Embodied Intelligence for Flexible Manufacturing: A Review Postprint

Authors: Xu Kai, Zhao Hang, Hu Ruizhen, Yang Min, Liu Hao, Zhang Hui, Yu Haibin, Xu Kai

Date: 2026-01-16T13:38:49+00:00

Abstract

Driven by the breakthrough progress of the new generation of artificial intelligence, embodied intelligence, as an important branch of artificial intelligence, is rapidly permeating industrial manufacturing scenarios. Owing to the semi-structured nature of manufacturing environments in industrial settings, the relative stability of operating conditions, and the relatively standardized process flows, it is easier to achieve rapid deployment of embodied intelligence technologies, making industrial manufacturing highly likely to become the first large-scale application domain of embodied intelligence. However, as the flexible manufacturing paradigm of "multi-variety, small-batch" production has become the mainstream trend in modern manufacturing, mixed-flow production of multiple product types on the same line, frequent product iteration and upgrading, and increasingly unstructured and non-standard manufacturing processes have become the new normal in the manufacturing industry. This imposes dual requirements on intelligent manufacturing systems, namely, enhanced capabilities for handling complex processes and ensuring manufacturing precision. Consequently, embodied intelligence in flexible manufacturing scenarios faces three core challenges: (1) the difficulty of accurate process modeling and monitoring under limited perception; (2) the challenge of dynamically balancing flexible adaptation and high-precision manipulation; and (3) the challenge of synergistic integration between general-purpose skills and specialized processes. These challenges also bring new opportunities for the development of embodied intelligence itself.

In response to the above opportunities and challenges, this paper provides a survey from three dimensions— "industrial eye, industrial hand, and industrial brain." At the perception layer (industrial eye), it focuses on multimodal data fusion and real-time modeling methods in complex dynamic environments; at the control layer (industrial hand), it conducts an in-depth analysis of flexible, adaptive, and precise manipulation methods for complex manufacturing processes; at

the decision-making layer (industrial brain), it systematically summarizes intelligent optimization methods for process planning and production line scheduling. From the perspective of multi-level technological synergy and interdisciplinary integration, the paper reveals the key technological pathways of embodied intelligence for closed-loop optimization of the "perception–decision–execution" cycle in manufacturing systems, proposes a three-stage evolution model for embodied intelligence in flexible manufacturing scenarios— "cognitive enhancement, skill leap, and system evolution" —and explores future development trends. The aim is to provide a theoretical framework and practical reference for the interdisciplinary and integrated development of industrial embodied intelligence under the trend of flexible manufacturing.

Full Text

1 Industrial Eye

Industrial Eye aims to achieve precise perception and real-time monitoring of manufacturing environments and operational objects. Confronted with the challenges of multi-variety and small-batch production in flexible manufacturing, its perceptual capabilities must continuously improve in terms of accuracy, robustness, and cross-scenario transferability. Recent advances in computer vision and graphics have provided solid support for Industrial Eye. 3D vision technology enables sub-millimeter-level imaging, detection, and measurement, overcoming the limitations of 2D vision—such as insufficient accuracy in complex assembly environments, lack of spatial information, and sensitivity to environmental changes—thereby effectively ensuring consistency and precision in manufacturing processes. Multimodal perception systems integrate visual, tactile, acoustic, and other multi-source information, enhancing perceptual stability and reliability under dynamic conditions, occlusions, and noise interference. Furthermore, large-scale pre-trained vision models possess powerful feature extraction and generalization capabilities, enabling Industrial Eye to rapidly adapt to new products, processes, and scenarios, achieving zero-shot or few-shot transfer across tasks. This section systematically elaborates on how Industrial Eye comprehensively perceives production environments through these key technological pathways, and demonstrates its practical applications through typical cases.

1.1 3D Vision High-Precision Imaging

In flexible manufacturing workshops, frequent production task switching imposes higher demands on defect detection and dimensional measurement, which have become critical for achieving closed-loop manufacturing and ensuring product quality. However, traditional methods suffer from low automation, poor efficiency, inconsistent results, and excessive reliance on manual experience, making them ill-suited for high-takt, high-precision, and full-coverage inspection requirements. Take the commonly used coordinate measuring machine (CMM) method [22] as an example: it exhibits poor adaptability to complex structural

parts, and its reliance on measurement programs and manual operations leads to low efficiency during product variety switching. In contrast, 3D vision, leveraging high-resolution sensors, can reconstruct 3D geometric models of workpiece surfaces, enabling automated defect recognition through geometric feature analysis and outputting dimensional and morphological parameters at millimeter or even sub-millimeter levels, thereby improving measurement efficiency, accuracy, and consistency. This subsection introduces key technologies of 3D vision in high-precision imaging, defect detection, and dimensional measurement.

Real-time Precision Imaging Based on 3D Vision. 3D vision captures visual data of environments or objects through sensors and converts it into 3D geometric information, enabling precise modeling of scenes or objects. Compared to 2D images, 3D imaging provides richer spatial information, allowing more accurate identification of complex object shapes, structures, and positional relationships to achieve fine defect detection and high-precision dimensional measurement, thereby enhancing product quality and manufacturing precision. In recent years, 3D imaging technology has developed rapidly, with diverse methods each having distinct focuses. This paper categorizes them into three types based on imaging principles: structured light, Time-of-Flight (TOF), and multi-view RGB imaging, as shown in Figure 3 [Figure 3: see original paper], with performance comparisons provided in Table 2 .

Structured Light-Based 3D Imaging projects coded light patterns (e.g., stripes, dot matrices) onto workpiece surfaces and calculates surface 3D coordinates through optical triangulation by analyzing light signal deformations caused by surface geometry [26]. This method does not rely on surface texture and offers high precision and fast imaging advantages, though performance degrades under long-distance measurement or strong light interference. The technology traces back to Takeda et al.'s Fourier Transform Profilometry (FTP) [27], which achieved high-density reconstruction by projecting sinusoidal stripes and performing phase Fourier analysis, but suffered from phase wrapping and light sensitivity issues. Phase-Shifting Profilometry (PSP) [28] employs multi-frame phase shifting to recover absolute phase, improving depth accuracy. Multi-frequency phase-shifting algorithms [23] combine low-frequency phase unwrapping with high-frequency decoding to achieve high-precision reconstruction with fewer frames. To adapt to dynamic scenes, De Bruijn sequence encoding [29] utilizes single-frame projection of multiple non-repeating patterns, providing unique identifiers for each pixel to support rapid scanning of moving objects. For highly reflective or smooth materials where traditional structured light is limited, Phase Measuring Deflectometry (PMD) [30] reconstructs surface gradients by projecting sinusoidal stripes and calculating phase shifts, then integrates them to reconstruct 3D morphology, offering greater precision and stability on highly reflective surfaces. Additionally, line-scan cameras [31] are often used for 3D imaging. Their principle involves projecting a laser or structured light line onto the workpiece surface, while a line-array camera continuously captures depth information along the movement direction and stitches it into a complete 3D model, making them suitable for PCB, coil material, and metal pipe inspec-

tion [32–34].

Time-of-Flight (TOF)-Based 3D Imaging emits modulated light (typically infrared), measuring the time difference or phase shift of light from emission to reflection and return to obtain depth information for each pixel in real time. For instance, LiDAR emits laser beams and receives reflected light, calculating propagation time differences while rotating to scan and record sparse point clouds [36]. It offers strong anti-interference capability, suitable for long-distance and bright-light environments, but is bulky, costly, and has limited reconstruction accuracy. In comparison, RGB-D cameras use area-array sensor designs, capturing dense depth maps of full scenes without mechanical scanning, featuring compact hardware suitable for real-time applications. KinectFusion [37] utilizes GPU-accelerated Truncated Signed Distance Function (TSDF) voxel fusion for real-time dense surface reconstruction. ROSEFusion [38] enhances reconstruction stability under high-speed camera motion through particle filter optimization. MIPS-Fusion [39] combines multi-submaps with gradient tracking for online reconstruction of large-scale scenes. Multi-view reconstruction requires registration of point clouds from different viewpoints [40], with Figure 4 [Figure 4: see original paper] showing an example of multi-view point cloud joint registration for aero-engine blade reconstruction. Ge et al. [41] built a dual-camera system, improving the ICP (Iterative Closest Point) algorithm with workpiece linear motion for rapid online modeling in robotic spray painting scenarios. Peng et al. [35] enhanced aero-engine blade reconstruction accuracy through variable-parameter graph optimization and unscented Kalman filtering. GeoTransformer [42] designs geometric invariance encoding for low-overlap point clouds, proposing an end-to-end registration method that achieves hundred-fold acceleration without RANSAC.

Multi-View RGB Image-Based 3D Imaging captures synchronized color images from multiple viewpoints, combining feature extraction and matching algorithms (e.g., SIFT [43], ORB [44]) with camera parameters and triangulation to recover 3D coordinates and fuse them into complete point cloud models. It requires no active light source and simultaneously obtains geometric and texture information, but reconstruction accuracy depends on the number of viewpoints, camera calibration, and texture quality. Multi-view reconstruction originates from Structure from Motion (SfM) [45], which first detects local features in images and performs cross-view matching, then recovers scene 3D structure through camera pose estimation and sparse point cloud triangulation. Building upon SfM' s initial camera poses and sparse point clouds, Multi-View Stereo (MVS) [46] generates dense point clouds. Neural networks have enhanced reconstruction capabilities in weak-texture and occluded regions. Representative works include MVSNet [47] using multi-view cost volumes for depth prediction, and SuperGlue [48] achieving high-quality feature matching through graph neural networks, improving measurement reliability. Mature tools such as COLMAP [49–50], OpenMVS [51], and AliceVision [52] provide complete image-based modeling pipelines. Recently, Neural Radiance Fields (NeRF) [53] and Gaussian Splatting (GS) [54] have achieved high-quality novel view synthe-

sis through implicit scene modeling and volume rendering, but their geometric accuracy and efficiency are limited, resulting in fewer industrial applications. Emerging foundation models DUSt3R [55] and VGGT [56] bypass traditional SfM pipelines, directly reconstructing scenes from multiple images and inferring image matching and camera parameters, providing high-quality inputs for industrial vision tasks.

Defect Detection and Dimensional Measurement. Based on 3D imaging results, key information such as surface defects and geometric features can be extracted, bridging "perception"and "execution"in flexible industrial production. Precise detection helps identify quality issues in real time, while geometric feature extraction supports positioning and measurement, improving efficiency and consistency. Faced with multi-variety and small-batch demands, rapid and accurate feedback of this key information is essential for achieving process switching and high-precision manufacturing.

Industrial vision system defect detection targets physical anomalies in manufacturing processes (e.g., cracks, pits, assembly misalignments), preventing economic losses and safety risks by locating surface flaws. Compared to traditional 2D image methods [57–58], 3D vision introduces depth information that accurately reflects surface geometric changes, avoiding texture and lighting interference, and quantifying defect dimensions, depth, and volume to support subsequent repair and optimization. Auerswald et al. [59] achieved micron-precision full-tooth 3D reconstruction of large gears based on laser line triangulation, enabling micron-level detection of chipping and scratches. Li et al. [60] detected metal thickness fluctuations based on point cloud registration, achieving 0.1mm-level detection precision. Yan et al. [61] implemented fine-grained defect point cloud segmentation through density clustering and region growing algorithms, effectively supporting pipeline inner wall defect detection. Huang et al. [62] designed an entropy-driven neighborhood fitting algorithm for sub-millimeter crack detection and fitting error evaluation on complex curved magnetic tiles. Vokhmintcev et al. [63] proposed the Fusion-ICP algorithm, optimizing point cloud registration through orthogonal transformations for assembly gap and bending deformation modeling.

3D vision measurement extracts geometric features from object surfaces to achieve high-precision, non-contact dimensional inspection, particularly suitable for large workpieces such as heavy machinery and aerospace components. Compared to traditional contact or point-based measurement methods like CMMs, laser trackers, or ultrasonic thickness gauges, 3D vision offers greater flexibility and efficiency, effectively avoiding scratches and deformation risks from contact, and achieving full-coverage measurement of complex curved surfaces through multi-view fusion. Leveraging high-resolution sensors and precise calibration, the system can achieve micron-level data acquisition in pipeline or online scenarios. Yin et al. [64] developed a large freeform surface scanning system integrating structured light, stereo imaging, and error compensation modules, achieving $±$0.2 mm precision non-contact measurement. Wang et

al. [65] utilized a robot-mounted stereo vision system combined with hand-eye calibration and pose tracking to automatically scan components such as wind turbine blades. Huang et al. [66] employed multi-view phase-shifting structured light with feature-constrained registration to effectively address metal highlight interference, achieving complete reconstruction of turbine blades. Ma et al. [67] applied a dual line-scan camera high-resolution system to continuous scanning of complex structures like engine housings, achieving reconstruction errors below 0.05mm with industrial-grade stability suitable for high-speed inspection and offline quality control.

Case Study 1: Automotive Paint Surface Reconstruction and Defect Detection. In automotive manufacturing, paint surface quality directly affects vehicle appearance and market competitiveness. Traditional detection relies on manual inspection with low efficiency, poor precision, and false detection rates as high as 15%-20%, with missed detection rates exceeding 30% under high-fatigue conditions. Typical defects such as dust particles, pinholes, orange peel, and sagging range in size from 0.05-0.3mm, making stable identification difficult. The development of 3D vision has enabled automatic paint defect identification, automatically detecting minute defects while reducing human interference to ensure paint quality control.

In paint 3D reconstruction, traditional structured light or laser systems are affected by highly reflective surfaces, generating noise and scanning voids that fail to meet minute defect localization requirements. In contrast, Phase Measuring Deflectometry (PMD)-based reconstruction technology accurately captures surface morphology under high-reflection conditions by measuring phase changes in reflected light. The deflectometry camera is shown in Figure 5(a) [Figure 5: see original paper]. Since vehicle bodies are much larger than a single camera' s field of view, multi-camera and multi-robot collaboration significantly improves measurement efficiency, achieving full-coverage vehicle paint inspection through path planning, as shown in Figure 5(b) [Figure 5: see original paper]. Faced with minute defects under highlight or shadow occlusion, traditional image processing struggles to identify them. Combining 3D defect detection and classification methods with body model registration enables high-precision defect localization and identification [68]. After visual AI system processing, missed detection rates can be reduced to <1% and false detection rates to <3%, providing precise support for subsequent repair processes such as grinding and respraying.

Case Study 2: Imaging in Unstructured Welding Scenarios for Shipbuilding. In shipbuilding, welding processes account for approximately 40% of total hull construction costs [69], yet remain primarily manual operations with low efficiency, unstable quality, and high costs. Intelligent methods are urgently needed to improve welding efficiency and consistency. In flexible manufacturing environments, welding tasks typically require dimensional precision within $±$0.05 mm to $±$0.5 mm. However, large vessels such as oil tankers, passenger ships, or warships often involve over a million components, including deck plates, pipe joints, and support beams with diverse shapes, as shown in Figure

6(a) [Figure 6: see original paper]. Meanwhile, weld structures are complex, massive, and lack standardization. Accurate acquisition of workpiece geometry, weld position, and key parameters such as groove angle and width is essential to provide reliable geometric basis for torch trajectory planning.

Deep stereo matching analyzes multi-view image pairs to rapidly generate pixel-level depth maps and reconstruct 3D scenes, offering lightweight acquisition and sub-pixel precision that demonstrates good adaptability in unstructured welding scenarios. However, deep stereo matching relies on scene-specific fine-tuning, making it difficult to adapt to frequent switching in welding production lines. FoundationStereo [70] proposes a method for high-precision dense depth estimation without fine-tuning (Figure 6(b) [Figure 6: see original paper]), achieving $\pm$0.2 mm-level depth reconstruction through million-scale high-fidelity synthetic image self-supervised pre-training, combined with side-tuning structures that introduce monocular priors and integrate spatial-disparity attention mechanisms to effectively suppress occlusion and noise interference. This method exhibits strong generalization capabilities, providing stable 3D perception foundations for workpiece recognition, weld detection, and path planning in flexible welding.

1.2 Multi-Source Data Fusion Perception

Flexible manufacturing involves diverse tasks, complex working conditions, and frequent switching, imposing higher requirements on perception system robustness and adaptability. Single-modality sensing (e.g., relying solely on vision, touch, or acoustics) is prone to information loss or misjudgment under occlusion, reflection, surface variations, or environmental noise, making it difficult to support stable perception and environmental understanding in complex tasks. Multimodal perception fusion technology [71–72] integrates information from vision, depth, force, acoustics, and infrared to achieve comprehensive perception of object states, environmental constraints, and interaction processes, offering advantages such as information complementarity, strong anti-interference capability, and rich expressive power. Figure 7 [Figure 7: see original paper] shows an example of image and point cloud fusion for downstream segmentation and detection, while Table 3 summarizes characteristics of common modalities.

Multimodal Feature Extraction. Industrial multi-source sensing data exhibits significant structural heterogeneity: RGB/RGB-D cameras output multi-channel images where depth information can be converted to point clouds or voxel grids; force-torque, Inertial Measurement Units (IMU), audio, voltage, and current are high-frequency 1D time-series data; infrared thermography represents temperature distributions as grayscale images. Different modalities require effective feature extraction and representation to provide foundations for subsequent alignment and fusion.

At the feature extraction stage, each modality relies on different deep learning architectures emphasizing distinct feature advantages. For RGB/RGB-D

images, deep convolutional neural networks (e.g., ResNet [74]) excel at capturing local texture and edge features, while Vision Transformer (ViT) [75] based on self-attention mechanisms can model global contextual information. Depth maps or point clouds often utilize PointNet [76] and its hierarchical extension PointNet++ [77] to learn global-local geometric features in unordered point sets, while Point Transformer [78] further employs self-attention to capture long-range dependencies and directional information. For high-frequency 1D signals such as force-torque, IMU, current, and voltage, 1D Convolutional Neural Networks (1D CNN) [79], Temporal Convolutional Networks (TCN) [80], or Gated Recurrent Units (GRU) [81] are typically adopted to efficiently model temporal dependencies through dilated convolutions or gating mechanisms. Acoustic and vibration data are first mapped to 2D spectrograms via Fourier transform, then processed by convolutional networks to extract frequency-domain features. Infrared thermography and other grayscale images often employ U-Net [82] or dual-branch networks [83] to preserve fine-grained features, thereby improving defect detection rates and model generalization.

Although mature feature extraction methods exist for each modality, their data acquisition and training time costs remain high for industrial applications. In flexible production environments, models require rapid iteration to adapt to frequently switching processes and equipment. Applying pre-trained models can effectively reduce training costs. In the RGB domain, released self-supervised ViT models MAE [84] and DINO [85] provide large-scale pre-trained weights that can adapt to defect detection tasks with minimal or no fine-tuning. For 3D point clouds, PointMAE [86] and PTv3 [87] demonstrate excellent zero-shot and few-shot generalization capabilities for industrial point cloud registration and grasp pose estimation after pre-training on ShapeNet and other datasets. Audio and vibration signals can directly utilize PANNs [89] and YAMNet [90] pre-trained on AudioSet [88], mapping spectrograms to universal acoustic embeddings for industrial fault diagnosis. For end-to-end temporal features, pre-trained backbones like PatchTST [91] and TimesNet [92] can be introduced for rapid fine-tuning on industrial time-series data. In the thermal infrared domain, InfMAE [93] provides a self-supervised pre-trained model based on million-scale thermal frames, offering good initialization for thermal defect detection and workpiece temperature monitoring. These pre-trained libraries provide reliable universal representations for multi-source data fusion in industrial scenarios, reducing annotation and training costs to meet the demands of rapid model iteration and deployment in flexible production systems.

Heterogeneous Feature Alignment. Completing single-modality feature extraction is only the "first mile" of multi-source perception; deep feature alignment and fusion are crucial for achieving multimodal perception. Multimodal alignment refers to mapping features from different modalities into a shared semantic space to achieve cross-modal semantic consistency. Multimodal feature fusion emphasizes hierarchical integration of complementary information from each modality according to task requirements after alignment, generating more complete and reliable comprehensive representations that cannot be provided

by single modalities.

Multimodal feature alignment methods encompass various strategies, including contrastive learning, joint embedding, and attention mechanisms, to accommodate distribution and structural differences between modalities. Contrastive learning pulls together homologous positive samples while pushing apart heterogeneous negative samples, mapping different modalities into a shared semantic space. The representative CLIP [94] tightly aligns image-text features and demonstrates powerful transfer capabilities in zero-shot defect recognition [95-96]. Joint embedding methods focus on designing projection heads or explicit latent variables (e.g., deep canonical correlation analysis [97], variational joint distribution [98]) to learn common representations, thereby eliminating dimensional and scale differences between modalities. Such methods have validated high efficiency and interpretability in multi-source sensor fusion and cross-modal retrieval [99].

Attention mechanisms [100] can explicitly capture inter-modality correspondence and complementary dependencies at the token-level features, preserving both local details and global semantics, and have been widely applied. MulT [101] is one of the earliest systematic works introducing cross-modal attention mechanisms, designing Crossmodal Attention modules that align different modality sequences (audio, vision, text) directly via Transformer without explicit temporal synchronization. EMT [102] achieves missing modality reconstruction and alignment through a dual-level recovery module, particularly suitable for real-world multimodal scenarios with incomplete data. AnyGPT [103] introduces "any-to-any" multimodal conversational capabilities, mapping all modalities related to large language models (audio, text, image) into discrete token sequences without modifying existing model structures or training paradigms. TEAL [104] proposes a "Tokenize and Embed All"strategy, discretizing arbitrary modalities into token sequences and mapping them to a shared embedding space, then generating outputs autoregressively, enabling frozen large language models to efficiently process multiple non-text modalities while preserving text capabilities. M2PT [105] proposes a cross-modal path enhancement method, constructing modality-agnostic cross-model parameter sharing mechanisms for heterogeneous modality knowledge transfer. Meta-Transformer [106] demonstrates its potential in multimodal learning through a unified framework, supporting inputs of various modalities including hyperspectral images, audio, video, time series, and point clouds.

Multimodal Feature Fusion. After completing heterogeneous feature alignment, multimodal fusion aims to organically integrate information from each modality to construct more comprehensive and robust representations. The fusion process must ensure spatiotemporal and semantic alignment, leverage complementary advantages of different modalities, avoid redundancy and information loss, balance real-time performance and computational complexity, and maintain overall performance when single modalities fail or suffer from noise interference. Based on fusion timing and methodology, common multimodal

feature fusion methods can be categorized into Early Fusion, Late Fusion, and Intermediate Fusion [107]. Different feature fusion methods have their own advantages and limitations in industrial applications, as summarized in Table 4 .

In early fusion, data from different modalities are directly concatenated to form a unified feature representation before being input to the model for processing. This approach is simple and intuitive, capable of capturing global relationships between modalities at the perception stage. In industrial robot assembly tasks, early fusion of visual and force information helps robots accurately locate objects in the environment and perform high-precision operations. Lee et al. [108] extracted compact joint representations of force-vision data through cross-modal contrastive learning, transferring pre-trained representations to policy networks to achieve strong robustness to hole shapes, assembly gaps, and external disturbances in peg-hole assembly tasks. However, early fusion suffers from high feature dimensionality and increased computational complexity, particularly when processing high-resolution visual and low-sampling-rate force data, potentially leading to inefficiency.

Late fusion first processes each modality' s data independently to obtain individual prediction results, then combines these results through weighted averaging, voting, or other methods to form final decisions. In a multimodal robot grinding system, image and acoustic sensors separately collect audio signals during grinding, and the processing results from both modalities are fused [109]. When the RMS power from audio feedback falls below a threshold, decisions are made by combining it with powder area radius measured from visual feedback—if the current radius is less than or equal to previously measured values, powder collection is triggered; otherwise, grinding continues. Late fusion' s modular design is easy to implement and debug, but may ignore interactive information between modalities, especially in tasks requiring highly synergistic information fusion, potentially failing to fully exploit complementary advantages.

Intermediate fusion combines features from different modalities at intermediate model layers, typically using attention mechanisms, gating mechanisms, or other methods for interaction. In multimodal industrial anomaly detection, intermediate fusion of visual and force information can effectively capture fine-grained interactive information, enabling robots to make precise decisions in dynamic environments. M3DM [110] proposes fusing multimodal features from RGB images and 3D point clouds, employing patch-level contrastive learning to promote modality interaction and reduce interference, and using multiple memory banks to store different modality features to avoid information loss, ultimately making decisions based on multiple memory banks, achieving leading performance on the MVTec-3D AD [111] industrial anomaly detection dataset. However, intermediate fusion has high design complexity, requiring careful adjustment of fusion strategies, which increases system development difficulty and computational overhead.

Case Study 1: Multimodal Fusion for Part Assembly. Part assembly

and other contact-rich tasks are extremely challenging in industrial scenarios due to high operational uncertainty. The diversity of part geometries (e.g., irregular pins), minute assembly gap differences (typically only 0.1 mm–0.5 mm), and instantaneous contact state changes (e.g., collisions, sliding) all increase manipulation difficulty. Taking automotive door hinge assembly as an example, its fitting tolerance must be controlled within $±$0.05 mm; otherwise, jamming or looseness may occur. Single-modality perception (e.g., relying only on vision or touch) struggles to stably handle occlusion, lighting changes, or spatial perception deficiencies. Multimodal fusion can compensate for modality limitations, enabling more comprehensive task environment understanding and improving robot robustness in complex assembly tasks.

A Stanford University research team [108] fused vision (RGB-D images), touch (six-axis force-torque), and proprioception (end-effector position and velocity) in assembly tasks to enhance state representation completeness, as shown in Figure 8 [Figure 8: see original paper]. To achieve efficient fusion, a variational autoencoder-based architecture was adopted, jointly modeling latent representations from different modalities through the Product of Experts [98]. The model also introduced self-supervised tasks such as optical flow estimation and contact event detection to capture dynamic associations between modalities, providing compact and semantically rich inputs for policy learning. This method improved assembly performance, achieving cross-shape transfer and strong anti-interference capabilities, validating its robustness in real robot systems. This demonstrates that multimodal perception combined with self-supervised learning can reduce dependence on manual annotation, providing support for industrial embodied intelligence deployment in complex dynamic environments.

Case Study 2: Multimodal Fusion for Welding Quality Monitoring. During welding, real-time state monitoring and early prediction of defects such as incomplete penetration, burn-through, and misalignment are required to avoid finished product flaws. Traditional single-modality sensors have perception blind spots: vision is susceptible to arc light interference with errors exceeding 1 mm; acoustics are affected by noise with signal-to-noise ratios often below 10 dB; current/voltage sensors can only reflect heat input information, making it difficult to accurately predict complex weld pool and seam geometries. Multimodal fusion [112–113] becomes key to addressing these challenges. By integrating visual, acoustic, and electrical signal information, it compensates for single-source limitations, enhances prediction robustness and generalization, and enables efficient perception and early warning of welding processes.

A Shanghai Jiao Tong University research team [114] designed a multimodal feature extraction and fusion architecture for arc welding processes to predict welding quality defects in advance and guide process adjustments. The architecture covers three modalities: visual, acoustic, and electrical signals, as shown in Figure 9 [Figure 9: see original paper]. The visual modality extracts molten pool area, symmetry, and other features from weld pool images via CNN. The acoustic modality combines time-domain and frequency-domain analysis to ex-

tract features including average energy, amplitude, standard deviation, and time-frequency statistics. The electrical signal modality extracts time-domain statistical features of current and voltage. These heterogeneous features are normalized and input to an LSTM network for fusion, leveraging its time-series modeling capability to capture dynamic correlations. Under mixed-modality inputs, the model can predict defects such as incomplete penetration and burn-through 0–2 seconds in advance, providing a critical time window for process adjustments.

1.3 Industrial Vision Foundation Models

In flexible manufacturing scenarios characterized by rapid product iteration, variable process flows, and complex environmental conditions, vision systems require high generalization and adaptability. Traditional vision models, typically custom-trained for specific tasks, struggle to flexibly transfer across scenes and processes, resulting in low model reusability and high maintenance costs. Consequently, industrial vision foundation models have gained widespread attention. Through large-scale pre-training, foundation models possess capabilities such as few-shot adaptation, cross-task generalization, and multimodal extension, effectively supporting rapid deployment and stable operation of various vision tasks in flexible production. Building a visual general capability hub through foundation models is expected to become an important direction for advancing industrial perception systems toward intelligence and platformization.

Learning and Generalization of Vision Foundation Models. Vision Foundation Models (VFM) are vision models pre-trained on large-scale image or video data, possessing broad adaptability and strong generalization capabilities. Through a single pre-training, they support tasks including image classification [85, 94, 115], object detection [116–117], image segmentation [118–119], and 3D understanding [55, 120–121], reducing development thresholds for task-specific models and enabling "plug-and-play" visual intelligence.

Self-supervised learning has become the mainstream approach for VFM pre-training. It designs pre-training tasks without manual annotations, allowing models to learn features from large-scale unlabeled data. Mainstream methods include Masked Image Modeling (MIM), contrastive learning, and teacher-student frameworks. Inspired by BERT [122], MIM randomly masks image regions for reconstruction, enhancing model semantic modeling capabilities, particularly suitable for Vision Transformer architectures. Contrastive learning methods such as SimCLR [123], MoCo [124], and BYOL [125] construct positive and negative sample pairs to learn discriminative global semantic features, emphasizing modeling of overall image structure and semantics to enhance cross-task transferability. Teacher-student methods (e.g., DINO [115]) leverage knowledge distillation [126] mechanisms to transfer teacher model knowledge to lightweight student models, learning high-quality representations without using negative samples while maintaining good stability and generalization.

VFM generalization ensures strong model adaptability when facing diverse tasks, primarily manifested in three aspects: (1) Unified task modeling: By constructing universal input-output pairs, models can jointly handle classification, segmentation, detection, and other tasks. For example, LaVin-DiT [127] uses a joint diffusion Transformer and spatiotemporal variational autoencoder to support 20+ tasks; (2) In-context learning: Models achieve zero-shot or few-shot task transfer by introducing task descriptions or examples. MetaVL [128] first transfers in-context learning from language models to vision-language models, enabling efficient adaptation of compact models; (3) Multi-task learning and shared representations: Uni-Perceiver v2 [129] uses a unified maximum likelihood strategy to process vision and vision-language tasks, demonstrating cross-modal generalization capabilities even without fine-tuning.

From 2D to 3D Vision Foundation Models. In 2D vision tasks, SAM (Segment Anything Model) [118] represents a significant breakthrough in VFM, enabling zero-shot segmentation based on various prompts (points, boxes, text) with good modularity and task extensibility. Its upgraded version SAM 2 [119] supports video input, achieving real-time segmentation (30 fps) of 4K videos through streaming memory Transformer and interactive data engines. Depth Anything [130] is a depth estimation model that generates pseudo-labels on unlabeled images through a teacher-student architecture, combining Vision Transformer for robust monocular depth estimation. Its V2 version [131] introduces synthetic data to train teacher models, improving depth prediction accuracy. For 2D image pose estimation, NVIDIA' s FoundationPose [132] provides a unified 6D pose estimation and tracking framework supporting CAD or image inputs, achieving cross-object generalization through neural implicit representations and language model policies in contrastive learning. For unified visual feature representation, Meta AI' s DINOv2 [85] employs improved multi-scale ViT trained collaboratively using global and local contrast, view jittering, and knowledge distillation, producing strongly generalizable features for downstream tasks including segmentation, object detection, and keypoint estimation. For multi-view pre-training, 3D-MVP [133] proposed by Stanford University and Google Robotics constructs a large-scale multi-view RGB-D corpus, jointly capturing semantic and geometric features through cross-view contrast and geometric consistency reconstruction, bringing significant performance improvements to pose estimation and assembly tasks.

As 2D foundation models mature, their concepts are expanding into 3D vision. Such models no longer rely on traditional point cloud convolutions or voxel processing, but instead establish spatial-semantic alignment through multi-layer attention mechanisms. 3D-VisTA [134], RangeViT [135], and UniT3D [136] demonstrate Transformer' s powerful expressive capabilities in 3D vision-language tasks, showing trends toward multi-task unification, modality fusion, and semantic generalization. For 3D visual modeling, DUSt3R [55] introduces pointmap representation, predicting 3D structure and relative pose from images without camera parameters, simplifying multi-view reconstruction pipelines. VGGT [56] builds upon this with an end-to-end multi-task

framework that simultaneously predicts camera parameters, depth, pointmaps, and 3D tracking features from single or multiple images, explicitly modeling inter-image geometric consistency through explicit interaction between visual and camera tokens, further improving 3D task performance and efficiency.

Model Fine-Tuning Mechanisms. Although foundation models exhibit good generalization in generic tasks, they still require fine-tuning in industrial-specific tasks to better meet precision and practical requirements. Current mainstream fine-tuning methods include Full Fine-tuning [2], Linear Probing [94], and Parameter-Efficient Fine-Tuning (PEFT) [1]. Full fine-tuning is suitable for resource-abundant scenarios pursuing ultimate performance; linear probing only trains task heads, suitable for rapid adaptation or resource-constrained environments. PEFT balances performance and efficiency, gradually becoming mainstream, especially for multi-task or edge device deployment. Representative methods include Low-Rank Adaptation (LoRA) [137], Adapter Tuning [138], and Prompt Tuning [139–140]. LoRA has small parameter overhead, suitable for low-resource scenarios with large models; Adapter Tuning offers good cross-task transfer capabilities; Prompt Tuning achieves minimal-cost adaptation through input guidance. These three provide flexible fine-tuning strategy choices, gradually becoming standard tools for industrial multi-task systems. Performance comparisons of different PEFT methods are summarized in Table 5 .

Case Study 1: Few-Shot Defect Detection Based on Vision Foundation Models. Defect detection is a core link in industrial manufacturing for ensuring product quality. However, defect data in real industrial scenarios often exhibits long-tail distributions, with common good samples accounting for 95%–99%, and defect types being complex and diverse (e.g., scratches, stains, deformations) with high annotation costs. Traditional deep learning methods rely on large amounts of precisely annotated data [141], making it difficult to meet industrial demands for algorithm generalization and rapid deployment. Vision foundation models possess strong transfer capabilities, enabling high-quality visual representation extraction and rapid adaptation to target tasks, improving defect detection performance.

AnomalyDINO [142] achieves efficient few-shot anomaly detection by introducing the vision foundation model DINOv2 [85], as shown in Figure 10 [Figure 10: see original paper]. The algorithm builds a memory bank based on 2D visual features extracted by DINOv2, while designing random rotations to enhance the generalization of few-shot memory banks. The method leverages DINOv2' s zero-shot segmentation capability to generate object masks for eliminating background noise, enabling localization of anomalous regions in industrial images and achieving performance breakthroughs across multiple industrial detection datasets. Vision foundation models possess geometric perception and semantic understanding capabilities, with pre-trained visual features effectively distinguishing minute defects from normal texture variations. Simultaneously, VFM offers both feature universality and adaptation flexibility, enabling rapid deployment in industrial scenarios to build efficient defect detection systems.

Case Study 2: Industrial Weak-Texture Image Matching. Image matching is a critical preprocessing step for industrial vision inspection, supporting tasks such as target localization, defect identification, and 3D reconstruction. Its core lies in accurately aligning corresponding regions or feature points across viewpoints, time, or conditions to provide precise geometric foundations. In recent years, deep learning-based matching methods have developed. SuperPoint [143] improves robustness by jointly learning features and matching strategies, while LoFTR [144] and COTR [145] introduce Transformers to enhance global modeling capabilities. However, in industrial scenarios with weak textures, repetitive patterns, or highly similar structures, feature points are difficult to distinguish, local matching information is insufficient, and traditional methods easily fail, with matching accuracy still facing severe challenges.

DUSt3R [55], targeting 3D reconstruction, also demonstrates excellent performance in image matching tasks. Its core innovation lies in introducing pointmap representation that maps pixels to 3D space, enabling cross-image registration without camera parameters. Unlike traditional 2D feature methods, DUSt3R understands images from a 3D perspective, embedding them into a globally consistent spatial framework that enhances geometric stability of features. Pointmaps can encode geometric constraints such as normals, depth consistency, and spatial adjacency, maintaining structural continuity and positional stability in matching even in weak-texture scenarios. Compared to traditional methods relying only on local features in the image plane, DUSt3R possesses stronger cross-view invariance and structural alignment capabilities, supporting high-quality matching and point cloud generation while simplifying detection workflows and improving efficiency. As shown in Figure 11 [Figure 11: see original paper], DUSt3R achieves precise stitching of weak-texture automotive chassis images, demonstrating its application potential in industrial scenarios.

2 Industrial Hand

Industrial Hand completes precision operations through automated equipment such as robots, serving as the core execution body of manufacturing tasks. As manufacturing shifts toward flexible, small-batch, and multi-variety production, production line layouts, product types, and process flows become increasingly dynamic. Traditional control modes relying on high-precision, fixed processes have become difficult to adapt, and industrial systems are gradually adopting low-cost, modular hardware. These trends impose higher requirements on Industrial Hand: achieving high-precision operations despite limited hardware precision; possessing flexible adaptation and transfer capabilities for strategies during rapid production line reconfiguration; and dynamically adjusting parameters based on perceptual feedback under complex process constraints to ensure process stability and product consistency. From the perspective of enhancing Industrial Hand's flexible production capabilities, this section discusses compensation methods for control precision, flexible operation schemes under variable production lines, and adaptive process parameter regulation.

2.1 Low-Precision Hardware Precise Control

Flexible manufacturing requires rapid production line reconfiguration, often at the cost of low-cost, low-precision hardware. To ensure system stability and machining precision, compensation must rely on high-precision perception and control technologies. This subsection discusses compensation methods for control precision from three aspects: precise identification of object pose, active adaptation of control strategies, and virtual-to-real transfer.

Precise Identification of Object Pose. Pose identification [147] is fundamental to improving control precision. By estimating the spatial position and orientation of targets from observation data, systems can dynamically adjust operation paths, gripping methods, and interaction strategies to achieve high-precision, robust intelligent operations [148]. In multi-variety, small-batch scenarios, accurate pose identification is particularly critical, providing real-time, generalizable geometric support for different workpieces. For example, Guo et al. [149] generated stable grinding trajectories based on point clouds and pose optimization. Industrial pose identification methods can be mainly categorized into 2D image-based methods and 3D geometry-integrated methods, discussed separately below.

2D image-based pose identification typically relies on RGB images, using deep neural networks to directly regress 6DoF poses of objects, often employing supervised training with synthetic data (generated from CAD models) [150–151] or real annotated data [152–153] to learn rotation and translation parameters. DOPE [154] combines deep networks with PnP algorithms for rapid solution, enhancing robustness to lighting changes and background interference through dual translation correction. To address pose estimation inaccuracies caused by large shape variations within object categories, SOCS [155] proposes a semantic keypoint-guided deformation alignment method, combining coordinate attention and diffusion generation result filtering to improve consistency and precision of complex workpiece matching. Wan et al. [156] further introduce diffusion models and A5 group equivariant structures to jointly estimate object pose and geometry, capable of handling uncertainty scenarios such as occlusion and ambiguity, stably outputting multiple pose solutions even with incomplete inputs.

3D pose identification schemes integrate depth maps and point clouds to explicitly model object spatial structures, particularly suitable for industrial scenarios with textureless surfaces, occlusions, stacking, and multi-instance interference. Liu et al. [158] propose a pixel-level prediction method based on local features, generating dense predictions through encoder-decoder structures to adapt to complex object geometries and occlusions. Research [159] utilizes edge regions and pose verification mechanisms to improve recognition robustness under occlusion and stacking. For industrial scenarios with multiple instances and significant background interference, traditional geometric feature extraction methods are easily disturbed by irrelevant points and instances. MIRETR [157]

effectively isolates target information through instance masking and superpoint feature extraction, improving operation precision, as shown in Figure 13 [Figure 13: see original paper]. Gao et al. [160] propose a differentiable template matching method that introduces edge-aware modules and optimizes point correspondences from coarse to fine to address cross-modal differences between masks and grayscale images, achieving sub-pixel-level registration and further improving industrial part positioning precision and stability.

Active Adaptation of Control Strategies. The widespread application of low-precision sensors and actuators poses challenges to precise control in industrial environments. Traditional control methods rely on fixed parameters and preset strategies, struggling to adapt to dynamic loads and process changes, often leading to decreased precision and stability. In contrast, learning-based methods can optimize control strategies through offline data or online interaction, adapting to dynamic working conditions and possessing capabilities to handle nonlinear, multivariable, and high-dimensional tasks. This subsection discusses these control methods, with characteristics and applicable scenarios of different methods summarized in Table 6 .

Traditional control methods are typically based on precise or linearized models, achieving closed-loop control through feedback/feedforward design, suitable for structurally simple and controllable environments. Proportional-Integral-Derivative (PID) control [161] is simple and practical, the most common industrial approach, suitable for systems with slow dynamics and incomplete models, but limited in performance for nonlinear, highly coupled, multivariable systems. Linear Quadratic Regulator (LQR) [162] minimizes control cost under known state spaces, suitable for linear systems but sensitive to modeling errors. Fuzzy Logic Control [163] operates without precise models, based on rule libraries, suitable for systems with sufficient experience but difficult modeling. Evolutionary and swarm intelligence optimization methods [164] are applicable when models are non-differentiable or objective functions are complex. However, traditional methods typically require offline parameter tuning, easily fall into local optima, and lack real-time performance, making them difficult to cope with rapid changes in industrial environments.

Learning-based control methods adjust strategy parameters through historical data or environmental interaction to adapt to dynamic loads in production environments. Imitation Learning (IL) [165–166] learns control strategies by imitating expert data, accelerating training and improving adaptability to high-dimensional tasks. For example, Ng et al. [167] captured skilled worker trajectories to extract process strategies, while Zhang et al. [168] integrated trajectories, force control, and impedance regulation to achieve high-quality human-like operations. These methods enhance robot adaptability in unstructured environments. However, IL suffers from distribution shift issues—when agents encounter new states significantly different from training data, they may produce unforeseen actions, leading to poor generalization. Reinforcement Learning (RL) [169] optimizes strategies through environmental interaction, demonstrating stronger

adaptability in complex industrial tasks. Xu et al. [170] implemented multi-robot collaborative path planning and obstacle avoidance based on MADDPG, Zhong et al. [171] introduced inverse kinematics priors to improve policy learning efficiency, and Hu et al. [172] combined maximum entropy RL with primary-secondary action structures to enhance welding path planning stability and sampling efficiency. Although RL methods are more flexible, they typically have high training overhead, slow learning processes, and unstable initial strategies, limiting rapid deployment in industrial scenarios.

Combining existing controllers with active exploration from RL has become a more efficient solution [173]. Robots can further improve based on mastered skills to complete tasks more precisely and efficiently. A common approach is to first initialize policies using imitation learning, then fine-tune them through RL. Deepmimic [174] uses imitation rewards to guide policies toward reference motions while combining environmental rewards to achieve multi-skill composition and complex scene adaptation. Luo et al. [175] propose a dynamic weighting mechanism that fuses offline demonstrations with online interaction data, enabling policies to inherit human experience while continuously optimizing. DDT [176] uses single demonstrations as dynamic references, combining RL for adaptive tracking of demonstrations and handling environmental disturbances. Early combinations of imitation and RL often relied on hand-designed imitation rewards, easily causing "reward hacking" [177] issues where agents learn policies that achieve high scores without completing actual tasks. AMP [178] uses discriminators to automatically generate adversarial imitation rewards [179], improving robustness and becoming a mainstream paradigm for robot manipulation learning. Residual Policy [180] learns residuals on top of imitation policies, correcting deficiencies through RL to achieve efficient transfer and rapid adaptation.

Virtual-to-Real Transfer. Industrial equipment is expensive and prone to damage, with any failure potentially causing significant losses. Therefore, control strategies are often trained in simulation environments to avoid real-world risks. Common simulation environments include IsaacGym [181], Pybullet [182], and MuJoCo [183]. However, differences between simulation and reality make trained strategies difficult to deploy directly in real scenarios—a problem known as Sim2Real.

Domain Randomization introduces extensive randomness in simulation rendering or physical parameters to enhance strategy robustness to real-world environmental variations. Peng et al. [184] randomized physical parameters to improve adaptation to dynamic changes. Miki et al. [185] trained quadruped robots in diverse simulated physical environments. During testing, robots first probe terrain through physical contact, then pre-plan and adapt gaits, achieving high robustness and speed. Tobin et al. [186] enhanced visual detection capabilities in simulation by randomizing rendering (e.g., textures, lighting, backgrounds) to improve real-world performance. Yue et al. [187] learned domain-invariant representations by randomizing synthetic image styles using real images from

auxiliary datasets. Dai et al. [188] proposed an automated pipeline that transforms real-world scenes into diverse interactive digital environments called "digital cousins," further improving real-world transfer success rates compared to digital twins [189].

Although domain randomization provides a simple Sim2Real pathway, excessive randomization lacks real-world priors, easily expanding simulation space and increasing strategy learning burden [190], leading to conservative or even degraded policy performance [191]. To address this, researchers propose mapping real-world scenarios to simulation (Real2Sim) for policy learning, then transferring simulation policies to real-world Sim2Real testing, forming a Real2Sim2Real closed loop. Effective robot simulation interaction environments primarily consider modeling geometry, visual appearance, and physics [192]. Gaussian Splatting [54, 193–194] integrates object geometry and appearance attributes into Gaussian particles, enabling simultaneous modeling and optimization of appearance and geometry, and is widely applied in Real2Sim2Real pipelines. SplatSim [195] uses Gaussian splatting to generate highly realistic visual rendering, narrowing the Sim2Real visual gap. RobogSim [196] combines Gaussian splatting to optimize scene appearance and geometry, using the IsaacGym physics engine as the underlying dynamics to train control strategies in simulation and transfer them to real-world testing. ManiGaussian [197] learns Gaussian models, then collects real interaction data to fit dynamics for auxiliary policy training, but data-driven dynamics fitting often faces Out-of-Distribution (OOD) challenges, requiring large amounts of real-world interaction data. PIN-WM [198] combines Gaussian splatting with differentiable physics [199], enabling physics-aware perturbations to build Physics-Aware Digital Cousins within the neighborhood of a single task-agnostic interaction trajectory, enhancing policy Sim2Real transfer capabilities.

Case Study 1: Residual Policy Learning for 3C Product Assembly. In computer, communication, and consumer electronics (3C) product assembly tasks, components have complex structures and extremely small tolerances ($\pm 0.05mm$–$\pm$\$0.2 mm), with frequent contact state changes. Slight errors can lead to assembly failure and affect product reliability. Improving execution precision and efficiency of robots on low-precision hardware is a key research problem. Residual policy models combining demonstration data with active exploration provide flexible and efficient solutions for industrial robots. Demonstration data (typically requiring $ $50 demonstration trajectories [200]) enables robots to obtain high-quality training materials in relatively short time, while exploratory learning allows robots to actively optimize policies in dynamic environments.

Tsinghua University [201] proposed an assembly solution combining digital twins and residual policies: using Virtual Reality (VR) devices to collect human multimodal demonstration data including tactile, visual, and voice information, covering grasping, placement, and strategies for handling variations; reconstructing real workshops in digital twin environments for large-scale simulation training

with curriculum learning mechanisms that gradually adapt from ideal conditions to disturbances such as lighting changes and component offsets. This method reduces dependence on large amounts of expert data, enabling stable robot learning and policy optimization in dynamic environments (Figure 14 [Figure 14: see original paper]), providing a feasible path for high-precision industrial assembly.

Case Study 2: Fine Manipulation with Low-Cost Hardware via Virtual-to-Real Transfer. Fine manipulation tasks such as battery insertion and drill installation often require robots to possess high-precision coordination capabilities. Traditional industrial robot systems, while capable of such tasks, are limited by high costs and complex calibration procedures, restricting their scalability in practical flexible production applications. Under low-precision hardware conditions, improving precise control capabilities of industrial robots in real tasks is an important challenge in automation and intelligent manufacturing.

Stanford University' s ALOHA system [202–203] achieves millimeter-precision bimanual coordination on low-cost hardware under $20,000 through imitation learning algorithm ACT, as shown in Figure 15 [Figure 15: see original paper]. To compensate for hardware precision and dynamics coupling issues, ALOHA builds a "perception-policy-execution" closed-loop system, fusing multi-camera visual inputs and Transformer encoders to predict multi-step action sequences, effectively mitigating error accumulation. The system adopts a three-stage Sim2Real transfer framework: first, building a digital twin environment aligned with real equipment; second, progressively increasing task complexity through progressive domain randomization; and finally, optimizing training data through mixed reality systems and human feedback. This method improves success rates in precision tasks such as battery insertion and drill installation from 60% to 96%, reducing operation cycles by 30–40%, providing an effective solution for large-scale application of low-cost equipment in flexible manufacturing.

2.2 Flexible Operation in Variable Production Lines

Traditional industrial production relies on fixed production lines and preset processes, suitable for large-scale single-product manufacturing but difficult to cope with diversification and customization demands. Modern industry requires production lines with higher flexibility, relying on Industrial Hand to achieve rapid adjustment and flexible operation to adapt to different products and process changes. This subsection explores flexible operation schemes for Industrial Hand in variable production lines from three perspectives: general control policy learning, interactive representation design, and decision foundation models.

General Control Policy Learning. The core objective of general control policies is to learn a policy framework enabling rapid transfer of robots across multiple tasks. This is elaborated from two perspectives: policy distillation and meta-learning.

Policy distillation [204–205] transfers policy knowledge from multiple deep net-

works into a single network, enabling the final combined policy to perform well across various environments. AutoMate [206] distills multiple expert assembly strategies into a single generalist policy, achieving generality across 20 assembly tasks. Wu et al. [207] propose a hybrid policy based on teacher-student frameworks, fusing expert behaviors into diffusion policies (Diffusion Policy) [208], supporting dynamic grasping of over 30 object categories. Mosbach et al. [209] address the challenge of dexterous object grasping from clutter, combining RL with policy distillation in a two-stage learning process that effectively grasps diverse objects and demonstrates zero-shot transfer to novel objects.

Meta-learning [210] enables models to effectively utilize existing knowledge for rapid learning and generalization in new tasks by learning from experience across multiple tasks. Meta-RL [211] enables agents to quickly adjust strategies across different environments through multi-task learning, achieving good performance with limited experience. Contact mechanics and friction effects in industrial insertion tasks are difficult to solve through traditional feedback control. Schoettler et al. [212] use meta-RL to capture latent structures of simulation tasks. After pre-training policies in simulation, they quickly adapt to industrial insertion tasks through few real-world trials and error correction. To address high exploration costs, low success rates, and difficulties in adapting to new tasks in insertion tasks, the ODA algorithm [213] fuses offline data, contextual meta-learning, and online fine-tuning, efficiently completing insertion task adaptation with only a small amount of demonstration data and about 30 minutes of online training, reducing exploration costs and improving success rates.

Interactive Representation Design. Interactive representation aims to extract general features across tasks and environments, enabling rapid robot transfer and efficient response to complex changes. By modeling relationships between perception and manipulation, system generalization and adaptation capabilities are enhanced. This is discussed from 2D and 3D perspectives.

In 2D representation, affordance [214] is used to locate interactive regions in images, serving as a bridge between perception and action. For example, in robot grasping scenarios, affordance anchoring helps models find optimal grasping positions on objects. VIMA [215] uses pre-trained detectors to identify target objects and extract segmentation regions, guiding grasping policies to focus on key regions. Mu et al. [216] propose a method combining image simplification and deep networks to optimize grasping poses, significantly improving grasping success rates under sensor blur and complex structures. Instruct2Act [217] combines open-object detection models like SAM [118] with task instructions and 3D coordinates to generate grasping actions that drive robotic arms.

In 3D interactive representation, existing work enhances robot manipulation capabilities in complex environments by designing strongly generalizable 3D interactive representations. She et al. [218] propose a state modeling method based on Interaction Bisector Surface (IBS) that finely expresses relationships between grippers and objects, enabling dexterous grasping of complex shapes. Xiao et al. [219] design dynamic representations of contact regions between end-

effectors and objects, driving pin-pressing grippers to achieve adaptive grasping and in-hand repositioning. Research [220] proposes a cross-gripper policy transfer framework that uses general policies to predict key point displacements, which are then converted to specific control signals by adaptation modules to accommodate different gripper structures. DP3 [221] focuses on using sparse point clouds as compact 3D representations, extracting 3D visual features through efficient point encoders, demonstrating good generalization capabilities.

Decision Foundation Models. Decision foundation models learn shared features and general policies through training across multiple tasks and environments, possessing cross-task generalization capabilities that break traditional method dependencies on single scenarios. Decision foundation models have formed two paradigms, respectively emphasizing "plan-then-act" and "end-to-end decision-making." This section elaborates on representative works from both routes.

The "plan-then-act" paradigm uses Large Language Models (LLM) as high-level planners, parsing natural language task instructions into low-level skill sequences executed by underlying controllers. This method leverages LLM' s world knowledge and logical reasoning for task decomposition [222–223], offering advantages such as strong interpretability, good modularity, and easy integration of symbolic rules, but requires high coverage of skill libraries and consistency of perception-action interfaces. Palm-E [224] encodes perceptual inputs like images and states into latent variables unified with text, processing them jointly through LLM self-attention to output language-form task plans. SayCan [225] combines skill candidates generated by LLM with feasibility scoring networks to select the most executable and effective action sequences. Ha et al. [226] propose an LLM-guided diffusion policy framework that generates diverse trajectories through high-level task planning and failure detection, improving multi-task policy generalization. Chain-of-thought (CoT) simulates human thinking processes, decomposing complex problems into smaller, more manageable steps [227]. Embodied-GPT [228] improves robot task execution success rates by generating more detailed and executable plans through CoT.

In the "plan-then-act" paradigm, LLMs are responsible for generating task plans and skill sequences, while underlying skill libraries are typically predefined. In this context, skill chaining becomes a critical issue—ensuring that the terminal state of the previous skill closely matches the initial state of the next skill to achieve seamless switching [229]. T-STAR [230] optimizes sub-task policies through reward regularization to maximize state overlap between termination and initiation. Huang et al. [231] train high-level schedulers to select goal-conditioned policies most likely to complete transitions for each stage, with schedulers also obtained through RL. However, most methods rely on fixed-order sub-policy retraining, struggling to cope with flexible process adjustments. DeCo [232] proposes a more practical solution: defining a starting keyframe for each skill, then using motion planning [233] algorithms to automatically transition to the next skill' s keyframe after completing each process, enabling

free skill composition.

"End-to-end decision-making" jointly encodes multimodal observations and language instructions to directly output low-level control signals, achieving an integrated closed loop of perception-understanding-control [234–235]. This method trains on large-scale interaction trajectories, enabling zero-shot transfer to new objects and scenes, though long-horizon reasoning and safety verification remain to be breakthrough. LaMo [236] first explores using small-scale GPT-2 as an offline RL policy, trained under a conditional imitation learning framework. Google' s Robot Transformer series [237–238] achieves excellent performance across multiple embodied tasks using larger models and datasets. OpenVLA [239], based on LLaMA 2 [240], DINOv2 [85], and SigLIP [241], trains on 970,000 real demonstration data, supporting consumer-grade GPU fine-tuning with multi-task and language instruction generalization capabilities. ManipLLM [242] fine-tunes multimodal large models to directly predict end-effector poses. 0.5 [243] builds upon 0 [244], integrating heterogeneous task data to support long-horizon execution in open worlds. 3D-VLA [121] fuses 3D scene understanding with action planning, enabling multi-step operation generation capabilities for real physical worlds.

Although end-to-end methods can learn general policies leveraging large-scale data, training a "one-size-fits-all" model for complex and diverse industrial tasks remains challenging. A more feasible path is: using pre-trained policies as foundations, combined with small amounts of task data for efficient fine-tuning to improve adaptability and reduce data requirements. RLDG [245] proposes using RL to generate high-quality demonstrations for fine-tuning fine-grained manipulation tasks, significantly improving model success rates by about 30–50%. ConRFT [246] addresses VLA model issues such as demonstration scarcity, distribution bias, and adaptation difficulties by designing a unified offline-online RL fine-tuning framework. The offline stage initializes models (e.g., OCTO [247]) through small amounts of demonstrations and consistency training, while the online stage achieves rapid safe adaptation through human intervention and interaction data. In 8 high-contact tasks, ConRFT improves average success rates to 96.3% with only 45–90 minutes of fine-tuning. These methods provide feasible paths for rapid adaptation and deployment of general policies in industrial embodied intelligence.

Case Study 1: Multi-Part Flexible Assembly. Faced with demands for multi-type, multi-configuration, and customized production in flexible manufacturing, traditional fixed production lines typically rely on highly customized engineering designs and fixed motion paths, struggling to flexibly handle diverse part assembly needs. Changes in part combination methods directly affect joint strength, requiring robots to possess high adaptive adjustment capabilities for various geometries and poses of parts, as shown in Figure 16 [Figure 16: see original paper]. This poses challenges to traditional robot control technologies.

NVIDIA proposes learning generalist assembly policies through policy distillation [206] to accommodate assembly of parts with diverse geometries and poses.

The method first initializes policies using Behavior Cloning (BC) [248], then fine-tunes them with DAgger [249] and curriculum RL, distilling multiple expert policies into a single generalist policy network. This policy achieves over 80% average success rate across 80 assembly tasks, with 88% success rate for unseen parts under ±0.5–1mm tolerances. In real-world validation, the generalist policy achieves 86.5% success rate across 20 different part categories, demonstrating millimeter-level assembly precision.

Case Study 2: LLM-Guided Generalized Grinding Control. Grinding is a critical process affecting product performance in industries such as wind power, aerospace, and shipbuilding, but faces challenges including diverse defect types and complex workpiece geometries. Post-grinding shape deviation of aero-engine blade airfoil surfaces must be controlled within ±0.03mm; wind turbine blade grinding requires surface roughness Ra 0.8μm and shape error within ±0.2mm; ship steel plate grinding generally allows flatness deviations of ±0.1 mm–±0.3mm. Traditional fixed-process methods struggle to adapt to variable demands, urgently requiring general and adaptive intelligent control technologies.

RoboGrind [250] builds a generalized grinding control system integrating 3D perception, ontological reasoning, natural language interaction, and force control execution based on LLM, as shown in Figure 17 [Figure 17: see original paper]. The system generates structured task instructions from workpiece point clouds and natural language inputs, adaptively completing information and performing contextual reasoning. The underlying layer dynamically optimizes trajectories and force control parameters through model-based RL fused with point cloud feedback. Even with temporary task changes or imprecise language descriptions, the system can stably complete task planning and control generation. This work demonstrates LLM' s potential in task understanding, strategic reasoning, and cross-task transfer, with broad value for extension to other industrial embodied tasks.

2.3 Adaptive Process Parameter Regulation

Unlike robot operation tasks emphasizing trajectory precision, process parameter regulation focuses on dynamic control of key parameters affecting performance and quality during production to ensure process stability and product consistency. Welding, grinding, and assembly are representative industrial processes, all involving complex physical processes and regulation of multiple process parameters. The welding stage requires comprehensive adjustment of arc parameters, wire feed speed, torch speed, and angle to achieve optimal matching of weld pool morphology and heat input. The grinding process requires dynamic control of contact force, cutting speed, grinding head pose, and path planning to ensure surface finish and shape accuracy. Assembly processes require precise management of position/pose, assembly force/torque, motion speed, and environmental conditions to ensure accurate part alignment and stable connections.

Traditional process parameter regulation primarily targets single products, relying on manual experience and trial-and-error to obtain stable parameters on highly structured production lines. For example, Wang et al. [251] investigated the effects of rotational speed, grinding head quantity, and grinding direction on grinding carbon fiber reinforced composites, suggesting parameter combinations that reduce surface roughness. As manufacturing shifts toward multi-variety, small-batch, and high-customization, preset parameters struggle to cope with frequent line changes and rapid adjustments. Flexible manufacturing urgently needs data-driven adaptive parameter tuning technology that dynamically optimizes control parameters through real-time perception of system states and environmental changes, achieving stable and efficient process execution. Data-driven control employs state-space modeling suitable for multi-input-multi-output systems, emphasizing learning parameter-performance mappings from historical data, with strong generalization and online optimization capabilities. The following discusses relevant research progress using welding, grinding, and assembly as examples.

In welding processes, parameters such as current, voltage, wire feed speed, and trajectory directly affect weld structural strength [252] and formation precision [253]. Kershaw et al. [254] propose an adaptive control method combining CNN and MLP that predicts weld width from weld pool images and adjusts speed accordingly. Wang et al. [255] further propose an adaptive control method based on advanced gradient descent that improves control stability by normalizing means and variances of historical and current gradients, reducing error bands and initial weld pool splitting rates. Jin et al. [256] propose a weld pool width control strategy based on RL, validating its effectiveness in GTAW and GMAW welding. Masinelli et al. [257] apply RL to closed-loop adaptive control of laser power in laser welding, achieving welding quality optimization without prior knowledge through agent-feedback systems.

In grinding processes, cut depth and feed rate relate to surface roughness and stability. Excessive cut depth increases grinding force and heat input, easily causing workpiece burning, abrasive clogging, and deteriorating surface roughness and geometric accuracy. Overly high feed rates cause vibration, affecting processing quality [149]. Cheng et al. build data-driven process parameter matching models for parameter optimization in process manufacturing [258]. Liu et al. [259] propose the IPSO-GRNN model that predicts surface roughness by fusing real-time machining data with machining mechanisms and dynamically optimizes cutting parameters. Li et al. [260–261] propose a removal model fusing multiple parameters and a hybrid force-position control method, effectively improving grinding precision and quality. Zhang et al. [168] achieve precise response under complex dynamics by modeling force control parameters and trajectory planning in manual grinding processes. Overall, intelligent control integrating physical modeling and human experience can enhance system flexibility and robustness.

Assembly processes require precise control of end-effector position, pose,

force/torque, and speed to ensure accurate part alignment and stable connections, avoiding damage or looseness [262]. Zhang et al. [263] propose an RL-based fastening method for aero-engine rotor assembly, improving assembly precision by modeling inter-bolt elastic interactions and using GRU networks to predict coaxiality changes. Zhou et al. [264] propose a bolt fastening method combining vision and force information, estimating threaded hole positions through elliptical arc fitting and three-point methods, using passive compliance to monitor and control radial force, and designing adaptive controllers to reduce force impact and improve tracking precision. Shtabel et al. [265] design an automatic control system for small spacecraft assembly based on vision and wireless tools, simplifying hardware configuration and validating its practicality. You et al. [266] propose a visual servoing algorithm fusing disturbance observers with finite-time control, outperforming traditional controllers such as PID, LQR, and MPC in steady-state precision.

Case Study 1: Real-Time Control of Arc Welding Process Parameters. Welding joins materials through heating, pressure, or both, causing melting or plastic deformation at contact surfaces that form connections upon cooling. In flexible manufacturing, welding intelligence directly affects production line adaptability and efficiency. Conventional weld tracking precision requirements are ±0.2–±0.5 mm, while high-precision scenarios (e.g., aerospace, precision piping) require control within ±0.1–±0.2 mm. If weld width or depth errors exceed ±0.5 mm, defects such as incomplete fusion and stress concentration may occur, weakening structural strength and increasing fatigue risk. However, welding processes are affected by multiple factors including materials, assembly errors, thermal deformation, and environmental disturbances. Traditional fixed-parameter control struggles to adapt to complex dynamic conditions, easily causing quality fluctuations and defects.

A University of Kentucky research team proposed a data-driven real-time control method for arc welding [255], with the algorithm framework shown in Figure 18 [Figure 18: see original paper]. They first capture weld pool width through optical imaging and precisely measure it using pixel-level image segmentation networks. Then, based on weld pool width variations, they employ a gradient descent-driven controller to online optimize welding parameters, achieving fast and continuous feedback regulation. This method improves adjustment efficiency, reduces steady-state errors, and can converge weld pool width to target ranges within only seven control cycles.

Case Study 2: Adaptive Parameter Regulation for Grinding Based on Meta-RL. Grinding removes surface irregularities through abrasives to improve flatness and smoothness. In aerospace, automotive, and other industries, critical components often require surface roughness Ra 0.8 μm and flatness error 0.05 mm. Traditional manual or rigid automated grinding exhibits poor adaptability and low efficiency when facing drastic curvature changes, irregular contact, and workpiece variations, making stable 达标 difficult. Online grinding control can dynamically adjust parameters based on workpiece state and tool

wear, improving removal precision and machining efficiency, as shown in Figure 19 [Figure 19: see original paper].

Huazhong University of Science and Technology proposed a meta-RL-based grinding parameter adaptive regulation method [267]. The method maintains a "high-quality experience pool" to prioritize high-return trajectories, improving material removal precision, and combines Model-Agnostic Meta-Learning (MAML) [210] with Proximal Policy Optimization (PPO) to obtain general initial policies through multi-round gradient updates. When facing new tasks, only small amounts of samples are needed for rapid adaptation to different workpiece characteristics and grinding tool conditions. Experiments show that MAML-PPOBE outperforms MAML, PPOBE, SAC, and fuzzy control in removal error and convergence speed, demonstrating stronger robustness and consistency, providing an effective path for high-precision, real-time adaptive grinding control.

3 Industrial Brain

In the industrial embodied intelligence system, Industrial Brain is responsible for global scheduling and decision-making across multiple processes, workstations, and tasks. Unlike traditional scheduling methods relying on experience and static rules, Industrial Brain is data-driven and model-reasoning centered, building an intelligent hub with dynamic adaptation and real-time optimization capabilities. Its core capabilities include: intelligent scheduling and resource allocation for multi-station tasks to cope with order changes and production line reconfiguration; digital modeling and virtual-real synchronization of the entire production process to achieve precise manufacturing state perception and rapid response; and physical modeling and reasoning of complex processes to support higher-precision and more autonomous production control. The following sections discuss core technologies including factory scheduling, digital twin virtual-real synchronization, and world model physical perception.

3.1 Factory Intelligent Scheduling and Production Planning

Flexible production requires systems to quickly adjust plans and dynamically coordinate resources with order changes, addressing complex process flows. In this context, factory-level intelligent scheduling and decision-making become critical for achieving flexible response. On one hand, dynamic constraints such as parallel processes, production line reconfiguration, and rush orders make global optimal scheduling crucial for capacity and delivery efficiency. On the other hand, efficient collaboration among equipment, personnel, and materials demands higher requirements for data-driven, model-aware, and intelligent optimization. To address scheduling challenges with multiple tasks, resources, and high dynamics, this section introduces relevant work from perspectives of task scheduling, path planning, and material warehousing in smart factories.

Job Shop Scheduling. The Job Shop Scheduling Problem (JSSP) [268] aims to allocate reasonable sequences and start/end times for jobs under multi-

process and multi-equipment constraints to minimize production cycles, resource occupation, or delay costs. As the core of production planning, JSSP not only relates to capacity utilization and delivery efficiency but also determines system responsiveness to rush orders and equipment failures. Due to complex constraints and NP-hard nature, it has long been a focus of combinatorial optimization research. Traditional methods such as branch-and-bound [269] are suitable for small-scale exact solutions, while metaheuristic methods like genetic algorithms [270–271], simulated annealing [272], and tabu search [273] can obtain high-quality approximate solutions for medium-scale problems. However, these methods mostly rely on offline optimization, lacking real-time perception and response capabilities for dynamic information, making them difficult to cope with frequently changing tasks and resource scheduling demands in flexible manufacturing, limiting their applicability in actual complex scenarios.

As traditional scheduling methods become increasingly limited in response speed and generalization, learning-based scheduling strategies have become a research hotspot. Zhang et al. [274] use Graph Isomorphism Networks (GIN) [275] to encode scheduling states, combining policy gradients to learn policies directly from data. Research [276] introduces a multi-agent RL framework for rush orders and equipment failures, treating equipment as edge agents and achieving collaborative scheduling through improved PPO and contract net protocols. Destouet et al. [277] propose a multi-objective RL scheduling method for sustainable flexible workshops, balancing efficiency, energy consumption, and delay. Li et al. [278] combine attention mechanisms with cost prediction to improve high-dimensional planning efficiency for synchronous dual-arm rearrangement. Zhang et al. [279] propose an online hierarchical scheduling algorithm for dual-arm robots, using RL for high-level task allocation and heuristic planning for low-level motion to avoid search explosion with increasing objects. Yao et al. [280] fuse critical path neighborhood search with knowledge-guided hybrid optimization to improve solution efficiency and quality. SeEvo [281] incorporates Large Language Models into automatic algorithm design, generating heuristic prompts and individual programs, then achieving dynamic generation and optimization of workshop scheduling strategies through individual and collective self-evolution reflection mechanisms.

Autonomous Mobile Unit Path Planning. In flexible production, path planning for autonomous mobile units (e.g., AGVs) is crucial for improving logistics efficiency and production line throughput. Facing complex environments with shared passages and intersection areas, systems must dynamically plan paths to avoid collisions, congestion, and deadlocks while ensuring timely delivery of critical materials. This problem can be reduced to the Traveling Salesperson Problem (TSP), designing the shortest closed-loop path among all predetermined pickup and drop-off points [282]. Its multi-vehicle extension, the Vehicle Routing Problem (VRP) [283], treats multiple vehicles as multiple "salespersons," solving optimal route sets from warehouses to customers, which aligns well with multi-transport-unit path planning in industrial scenarios. Traditionally, TSP and VRP also rely on exact algorithms [284–285] or metaheuristic

methods [286–288]. However, in large-scale, dynamic, multi-objective scenarios, they often face challenges such as poor solution stability and insufficient real-time performance [289].

To break these bottlenecks, researchers have begun introducing RL, supervised learning, and graph neural networks to explore new paradigms for intelligent, efficient, and scalable path planning. Xue et al. [290] propose real-time collaborative decision-making for multiple AGVs in flow shops through full workshop information sharing and RL, reducing overall completion time. AM [291] combines Transformer attention mechanisms with the REINFORCE algorithm [169] to achieve end-to-end learned solutions for VRP. ELG [292] improves model generalization across different instances by fusing transferable local and global policies. DIFUSCO [293] models TSP as a discrete 0-1 vector optimization problem, using graph denoising diffusion models to generate high-quality paths. DISCO [294] further builds an efficient diffusion solver through residual guidance and analytical acceleration, combining divide-and-conquer strategies to achieve efficient inference for ultra-large-scale NP-hard problems.

Material Warehousing Optimization. Efficiently storing multi-variety materials within limited space to improve system flexibility and reduce inventory costs is also key to enhancing production system flexibility and reducing inventory costs. This problem can be formalized as the Bin Packing Problem (BPP) [295], i.e., optimally placing items in limited containers to maximize space utilization, as shown in Figure 20 [Figure 20: see original paper]. However, industrial bin packing tasks face online decision-making challenges with unknown item arrival sequences [296], dynamic constraints requiring real-time response

References

[253] MA B, GAO X, WANG L, et al. Effect of current stability on surface formation of gmaw-based multi-layer single-pass additive deposition[J]. Journal of Mechanical Science and Technology, 2021, 35(6): 2449-2458.

[254] KERSHAW J, YU R, ZHANG Y, et al. Hybrid machine learning-enabled adaptive welding speed control[J]. Journal of Manufacturing Processes, 2021, 71: 374-383.

[255] WANG P, KERSHAW J, RUSSELL M, et al. Data-driven process characterization and adaptive control in robotic arc welding[J]. CIRP Annals, 2022, 71(1): 45-48.

[256] JIN Z, LI H, GAO H. An intelligent weld control strategy based on reinforcement learning approach[J]. The International Journal of Advanced Manufacturing Technology, 2019, 100: 1-15.

[257] MASINELLI G, LE-QUANG T, ZANOLI S, et al. Adaptive laser welding control: A reinforcement learning approach[J]. IEEE Access, 2020, 8: 103803-103814.

[258] CHENG J, WANG J. Data-driven matching method for processing parameters in process manufacturing[J]. Computer Integrated Manufacturing Systems, 2017, 23(11): 2361-2372.

[259] LIU L, ZHANG X, WAN X, et al. Digital twin-driven surface roughness prediction and process parameter adaptive optimization[J]. Advanced Engineering Informatics, 2022, 51: 101470.

[260] LI D, YANG J, ZHAO H, et al. Contact force plan and control of robotic grinding towards ensuring contour accuracy of curved surfaces[J]. International Journal of Mechanical Sciences, 2022, 227: 107449.

[261] LI D, YANG J, DING H. Process optimization of robotic grinding to guarantee material removal accuracy and surface quality simultaneously[J]. Journal of Manufacturing Science and Engineering, 2024, 146(5): 051005.

[262] WU Z, ZHANG G, DU W, et al. Torque control of bolt tightening process through adaptive-gain second-order sliding mode[J]. Measurement and Control, 2020, 53(7-8): 1131-1143.

[263] ZHANG H, WANG M, DENG W, et al. Semi-physical simulation optimization method for bolt tightening process based on reinforcement learning[J]. Machines, 2022, 10(8): 637.

[264] ZHOU Y, WANG X, ZHANG L. Research on assembly method of threaded fasteners based on visual and force information[J]. Processes, 2023, 11(6): 1770.

[265] SHTABEL N, SARAMUD M, TKACHEV S, et al. Automated machine vision control system for technological node assembly process[C]//AIP Conference Proceedings. 2024.

[266] YOU J, DU H, CHEN C C, et al. Disturbance observer-based finite-time control algorithm for robotic bolt-tightening via visual feedback[J]. IEEE Transactions on Automation Science and Engineering, 2024, 22: 7226-7237.

[267] PAN J, CHEN F, HAN D, et al. Adaptive process parameters decision-making in robotic grinding based on meta-reinforcement learning[J]. Journal of Manufacturing Processes, 2025, 137: 376-396.

[268] MANNE A S. On the job-shop scheduling problem[J]. Operations Research, 1960, 8(2): 219-223.

[269] BRUCKER P, JURISCH B, SIEVERS B. A branch and bound algorithm for the job-shop scheduling problem[J]. Discrete Applied Mathematics, 1994, 49(1-3): 107-127.

[270] ZHANG F, MEI Y, NGUYEN S, et al. Surrogate-assisted evolutionary multitask genetic programming for dynamic flexible job shop scheduling[J]. IEEE Transactions on Evolutionary Computation, 2021, 25(4): 651-665.

[271] SUN X, SHEN W, FAN J, et al. An improved non-dominated sorting genetic algorithm ii for distributed heterogeneous hybrid flow-shop scheduling

with blocking constraints[J]. Journal of Manufacturing Systems, 2024, 77: 990-1008.

[272] FONTES D B, HOMAYOUNI S M, GONÇALVES J F. A hybrid particle swarm optimization and simulated annealing algorithm for the job shop scheduling problem with transport resources[J]. European Journal of Operational Research, 2023, 306(3): 1140-1157.

[273] XIE J, LI X, GAO L, et al. A hybrid genetic tabu search algorithm for distributed flexible job shop scheduling problems[J]. Journal of Manufacturing Systems, 2023, 68: 1-18.

[274] ZHANG C, SONG W, CAO Z, et al. Learning to dispatch for job shop scheduling via deep reinforcement learning[C]//Advances in Neural Information Processing Systems. 2020.

[275] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks?[C]//International Conference on Learning Representations. 2019.

[276] ZHANG Y, ZHU H, TANG D, et al. Dynamic job shop scheduling based on deep reinforcement learning for multi-agent manufacturing systems[J]. Robotics and Computer-Integrated Manufacturing, 2022, 78: 102412.

[277] DESTOUET C, TLAHIG H, BETTAYEB B, et al. Flexible job shop scheduling problem under industry 5.0: A survey on human reintegration, environmental consideration and resilience improvement[J]. Journal of Manufacturing Systems, 2023, 67: 102-120.

[278] LI W, ZHANG S, DAI S, et al. Synchronized dual-arm rearrangement via cooperative mtsp[C]//IEEE International Conference on Robotics and Automation. 2024.

[279] ZHANG S, SHE Q, LI W, et al. Learning dual-arm object rearrangement for cartesian robots[C]//IEEE International Conference on Robotics and Automation. 2024.

[280] YAO Y, WANG C, LI X, et al. A knowledge-driven hybrid algorithm for solving the integrated production and transportation scheduling problem in job shop[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 26(2): 2707-2720.

[281] HUANG J, LI X, GAO L, et al. Automatic programming via large language models with population self-evolution for dynamic job shop scheduling problem[J]. arXiv preprint arXiv:2410.22657, 2024.

[282] GAVISH B, GRAVES S C. The travelling salesman problem and related problems[J]. 1978.

[283] TOTH P, VIGO D. The vehicle routing problem[M]. SIAM, 2002.

[284] APPLEGATE D, BIXBY R, CHVATAL V, et al. Concorde tsp solver[Z]. 2006.

[285] LLC Gurobi Optimization. Gurobi optimizer reference manual[M]. 2018.

[286] COOK W J, APPLEGATE D L, BIXBY R E, et al. The traveling salesman problem: A computational study[M]. Princeton University Press, 2011.

[287] HELSGAUN K. An extension of the lin-kernighan-helsgaun tsp solver for constrained traveling salesman and vehicle routing problems[J]. Roskilde: Roskilde University, 2017, 12: 966-977.

[288] CROES G A. A method for solving traveling-salesman problems[J]. Operations Research, 1958, 6(6): 791-812.

[289] MAHMUDY W F, WIDODO A W, HAIKAL A H. Challenges and opportunities for applying meta-heuristic methods in vehicle routing problems: A review[J]. Engineering Proceedings, 2024, 63(1): 12.

[290] XUE T, ZENG P, YU H. A reinforcement learning method for multi-agv scheduling in manufacturing[C]//IEEE International Conference on Industrial Technology. 2018.

[291] KOOL W, VAN HOOF H, WELLING M. Attention, learn to solve routing problems[C]//International Conference on Learning Representations. 2019.

[292] GAO C, SHANG H, XUE K, et al. Towards generalizable neural solvers for vehicle routing problems via ensemble with transferrable local policy[C]//International Joint Conference on Artificial Intelligence. 2024.

[293] SUN Z, YANG Y. DIFUSCO: graph-based diffusion solvers for combinatorial optimization[C]//Advances in Neural Information Processing Systems. 2023.

[294] ZHAO H, YU K, HUANG Y, et al. Disco: Efficient diffusion solver for large-scale combinatorial optimization problems[J]. arXiv preprint arXiv:2406.19705, 2024.

[295] MARTELLO S, PISINGER D, VIGO D. The three-dimensional bin packing problem[J]. Operations Research, 2000, 48(2): 256-267.

[296] SEIDEN S S. On the online bin packing problem[J]. Journal of the ACM, 2002, 49(5): 640-671.

[297] FAROE O, PISINGER D, ZACHARIASEN M. Guided local search for the three-dimensional bin-packing problem[J]. INFORMS Journal on Computing, 2003, 15(3): 267-283.

[298] SILVA J L D C, SOMA N Y, MACULAN N. A greedy search for the three-dimensional bin packing problem: The packing static stability case[J]. International Transactions in Operational Research, 2003, 10(2): 141-153.

[299] ZHAO H, SHE Q, ZHU C, et al. Online 3d bin packing with constrained deep reinforcement learning[C]//AAAI Conference on Artificial Intelligence. 2021.

[300] ZHAO H, ZHU C, XU X, et al. Learning practically feasible policies for online 3d bin packing[J]. Science China Information Sciences, 2022, 65(1): 112105.

[301] ZHAO H, YU Y, XU K. Learning efficient online 3d bin packing on packing configuration trees[C]//International Conference on Learning Representations. 2021.

[302] ZHAO H, XU J, YU K, et al. Deliberate planning of 3d bin packing on packing configuration trees[J]. arXiv preprint arXiv:2504.04421, 2025.

[303] ZHAO H, PAN Z, YU Y, et al. Learning physically realizable skills for online packing of general 3d shapes[J]. ACM Transactions on Graphics, 2023, 42(5): 1-21.

[304] HU R, XU J, CHEN B, et al. Tap-net: transport-and-pack using reinforcement learning[J]. ACM Transactions on Graphics, 2020, 39(6): 1-15.

[305] XU J, GONG M, ZHANG H, et al. Neural packing: from visual sensing to reinforcement learning[J]. ACM Transactions on Graphics, 2023, 42(6): 1-11.

[306] GROVE E F. Online bin packing with lookahead[C]//Annual ACM-SIAM Symposium on Discrete Algorithms. 1995.

[307] PUCHE A V, LEE S. Online 3D bin packing reinforcement learning solution with buffer[C]//IEEE International Conference on Intelligent Robots and Systems. 2022.

[308] CHOSET H, LYNCH K M, HUTCHINSON S, et al. Principles of robot motion: theory, algorithms, and implementations[M]. MIT press, 2005.

[309] HOF A L, GAZENDAM M, SINKE W. The condition for dynamic stability[J]. Journal of Biomechanics, 2005, 38(1): 1-8.

[310] YANG Z, YANG S, SONG S, et al. Packerbot: Variable-sized product packing with heuristic deep reinforcement learning[C]//International Conference on Intelligent Robots and Systems. 2021.

[311] DONG J, REN L. A digital twin modeling code generation framework based on large language model[C]//Annual Conference of the IEEE Industrial Electronics Society. 2024.

[312] LIU Q, ZHANG H, LENG J, et al. Digital twin-driven rapid individualised designing of automated flow-shop manufacturing system[J]. International Journal of Production Research, 2019, 57(12): 3903-3919.

[313] LENG J, ZHANG H, YAN D, et al. Digital twin-driven manufacturing cyber-physical system for parallel controlling of smart workshop[J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 10: 1155-1166.

[314] SHAO G. Manufacturing digital twin standards[C]//ACM International Conference on Model Driven Engineering Languages and Systems. 2024.

[315] NGUYEN T O, TABBONE S, BOUCHER A. A symbol spotting approach based on the vector model and a visual vocabulary[C]//International Conference on Document Analysis and Recognition. 2009.

[316] FAN Z, CHEN T, WANG P, et al. Cadtransformer: Panoptic symbol spotting transformer for cad drawings[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2022.

[317] MU J, YANG F, ZHANG Y, et al. Cadspotting: Robust panoptic symbol spotting on large-scale cad drawings[J]. arXiv preprint arXiv:2412.07377, 2024.

[318] WANG X, WANG L, WU H, et al. Parametric primitive analysis of cad sketches with vision transformer[J]. IEEE Transactions on Industrial Informatics, 2024, 20(10): 12041-12050.

[319] WU R, XIAO C, ZHENG C. Deepcad: A deep generative network for computer-aided design models[C]//IEEE International Conference on Computer Vision. 2021.

[320] SUN Z, ZHENG H, LYU C, et al. Neural rendering-based fast scene geometry modeling and retrieval method for digital twin assets[J]. Computer Integrated Manufacturing Systems, 2024, 30(4): 1189-1200.

[321] AGAPAKI E, BRILAKIS I. Geometric digital twinning of industrial facilities: Retrieval of industrial shapes[J]. arXiv preprint arXiv:2202.04834, 2022.

[322] LONG L, XIA Y, YANG M, et al. Retrieval of a 3d cad model of a transformer substation based on point cloud data[J]. Automation, 2022, 3(4): 563-578.

[323] JAOUA A, NEGRI E, JAOUA M, et al. Novel methods for teaching simulation: Strengthening digital twin development[C]//Winter Simulation Conference. 2024.

[324] MA L, YANG Z, YAN H, et al. Research on digital twin system driven by assembly action sequence database[J]. The International Journal of Advanced Manufacturing Technology, 2025: 1-16.

[325] MACÍAS A, MUÑOZ D, NAVARRO E, et al. Data fabric and digital twins: An integrated approach for data fusion design and evaluation of pervasive systems[J]. Information Fusion, 2024, 103: 102139.

[326] LEE S, LEE S, YANG Y. Occlusion-robust and efficient 6d pose estimation with scene-level segmentation refinement and 3d partial-to-6d full point cloud transformation[J]. Proceedings Copyright, 2024, 763: 771-780.

[327] WANG J, LUO L, LIANG W, et al. Oa-pose: Occlusion-aware monocular 6-dof object pose estimation under geometry alignment for robot manipulation[J]. Pattern Recognition, 2024, 154: 110576.

[328] QIN W, HU Q, ZHUANG Z, et al. Ippe-pcr: a novel 6d pose estimation method based on point cloud repair for textureless and occluded industrial

parts[J]. Journal of Intelligent Manufacturing, 2023, 34(6): 2797-2807.

[329] ZHUANG C, NIU W, WANG H. Sparse convolution-based 6d pose estimation for robotic bin-picking with point clouds[J]. Journal of Mechanisms and Robotics, 2024, 17(3): 031007.

[330] DING H, ZHAO L, YAN J, et al. Implementation of digital twin in actual production: intelligent assembly paradigm for large-scale industrial equipment[J]. Machines, 2023, 11(11): 1015.

[331] ZHU Z, XU X, ZHU J. Intelligent management and control of automatic loading and unloading system based on digital twin[C]//International Conference on Artificial Intelligence and Advanced Manufacture. 2021.

[332] YANG M, HUANG Z, SUN Y, et al. Digital twin driven measurement in robotic flexible printed circuit assembly[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.

[333] LIU Q, WAN J, ZHOU K. Cloud manufacturing service system for industrial-cluster-oriented application[J]. Journal of Internet Technology, 2014.

[334] Siemens AG. Siemens Xcelerator: Software for industry[Z]. 2023.

[335] NVIDIA Corporation. NVIDIA Omniverse: Platform for openusd and rtx rendering[Z]. 2025.

[336] YUE P, HU T, WEI Y, et al. A disturbance evaluation method for scheduling mechanisms in digital twin-based workshops[J]. The International Journal of Advanced Manufacturing Technology, 2024, 131(7): 4071-4088.

[337] WANG Y, WANG C, XU Y, et al. Digital twin task scheduling method for jobs of intelligent manufacturing unit under edge-cloud collaboration[J]. Journal of Mechanical Engineering, 2024, 60(6): 137-152.

[338] JIA Z, DONG J, LI S, et al. A digital twin system for predictive maintenance of complex equipment[C]//IEEE Smart World Congress. 2024.

[339] JIN S, YU F, WANG B, et al. Research on a real-time control system for discrete factories based on digital twin technology[J]. Applied Sciences, 2024, 14(10): 4076.

[340] XU C, TANG Z, YU H, et al. Digital twin-driven collaborative scheduling for heterogeneous task and edge-end resource via multi-agent deep reinforcement learning[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(10): 3056-3069.

[341] China News Service. From automotive paint inspection to shipbuilding: "kunwu platform" tackles industrial ai deployment challenges[Z]. 2025.

[342] HA D, SCHMIDHUBER J. World models[J]. arXiv preprint arXiv:1803.10122, 2018.

[343] HAFNER D, LILLICRAP T P, BA J, et al. Dream to control: Learning behaviors by latent imagination[C]//International Conference on Learning Representations. 2020.

[344] HANSEN N, SU H, WANG X. Temporal difference learning for model predictive control[C]//International Conference on Machine Learning. 2022.

[345] RAISSI M, PERDIKARIS P, KARNIADAKIS G E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations[J]. Journal of Computational Physics, 2019, 378: 686-707.

[346] WU P, ESCONTRELA A, HAFNER D, et al. Daydreamer: World models for physical robot learning[C]//Conference on Robot Learning. 2022.

[347] ZHOU G, PAN H, LECUN Y, et al. Dino-wm: World models on pre-trained visual features enable zero-shot planning[J]. arXiv preprint arXiv:2411.04983, 2024.

[348] HANSEN N, SU H, WANG X. TD-MPC2: scalable, robust world models for continuous control[C]//International Conference on Learning Representations. 2024.

[349] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems. 2020.

[350] MENDONCA R, BAHL S, PATHAK D. Structured world models from human videos[C]//Robotics: Science and Systems. 2023.

[351] YU T, THOMAS G, YU L, et al. Mopo: Model-based offline policy optimization[C]//Advances in Neural Information Processing Systems. 2020.

[352] RAFAILOV R, YU T, RAJESWARAN A, et al. Offline reinforcement learning from images with latent space models[C]//Learning for Dynamics and Control. 2021.

[353] LIAO S, XUE T, JEONG J, et al. Hybrid thermal modeling of additive manufacturing processes using physics-informed neural networks for temperature prediction and parameter identification[J]. Computational Mechanics, 2023, 72(3): 499-512.

[354] ZHU Q, LU Z, HU Y. A reality-augmented adaptive physics informed machine learning method for efficient heat transfer prediction in laser melting[J]. Journal of Manufacturing Processes, 2024, 124: 444-457.

[355] SHARMA R, GUO Y, RAISSI M, et al. Physics-informed machine learning of argon gas-driven melt pool dynamics[J]. Journal of Manufacturing Science and Engineering, 2024, 146(5): 051001.

[356] ZHU Q, LU Z, HU Y. Transfer learning-enhanced physics informed neural network for accurate melt pool prediction in laser melting[J]. Advanced Manufacturing, 2025, 2(1): 1-22.

[357] LUTTER M, RITTER C, PETERS J. Deep lagrangian networks: Using physics as model prior for deep learning[C]//International Conference on Learning Representations. 2019.

[358] HEIDEN E, MILLARD D, COUMANS E, et al. Neural-sim: Augmenting differentiable simulators with neural networks[C]//IEEE International Conference on Robotics and Automation. 2021.

[359] MURTHY J K, MACKLIN M, GOLEMO F, et al. gradsim: Differentiable simulation for system identification and visuomotor control[C]//International Conference on Learning Representations. 2021.

[360] DE AVILA BELBUTE-PERES F, SMITH K, ALLEN K, et al. End-to-end differentiable physics for learning and control[C]//Advances in Neural Information Processing Systems. 2018.

[361] CHEN R, ZHAO J, ZHANG F L, et al. Neural radiance fields for dynamic view synthesis using local temporal priors[C]//Computational Visual Media. 2024.

[362] JING X, YU T, HE R, et al. Frnerf: Fusion and regularization fields for dynamic view synthesis[C]//Computational Visual Media. 2024.

[363] YANG G W, LIU Z N, LI D Y, et al. Jnerf: An efficient heterogeneous nerf model zoo based on jittor[J]. Computational Visual Media, 2023, 9(2): 401-404.

[364] LI X, QIAO Y, CHEN P Y, et al. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification[C]//International Conference on Learning Representations. 2023.

[365] CAO J, GUAN S, GE Y, et al. Neuma: Neural material adaptor for visual grounding of intrinsic dynamics[C]//Advances in Neural Information Processing Systems. 2024.

[366] KANDUKURI R K, STRECKE M, STUECKLER J. Physics-based rigid body object tracking and friction filtering from rgb-d videos[C]//International Conference on 3D Vision. 2024.

[367] MEMMEL M, WAGENMAKER A, ZHU C, et al. ASID: Active exploration for system identification in robotic manipulation[C]//International Conference on Learning Representations. 2024.

[368] BAUMEISTER F, MACK L, STUECKLER J. Incremental few-shot adaptation for non-prehensile object manipulation using parallelizable physics simulators[C]//International Conference on Robotics and Automation. 2025.

[369] SONG C, BOULARIAS A. Learning to slide unknown objects with differentiable physics simulations[C]//Robotics: Science and Systems. 2020.

[370] HUANG B, YU Z, CHEN A, et al. 2d gaussian splatting for geometrically accurate radiance fields[C]//SIGGRAPH. 2024.

[371] MAYNE D Q, RAWLINGS J B, RAO C V, et al. Constrained model predictive control: Stability and optimality[J]. Automatica, 2000, 36(6): 789-814.

[372] WILLIAMS G, WAGENER N, GOLDFAIN B, et al. Information theoretic mpc for model-based reinforcement learning[C]//IEEE International Conference on Robotics and Automation. 2017.

[373] LUCIA S, KARG B. A deep learning-based approach to robust nonlinear model predictive control[J]. International Federation of Automatic Control, 2018, 51(20): 511-516.

[374] SANYAL S, ROY K. Ramp-net: A robust adaptive mpc for quadrotors via physics-informed neural network[C]//IEEE International Conference on Robotics and Automation. 2023: 1019-1025.

[375] HAFNER D, LILLICRAP T, FISCHER I, et al. Learning latent dynamics for planning from pixels[C]//International Conference on Machine Learning. 2019.

[376] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering atari, go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588(7839): 604-609.

[377] SUTTON R S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming[C]//Machine Learning Proceedings. 1990.

[378] UKASZ KAISER, BABAEIZADEH M, MIOS P, et al. Model based reinforcement learning for atari[C]//International Conference on Learning Representations. 2020.

[379] HAFNER D, LILLICRAP T, NOROUZI M, et al. Mastering atari with discrete world models[C]//International Conference on Learning Representations. 2021.

[380] HAFNER D, PASUKONIS J, BA J, et al. Mastering diverse domains through world models[J]. arXiv preprint arXiv:2301.04104, 2023.

[381] ZAMIELA C, STOKES R, TIAN W, et al. Physics-informed approximation of internal thermal history for surface deformation predictions in wire arc directed energy deposition[J]. Journal of Manufacturing Science and Engineering, 2024, 146(5): 051003.

[382] NARANG Y S, STOREY K, AKINOLA I, et al. Factory: Fast contact for robotic assembly[C]//Robotics: Science and Systems. 2022.

[383] TANG B, LIN M A, AKINOLA I, et al. Industreal: Transferring contact-rich assembly tasks from simulation to reality[C]//Robotics: Science and Systems. 2023.

[384] HOELLER D, RUDIN N, SAKO D, et al. Anymal parkour: Learning agile navigation for quadrupedal robots[J]. Science Robotics, 2024, 9(88): eadi7566.

[385] KIM H, OH H, PARK J, et al. High-speed control and navigation for quadrupedal robots on complex and discrete terrain[J]. Science Robotics, 2025, 10(102): eads6192.

[386] LEE J, BJELONIC M, RESKE A, et al. Learning robust autonomous navigation and locomotion for wheeled-legged robots[J]. Science Robotics, 2024, 9(89): eadi9641.

[387] KIM Y, OH H, LEE J, et al. Not only rewards but also constraints: Applications on legged robot locomotion[J]. IEEE Transactions on Robotics, 2024, 40: 2984-3003.

[388] RADOSAVOVIC I, XIAO T, ZHANG B, et al. Real-world humanoid locomotion with reinforcement learning[J]. Science Robotics, 2024, 9(89): eadi9579.

[389] LI Z, PENG X B, ABBEEL P, et al. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control[J]. The International Journal of Robotics Research, 2025, 44(5): 840-888.

[390] WANG H, WANG Z, REN J, et al. Beamdojo: Learning agile humanoid locomotion on sparse footholds[C]//Robotics: Science and Systems. 2025.

[391] ZHUANG Z, YAO S, ZHAO H. Humanoid parkour learning[C]//Conference on Robot Learning. 2024.

[392] XUE Y, DONG W, LIU M, et al. A unified and general humanoid whole-body controller for fine-grained locomotion[C]//Robotics: Science and Systems. 2025.

[393] LI J, CHENG X, HUANG T, et al. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control[C]//Robotics: Science and Systems. 2025.

[394] BEN Q, JIA F, ZENG J, et al. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit[C]//Robotics: Science and Systems. 2025.

[395] SHAO Y, HUANG X, ZHANG B, et al. Langwbc: Language-directed humanoid whole-body control via end-to-end learning[C]//Robotics: Science and Systems. 2025.

[396] LI J, ZHU Y, XIE Y, et al. OKAMI: Teaching humanoid robots manipulation skills through single video imitation[C]//Conference on Robot Learning. 2024.

[397] HE X, DONG R, CHEN Z, et al. Learning getting-up policies for real-world humanoid robots[C]//Robotics: Science and Systems. 2025.

[398] HUANG T, REN J, WANG H, et al. Learning humanoid standing-up control across diverse postures[C]//Robotics: Science and Systems. 2025.

[399] HE T, GAO J, XIAO W, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills[C]//Robotics: Science and Systems. 2025.

[400] SEO Y, SFERRAZZA C, GENG H, et al. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control[J]. arXiv preprint arXiv:2505.22642, 2025.

[401] SFERRAZZA C, HUANG D M, LIN X, et al. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation[C]//Robotics: Science and Systems. 2024.

[402] CHI Y, LIAO Q, LONG J, et al. Demonstrating berkeley humanoid lite: An open-source, accessible, and customizable 3d-printed humanoid robot[C]//Robotics: Science and Systems. 2024.

[403] INC. T. Tesla optimus: Humanoid robot progress update[Z]. 2024.

[404] REUTERS. Figure signs deal with bmw to deploy humanoid robots at us plant[Z]. 2024.

[405] VENTUREBEAT. Sanctuary ai launches phoenix, a humanoid general-purpose robot[Z]. 2024.

[406] TechNode. Ubtech's industrial humanoid robot walker s undergoes training on nio' s ev assembly line[Z]. 2024.

[407] EqualOcean. Four promising application scenarios for humanoid robots: Deployment of yuanzheng a1 in 3c assembly, automotive chassis, and final inspection lines[Z]. 2024.

[408] EET China. Nio factory testing china' s first humanoid robot "kuafu" equipped with harmonyos[Z]. 2024.

Author Biography

Xu Kai (born 1982), male, Ph.D., Professor, Doctoral Supervisor. Research interests: computer graphics, 3D vision, embodied intelligence, digital twins, etc.

Note: Figure translations are in progress. See original paper for figures.