Contents lists available at SciVerse ScienceDirect

CIRP Annals Manufacturing Technology

Journal homepage: www.elsevier.com/locate/cirp

Sim2Know: new paradigm of digital twins to design and inform human-centric knowledge system

Bingbing Li[a], Haolin Fan[a], Zhen Fan[a], John Ahmet Erkoyuncu (2)[b], Hong-Chao Zhang (1)[c,*]

[a] *Autonomy Research Center for STEAHM (ARCS), California State University Northridge, Los Angeles, United States*
[b] *Faculty of Engineering and Applied Sciences, Cranfield University, Cranfield, United Kingdom*
[c] *School of Mechanical Engineering, Dalian University of Technology, Dalian, China*

Abstract: The novel framework, Sim2Know, tackles two major challenges of designing and informing the human-centric knowledge system adaptively: the lack of labeled real-world training data and the difficulty of capturing implicit knowledge. First, a digital twin demonstrator is built to produce high-quality synthetic training data. Next, we create a hybrid training approach that merges transfer learning from pre-trained self-supervised models with synthetic data augmentation, achieving a precision rate of 90.31% in identifying 11 essential human action patterns in metal additive manufacturing. Finally, we design the human-centric knowledge system to capture the implicit knowledge through contextualizing human machine interaction beyond the domain knowledge.

Keywords: Digital twin, Artificial intelligence, Human-centric knowledge

1. Introduction

In the evolution towards Industry 5.0, manufacturing paradigms are undergoing a co-evolution, integrating data-driven intelligence with human-centered approaches. [1]. This transition prioritizes the "human in the loop" approach which will benefit both non-expert and expert training [2]. However, one significant challenge in developing high-fidelity mixed reality (MR) training for manufacturing lies in creating accurate and robust adaptive knowledge systems. These systems must effectively capture both explicit and implicit knowledge using advanced artificial intelligence (AI) techniques, particularly through large multimodal models (LMMs) and physics-informed machine learning [3]. However, implicit knowledge is very difficult to capture and articulate through an experienced human, while explicit knowledge can be easily expressed through knowledge graphs (KGs) [4]. To address the challenge, a human-centric knowledge informed approach offers promising solutions by providing expert guidance for applications such as process control [5], novice training [6], safety and quality assurance [7], and to expand human skills in MR training.

Within the evolving landscape of a human-centric knowledge informed approach, human action recognition in task has emerged as a promising method to capture implicit knowledge through cross-modal collaborative learning [8]. This approach combines advanced sensor systems with machine learning algorithms to accurately interpret human movements, enabling more natural, productive and efficient human-machine interaction (HMI) [9]. In manufacturing environments, action recognition systems have demonstrated significant benefits across multiple applications. For example, Mehta et al. [10] developed systems for detecting worker deviations from standard operating procedures, enhancing quality control and safety. Additionally, Yan et al. [11] implemented automated action recognition in standardized manufacturing workshops to verify operational compliance and timing accuracy.

The new paradigm of digital twin (DT) goes beyond simple virtual modeling to a more dynamic and interactive representation of physical systems, which could offer the service to collect data, evaluate, fill gaps, and share it with relevant models [12]. To leverage action recognition systems to inform human-centric knowledge system, the first major challenge involves dataset development and model training. Traditional approaches rely heavily on real-world data collection, which is both time-consuming and resource intensive. Moreover, the supervised learning paradigm requires extensive manual annotation, limiting model generalization and necessitating retraining for new applications. The second challenge concerns the implicit knowledge extraction and its integration with explicit knowledge. Currently the explicit knowledge in MR training heavily relies on the limited operation documentations such as user manuals, tutorials, and guides, while most of the implicit knowledge can only be captured through human practices and tasks, limiting their ability to provide contextualized guidance and adapt to diverse operational scenarios.

To address these two challenges, we propose an innovative framework, Sim2Know, to design and inform the human-centric knowledge system adaptively by bridging simulation and real-world insights through three key approaches: DT system for synthetic data generation, pre-trained self-supervised models for action recognition, and multimodal KGs to capture both explicit and implicit knowledge. Using metal additive manufacturing (AM) as a case study, we investigate the feasibility of DT-based synthetic data generation for training action recognition models, develop an efficient hybrid training paradigm combining pre-trained self-supervised learning models with synthetic data, and demonstrate how KG integration provides context-aware expert guidance in metal AM processes. This comprehensive approach significantly reduces dependency on real-world datasets, while maintaining high recognition accuracy, establishing a framework adaptable to broader manufacturing applications.

2. Methodology

As illustrated in **Figure 1**, the Sim2Know framework proposed in this paper includes three stages: dataset construction process, action recognition model training, and human-centric knowledge system. In the dataset construction stage, synthetic data is collected through interactive DT, combined with practical data collected from real operations. In the model training stage, preprocessed data from synthetic motions are used to train the action recognition model. During this training process, the

pretrained model is frozen and used to extract motion features from the input synthetic data. The predicted action will be considered as the input of the KG to provide more contextual information for decision-support during the practical operations. This further explores the action recognition results, provides more domain-knowledge, and facilitates HMI.

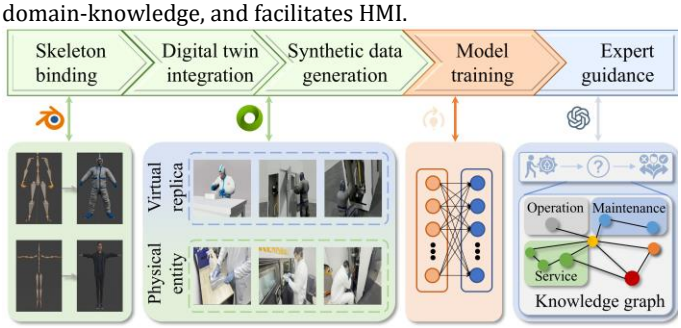


Figure 1. An overview of the proposed Sim2Know framework. Three key contributions include synthetic data generation though digital twins (DT), action recognition training and in-depth knowledge providence through a knowledge graph (KG).

2.1. Dataset construction process

The dataset construction process is divided into three steps: skeleton-action binding, action simulation through DT integration, and synthetic data creation.

In the first step, skeleton data are extracted from human motion videos and imported into Blender for precise binding to virtual character models. This ensures that the virtual model accurately replicates human actions.

The second step focuses on developing an interactive DT environment for metal AM processes using NVIDIA's Omniverse platform. We created a high-fidelity virtual representation by integrating a laser powder bed fusion machine Renishaw AM400 model with an interactive virtual operator, enabling comprehensive simulation of real-world manufacturing operations. The DT leverages GPU-accelerated simulation to enable rapid validation and synthetic data generation. In process validation, our DT environment simulates complex safety protocols and system checks that traditionally require manual verification. By running these pre-check procedures in parallel through the GPU-accelerated simulation, we reduce the time from 20 minutes to approximately 1 minute per operation. Similarly, our DT-based synthetic data generation pipeline circumvents the physical constraints and safety considerations that typically slow down real-world operational data collection. This virtual environment allows us to compress what would normally be a 30-hour data collection process into just 3 hours, achieving a 90% reduction in time while maintaining data quality standards.

To enhance dataset diversity, the simulation parameters are systematically varied. Camera configurations are randomized across three dimensions: focal length, position, and viewing angles. These variations ensure comprehensive coverage of operational viewpoints while maintaining clear visibility of critical actions. The simulations capture both first-person and third-person perspectives, with each synthetic video sequence lasting 2-3 seconds at 60 frames per second.

Real-world videos are also collected to complement the synthetic data, creating a comprehensive dataset that includes 11 key AM operations, indexed from 0 to 10: clean build plate (0), clean recoater (1), insert build plate screw (2), measure build plate thickness (3), open process chamber (4), operate valves (5), build plate installation (6), recoater installation (7), press reset button (8), press emergency stop button (9), and wipe laser window (10). These 11 actions were selected as they represent key tasks in the metal AM workflow, including preparation, operation, and maintenance, which can influence process efficiency, machine reliability, and print quality. The complete dataset comprises 4,400 synthetic videos supplemented by 560 real-operation videos, providing a balanced distribution of each operation class. An image dataset is also derived by extracting the central frame from each video. This dataset construction pipeline leverages DT to efficiently generate high-quality, diverse synthetic data. It effectively addresses the challenges of acquiring large-scale real-world datasets while maintaining operational fidelity and visual diversity.

2.2. Action recognition model training

The model training consists of two phases: using the pretrained V-JEPA model [13] for feature extraction and training downstream classifiers for action recognition, illustrated in **Figure 2**. V-JEPA employs a self-supervised learning framework that learns visual representations from vast amounts of videos. Its architecture consists of an encoder $E_\theta(\cdot)$ that maps input sequences x and y into feature representations s_x and s_y, and a predictor $P_\phi(\cdot)$ that estimates s_y from s_x, guided by the spatio-temporal transformation Δy. The training objective is defined as:

$$\min_{\theta,\phi} \| P_\phi(E_\theta(x), \Delta y) - \mathrm{sg}\,(E_\theta(y)) \|_{\ell_1}$$

where $\mathrm{sg}\,(\cdot)$ is a stop-gradient operation to prevent trivial solutions and ensure stable learning, and $\|\cdot\|_{\ell_1}$ represents the L1 norm used for computing the prediction loss.

During the pretraining phase, V-JEPA masks large spatiotemporal blocks of video data to create (x, y) pairs. This approach forces the model to learn robust, context-aware feature representations rather than focusing on low-level pixel similarities. The architecture implements vision transformers (ViTs) for encoding and prediction tasks in latent space, enabling the extraction of hierarchical visual representations that capture spatial and temporal dependencies.

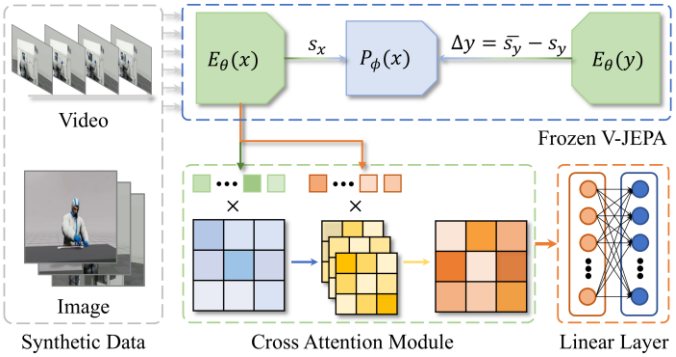


Figure 2. Training paradigm of the proposed method. The frozen V-JEPA model is used to extract features from the synthetic dataset, and the final classification is performed through a linear projection layer.

We implemented an attention-based classifier for downstream action recognition that builds upon the frozen pretrained V-JEPA weights, as depicted in [13]. The classifier has two main components: an attentive pooler and a linear classification head. The attentive pooler employs learnable query tokens and cross-attention mechanisms to aggregate spatiotemporal information from V-JEPA's feature representations. The architecture follows a transformer-based design with specific configurations (embedding dimension = 768, number of attention heads = 12, MLP ratio = 4.0) and incorporates weight rescaling to ensure stable gradient propagation during training. This design efficiently processes V-JEPA's pre-trained representations while maintaining

computational feasibility for video and image-based action recognition tasks.

2.3. Human-centric knowledge system

The action recognition system interfaces with a structured and multi-modal KG developed in our prior research to provide contextually relevant user guidance. This KG is a comprehensive repository of domain-specific knowledge, systematically organizing information extracted from machine user guides, including operational procedures, equipment specifications, and maintenance protocols. Furthermore, the KG is enhanced with a visual guidance component that retrieves implicit knowledge and presents relevant instructional images corresponding to specific operations and procedures.

The knowledge retrieval process implements a two-stage approach to generating user guidance. The system employs a semantic mapping algorithm in the first stage to transform recognized actions into natural language queries. These queries are then processed using a retrieval augmented generation (RAG) technique to extract relevant information from the KG entities and the original user guides. In the second stage, the system utilizes GPT-4 to process the retrieved content, incorporating domain-specific knowledge to generate coherent and contextually appropriate responses.

The system's capability can be illustrated through a practical example: when it recognizes an action such as "build plate installation in the AM400 printer process chamber," it automatically generates contextually relevant queries about procedural details, safety precautions, and operational requirements. These queries extract comprehensive information from the KG, ensuring that users receive relevant expert guidance for their specific operational context. The generated questions can range from basic operational inquiries to specific safety-critical considerations, providing a systematic approach to accessing context-aware expert guidance.

3. Experimental evaluation

3.1. Action recognition model

We comprehensively evaluated image- and video-based action classifiers using varying real and synthetic training data proportions. The real data ratio was systematically adjusted from 5% to 30% in 5% increments, with the remaining data allocated to the test set. This design led to six unique training configurations, enabling us to evaluate the effects of real data scarcity and the effectiveness of synthetic data in improving model performance. All images and videos underwent standardized preprocessing at two resolution settings: 384^2 and 224^2 pixels. Model performance was evaluated using precision, recall, and F1-score metrics against a consistent test set of real data.

The experimental results, detailed in **Table 1**, demonstrate that video-based models perform better than image-based approaches. The highest action recognition precision of 90.31% was achieved using a training dataset comprising 30% real video data supplemented with synthetic data at 384^2 resolution, followed closely by 89.82% at 224^2 resolution. Notably, even with only 5% real video data (28 videos), the model achieves a relatively satisfactory precision of 68.11% at 224^2 resolution, despite the complex backgrounds in manufacturing scenes. This result validates the effectiveness of synthetic data in compensating for limited real-world data availability.

Table 1 Precision results under two resolution settings, comparing training performance using video and image data with varying ratios of real data in the training dataset. (The best-performing results are highlighted in bold.)

Real data ratio	384^2		224^2	
	Video	Image	Video	Image
5%	65.41%	41.65%	68.11%	46.34%
10%	74.85%	55.84%	73.66%	56.04%
15%	76.31%	62.05%	79.04%	62.89%
20%	84.60%	63.70%	77.95%	66.59%
25%	87.83%	70.78%	87.62%	**71.02%**
30%	**90.31%**	**72.26%**	**89.82%**	69.97%

Our analysis reveals two key findings regarding resolution impact and modality choice. First, increasing resolution does not consistently improve performance, suggesting that while higher resolutions provide more granular information, they may compromise the model's ability to capture global action patterns. Second, video-based models consistently outperform image-based approaches across all configurations, which is attributable to video's inherent capability to capture spatiotemporal information that better aligns with the sequential nature of actions.

Figure 3 presents the confusion matrices for the four best-performing configurations, where the diagonal values represent the proportion of correctly classified instances for each class, and the off-diagonal values indicate the proportion of misclassifications. The matrices reveal consistent misclassification errors between physically proximate actions, particularly in the confusion between pressing the reset button (class 8 with an average recall of 0.72) and the emergency stop button (class 9 with an average recall of 0.56). This pattern suggests limitations in the model's ability to distinguish actions performed on closely positioned interface elements. The image-based models at both resolutions exhibit significant confusion between cleaning the build plate (class 0 with an average recall of 0.68) and cleaning the recoater (class 1 with an average recall of 0.72) actions. This misclassification is less prominent in video-based models, indicating that static images alone cannot capture the distinctive characteristics of these cleaning actions. In these cases, the superior performance of video-based models demonstrates the importance of temporal information in disambiguating visually similar metal AM operations.

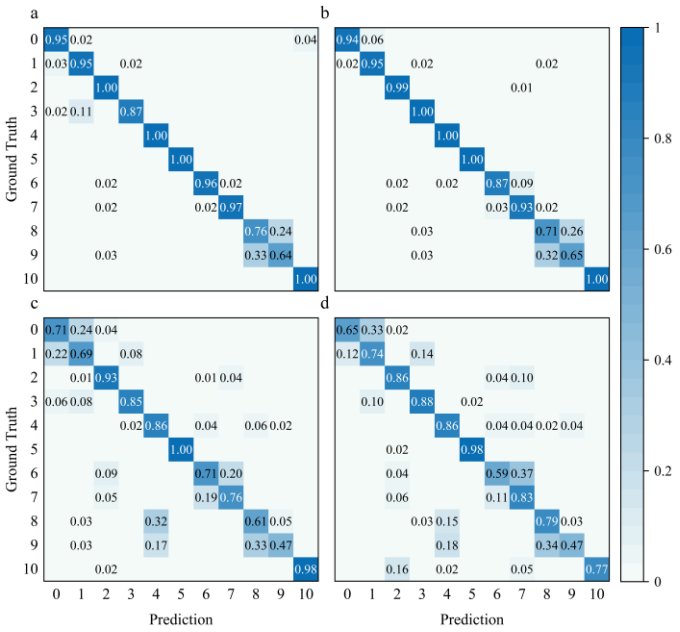


Figure 3. Confusion matrices showing action recognition performance across different training configurations (index from 0~10 refers to 11 actions). (a) video dataset at 384^2, (b) video dataset at 224^2, (c) image dataset at 384^2, and (d) image dataset at 224^2.

To establish a baseline and assess synthetic data's impact, we conducted parallel training experiments using limited real data (5% to 10% in 5% increments), without supplementing it with synthetic data. The F1 score was the primary evaluation metric for quantifying synthetic data's contribution to model performance.

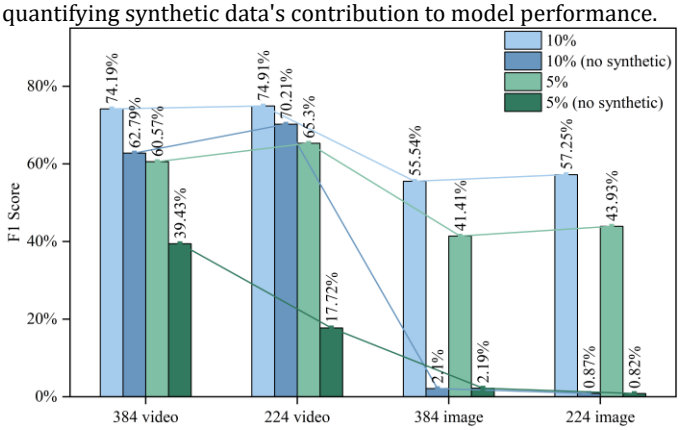


Figure 4. Performance comparison (F1 scores) across training configurations with varying real data ratios (5%-10%) with and without the integration of synthetic data.

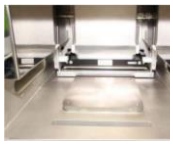
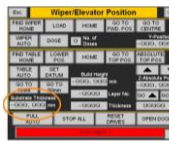
As depicted in **Figure 4**, synthetic data significantly enhances action recognition performance across all experimental configurations. The most substantial improvement is observed in image-based recognition, where synthetic data augmentation increases the F1 score from 2.1% to 55.54% at 384^2 resolution with 10% real data (a 53.44% improvement), compared to the video modality's improvement from 62.79% to 74.19% (an 11.4% increase). Video modality consistently outperforms image-based recognition, achieving a peak F1 score of 74.91% at 224^2 resolution compared to 57.25% for image modality, demonstrating video's superior capability in capturing temporal dynamics of operational procedures. Notably, models trained with 10% real data plus synthetic data consistently outperform those trained with 5% real data alone across both modalities and resolutions, indicating synthetic data's effectiveness in compensating for limited real training data. The consistent performance improvement across various resolutions demonstrates that synthetic data augmentation effectively enhances action recognition in low-data industrial environments, particularly in metal AM, leading to more reliable knowledge-based guidance and operational support.

3.2. Knowledge graphs for expert guidance

The system utilizes KGs to transform recognized actions into comprehensive expert guidance through question generation. The KG maps interconnections between various entities, enabling more contextually relevant responses. Powered by GPT-4 (or customizable alternatives), the system generates in-depth questions covering multiple aspects of each action, including procedural steps, equipment requirements, safety precautions, and maintenance guidelines. For instance, when the recognized action is "install the build plate," it generates contextually relevant questions such as "Is there a specific orientation required for the build plate?" and "What software settings need to be configured after installing the build plate?"

The system generates responses through a multi-step process: first, it embeds text and images into high-dimensional vector representations, then retrieves contextually relevant information using cosine similarity measurements. This retrieved information serves as a grounding context for GPT-4's response generation. Using vector-based retrieval, the system produces comprehensive answers incorporating precise textual explanations and relevant reference images, ensuring thorough coverage of all critical task aspects while maintaining semantic coherence. Implicit expert knowledge can be integrated into KGs through demonstrations or interactions, structuring their guidance into new entities for better knowledge representation. **Table 2** showcases qualitative results highlighting the system's ability to deliver expert-level guidance using the domain-specific KG. It also demonstrates an adaptive approach to incorporating an expert's implicit knowledge as input.

Table 2. An intuitive illustration of the usage of the multimodal knowledge graph (KG) to provide in-depth guidance and integrate implicit knowledge from human experts as new nodes within the KG.

Prediction	Question	Answer (The average response time is 5s)	
build plate installation	Is a specific build plate orientation required?	The build plate can be installed in four positions, but a shallow blind hole marks the front for correct orientation, ensuring proper installation to prevent damage to the AM400 system's z-axis.	
measure build plate thickness	Why is measuring build plate thickness required?	This operation ensures the correct thickness value is input into the control panel before final fixing. An incorrect thickness value can damage the system or cause process inconsistency.	
clean build plate	What concentration is ideal for cleaning the build plate?	(By expert) The build plate should be cleaned with 99% isopropyl alcohol. This concentration provides optimal cleaning effectiveness while maintaining quick evaporation and minimal residue.	

4. Conclusion

Our investigation yielded three key findings to better design and inform the human-centric knowledge system. First, our synthetic data generated from DT demonstrated superior performance in

training action recognition models compared to conventional real-world datasets, achieving a 90% reduction in data collection time, while maintaining comparable or better performance metrics. Second, our novel hybrid training paradigm, which synthesizes pre-trained self-supervised learning through V-JEPA with synthetic data, achieved remarkable precision rates of 90.31% and 89.82% for video-based and image-based action recognition, respectively, across 11 critical AM operations, despite limited real-world training data. Third, integrating our multi-modal KGs enhanced the framework's capability by providing context-aware expert guidance, adaptively capturing the implicit knowledge from performing tasks by the experts or skilled operators.

These findings demonstrate the framework's effectiveness in addressing fundamental challenges in dataset construction and human-centric knowledge system design and highlight its potential for broader applications. The successful implementation in the case study of metal AM, particularly in reducing the dependency on extensive real-world data collection, suggests promising applications across various advanced manufacturing domains where data acquisition is costly, hazardous, or impractical. Future work should focus on extending this framework to other manufacturing processes and investigating its adaptability to emerging manufacturing systems through human-centric knowledge.

Acknowledgement

This work is mainly funded by the MUREP High Volume project (80NSSC22M0132) through the U.S. NASA's Office of STEM Engagement, and the SMART IAC project (DE-EE0009726) through the U.S. Department of Energy's Office of Manufacturing and Energy Supply Chains.

References

[1] Li, X., Nassehi, A., Wang, B., Hu, S.J., Epureanu, B.I., 2023, Human-centric manufacturing for human-system coevolution in Industry 5.0, CIRP Annals, 72/1: 393-396.

[2] Simeone, A., Fan, Y., Antonelli, D., Catalano, A.R., Priarone, P.C., Settineri, L., 2024, Inclusive manufacturing: A contribution to assembly processes with human-machine reciprocal learning, CIRP Annals, 73/1: 5-8.

[3] Fan, H., Liu, C., Bian, S., Ma, C., Huang, J., Liu, X., Doyle, M., Lu, T., Chow, E., Chen, L., Fuh, J.Y.H., Lu, W.F., Li, B., 2025, New Era Towards Autonomous Additive Manufacturing: A Review of Recent Trends and Future Perspectives, International Journal of Extreme Manufacturing, 7/3: 032006.

[4] Liu, A., Zhang, D., Wang, Y., Xu, X., 2022, Knowledge graph with machine learning for product design, CIRP Annals, 71/1: 117-120.

[5] Fan, H., Liu, X., Fuh, J.Y.H., Lu, W.F., Li, B., 2025, Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics, Journal of Intelligent Manufacturing, 36/2: 1141–1157.

[6] Fan, H., Zhang, H., Ma, C., Wu, T., Fuh, J.Y.H., Li, B., 2024, Enhancing metal additive manufacturing training with the advanced vision language model: a pathway to immersive augmented reality training for non-experts, Journal of Manufacturing Systems, 75: 257-269.

[7] Zhong, Y., Karthikeyan, A., Pagilla, P., Mehta, R.K., Bukkapatnam, S.T.S., 2024, Human-centric integrated safety and quality assurance in collaborative robotic manufacturing systems, CIRP Annals, 73/1: 345-348.

[8] Kong, W., Liu, J., Hong, Y., Li, H., Shen, J., 2024, Cross-modal collaborative feature representation via Transformer-based multimodal mixers for RGB-T crowd counting, Expert Systems with Applications, 255/A: 124483.

[9] Lin, X., Xu, L., Zhuang, S., Wang, Q. 2023, VW-SC3D: a sparse 3D CNN-based spatial–temporal network with view weighting for skeleton-based action recognition, Electronics, 12/1: 117.

[10] Mehta, N.K., Prasad, S.S., Saurav, S., Saini, R., Singh, S., 2024, IAR-Net: a human-object context guided action recognition network for industrial environment monitoring, IEEE Transactions on Instrumentation and Measurement, 73: 1-8.

[11] Yan, J., Wang, Z., 2022, YOLO V3 + VGG16-based automatic operations monitoring and analysis in a manufacturing workshop under Industry 4.0, Journal of Manufacturing Systems, 63: 134-142.

[12] Erkoyuncu, J.A., del Amo, I.F., Ariansyah, D., Bulka, D., Vrabic, R., Roy, R., 2020, A design framework for adaptive digital twins, CIRP Annals, 69/1: 145-148.

[13] Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N., 2024, Revisiting feature prediction for learning visual representations from video. arXiv:2404.08471.