025.01.004

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 51(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2019). Industrial artificial intelligence for industry 4.0-based manufacturing systems. Manufacturing Letters, 18, 20–23. https://doi.org/10.1016/j.mfglet.2018.09.002

Li, B., Hou, Y., Che, W., et al. (2022a). Industry 4.0 in smart manufacturing: Trends and key technologies. Journal of Industrial Information Integration, 26, 100257. https://doi.org/10.1016/j.jii.2021.100257

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., et al. (2022b). Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10965–10975). https://doi.org/10.1109/CVPR52688.2022.01069

Li, Z., Zheng, Y., Liu, C., Jiang, X., & Tan, J. (2026). Heterogeneous data integration in digital twins for manufacturing: A survey. Engineering Applications of Artificial Intelligence, 142, 109728. https://doi.org/10.1016/j.engappai.2025.109728

Lim, K. Y. H., Zheng, P., & Chen, C.-H. (2020). A state-of-the-art survey of digital twin: Techniques, engineering product

lifecycle management and business innovation perspectives. Journal of Intelligent Manufacturing, 31(6), 1313–1337. https://doi.org/10.1007/s10845-019-01512-w

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In European Conference on Computer Vision (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, M., Fang, S., Dong, H., & Xu, C. (2021). Review of digital twin about concepts, technologies, and industrial applications. Journal of Manufacturing Systems, 58, 346–361. https://doi.org/10.1016/j.jmsy.2020.06.017

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., et al. (2024). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In European Conference on Computer Vision. https://doi.org/10.48550/arXiv.2303.05499

Liu, Z., Davoli, F., & Borsatti, D. (2025). Industrial Internet of Things (IIoT): Trends and technologies. Future Internet, 17(5), 213. https://doi.org/10.3390/fi17050213

Lu, Y. (2017a). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

Lu, Y. (2017b). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

Lu, Y., Liu, C., Wang, K. I.-K., Huang, H., & Xu, X. (2020). Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. Robotics and Computer-Integrated Manufacturing, 61, 101837. https://doi.org/10.1016/j.rcim.2019.101837

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024c). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Mayershofer, C., Holm, D.-M., Molter, B., & Fottner, J. (2020). LOCO: Logistics objects in context. In Proceedings of the IEEE International Conference on Machine Learning and Applications (pp. 612–617). https://doi.org/10.1109/ICMLA51294.2020.00102

Mohanraj, S., & Vaishnavi, R. (2025). Industrial digital twin deployment: From data integration to actionable analytics. International Journal of Production Research, 63(5), 1842–1867. https://doi.org/10.1080/00207543.2024.2367894

Munappy, A. R., Bosch, J., Olsson, H. H., Arpteg, A., & Brinne, B. (2022). Data management for production-quality deep learning models: Challenges and solutions. Journal of Systems and Software, 191, 111359. https://doi.org/10.1016/j.jss.2022.111359

Nagasubramanian, K., Singh, A., & Sarkar, S. (2026). Computational efficiency of deep learning pipelines for industrial vision: A benchmark study. Computers in Industry, 158, 104247. https://doi.org/10.1016/j.compind.2025.104247

Negri, E., Fumagalli, L., & Macchi, M. (2017). A review of the roles of digital twin in CPS-based production systems. Procedia Manufacturing, 11, 939–948. https://doi.org/10.1016/j.promfg.2017.07.198

Pan, Y., Hu, T., Zhang, C., & Liu, S. (2021). A multi-level digital twin architecture for cyber-physical production systems. Robotics and Computer-Integrated Manufacturing, 70, 102135. https://doi.org/10.1016/j.rcim.2021.102135

Paulin, G., & Ivasic-Kos, M. (2023). Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. Artificial Intelligence Review, 56(9), 9221–9265. https://doi.org/10.1007/s10462-022-10358-3

Peng, C., Yin, S., & Gao, H. (2018). Survey on recent advances in networked control systems. IEEE Transactions on Industrial Informatics, 14(3), 1043–1052. https://doi.org/10.1109/TII.2018.2809315

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., & Saenko, K. (2018). Visual domain adaptation challenge (VisDA-2017). arXiv preprint. https://doi.org/10.48550/arXiv.1710.06924

Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and Industry 4.0: 360 degree comparison. IEEE Access, 6, 3585–3593. https://doi.org/10.1109/ACCESS.2018.2793265

Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., Wang, L., & Nee, A. (2021). Enabling technologies and tools for digital twin. Journal of Manufacturing Systems, 58, 3–21. https://doi.org/10.1016/j.jmsy.2019.10.001

Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers from a modeling perspective. IEEE Access, 8, 21980–22012. https://doi.org/10.1109/ACCESS.2020.2970143

Rawal, R., Khan, M., & Krishnamurthi, R. (2023). A photorealistic synthetic data pipeline for industrial object detection using physically based rendering and YOLOv7. Computers and Industrial Engineering, 183, 109459. https://doi.org/10.1016/j.cie.2023.109459

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788).

https://doi.org/10.1109/CVPR.2016.91

Rosen, R., von Wichert, G., Lo, G., & Bettenhausen, K. D. (2015). About the importance of autonomy and digital twins for the future of manufacturing. IFAC-PapersOnLine, 48(3), 567–572. https://doi.org/10.1016/j.ifacol.2015.06.141

Schäfer, M., Strohmeier, M., & Lenders, V. (2023). Privacy-preserving image acquisition in industrial environments: A systematic review. Computers & Security, 130, 103245. https://doi.org/10.1016/j.cose.2023.103245

Schleich, B., Anwer, N., Mathieu, L., & Wartzack, S. (2017). Shaping the digital twin for design and production engineering. CIRP Annals, 66(1), 141–144. https://doi.org/10.1016/j.cirp.2017.04.040

Shao, G., & Helu, M. (2020). Framework for a digital twin in manufacturing: Scope and requirements. Manufacturing Letters, 24, 105–107. https://doi.org/10.1016/j.mfglet.2020.04.004

Sharma, A., Kosasih, E., Zhang, J., Brintrup, A., & Calinescu, A. (2022). Digital twins: State of the art theory and practice, challenges, and open research questions. Journal of Industrial Information Integration, 30, 100383. https://doi.org/10.1016/j.jii.2022.100383

Sheuly, S. S., Ahmed, M. U., & Begum, S. (2022). Machine-learning-based digital twin in manufacturing: A bibliometric analysis and evolutionary overview. Applied Sciences, 12(13), 6512. https://doi.org/10.3390/app12136512

Singhal, A., Kumar, P., & Rastogi, P. (2023). Bridging the simulation-to-reality gap in industrial computer vision. Pattern Recognition Letters, 174, 130–139. https://doi.org/10.1016/j.patrec.2023.08.015

Sisinni, E., Saifullah, A., Han, S., Jennehag, U., & Gidlund, M. (2018). Industrial Internet of Things: Challenges, opportunities, and directions. IEEE Transactions on Industrial Informatics, 14(11), 4724–4734. https://doi.org/10.1109/TII.2018.2852491

Sommer, M., Stjepandić, J., Stobrawa, S., & von Soden, M. (2023). Automated generation of digital twins for built environments using 3D point clouds and deep learning. Computers in Industry, 145, 103823. https://doi.org/10.1016/j.compind.2022.103823

Soori, M., Arezoo, B., & Dastres, R. (2023). Digital twin for smart manufacturing: A review. Sustainable Manufacturing and Service Economics, 2, 100017. https://doi.org/10.1016/j.smse.2023.100017

Sundermeyer, M., Marton, Z.-C., Durner, M., Brucker, M., & Triebel, R. (2018). Implicit 3D orientation learning for 6D object detection from RGB images. In European Conference on Computer Vision (pp. 712–729). https://doi.org/10.1007/978-3-030-01231-1_43

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. International Journal of Advanced Manufacturing Technology, 94, 3563–3576. https://doi.org/10.1007/s00170-017-0233-1

Tao, F., Zhang, M., Liu, Y., & Nee, A. Y. C. (2019). Digital twin driven prognostics and health management for complex equipment. CIRP Annals, 68(1), 169–172. https://doi.org/10.1016/j.cirp.2019.04.055

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 23–30). https://doi.org/10.1109/IROS.2017.8202133

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., et al. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 969–977). https://doi.org/10.1109/CVPRW.2018.00143

Urbina Coronado, P. D., Lynn, R., Louhichi, W., Parto, M., Wescoat, E., & Kurfess, T. (2018). Part data integration in the shop floor digital twin: Mobile and cloud technologies to enable a manufacturing execution system. Journal of Manufacturing Systems, 48, 25–33. https://doi.org/10.1016/j.jmsy.2018.02.002

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7464–7475). https://doi.org/10.1109/CVPR52729.2023.00721

Ward, R., Soulatiantork, P., Finneran, S., Hughes, R., & Tiwari, A. (2021). Real-time vision-based pallet tracking and digital twin synchronisation for production-line monitoring. Robotics and Computer-Integrated Manufacturing, 70, 102145. https://doi.org/10.1016/j.rcim.2021.102145

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Xu, X., Lu, Y., Vogel-Heuser, B., & Wang, L. (2024). Industry 4.0 and Industry 5.0—Inception, conception and perception. Journal of Manufacturing Systems, 75, 530–543. https://doi.org/10.1016/j.jmsy.2024.05.020

Yasin, A., Pang, T. Y., Cheng, C.-T., & Miletic, M. (2021). Industry 4.0 implementation in SMEs: A systematic literature review. Sustainability, 13(13), 7194. https://doi.org/10.3390/su13137194

Yu, P., Xu, L., Sun, X., & Tian, Y. (2025). Open-vocabulary object detection: A comprehensive survey. International Journal of Computer Vision, 133(1), 1–40. https://doi.org/10.1007/s11263-024-02184-7

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., et al. (2024). LLaMA-Adapter: Efficient fine-tuning of large language models with zero-init attention. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2303.16199

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050

Zhou, J., Hong, Y., Lu, X., Gao, J., & Wang, J. (2022). SOD-DT: Small object detection-based digital twin framework for industrial environments. Computers in Industry, 142, 103746. https://doi.org/10.1016/j.compind.2022.103746

Zhu, X., Bain, M., & Vedaldi, A. (2023). SIP-17: A synthetic-to-real benchmark for industrial parts classification. Pattern Recognition Letters, 165, 30–37. https://doi.org/10.1016/j.patrec.2022.12.011