This is the Pre-Published Version.

The following publication S. Li, P. Zheng, J. Fan and L. Wang, "Toward Proactive Human–Robot Collaborative Assembly: A Multimodal Transfer-Learning-Enabled Action Prediction Approach," in IEEE Transactions on Industrial Electronics, vol. 69, no. 8, pp. 8579-8588, Aug. 2022 is available at https://dx.doi.org/10.1109/TIE.2021.3105977.

Towards Proactive Human-Robot Collaborative Assembly: A Multimodal Transfer Learning-Enabled Action Prediction Approach

Shufei Li, *Student Member*, Pai Zheng*, *Member*, Junming Fan, *Student Member*, and Lihui Wang

***Abstract*—Human-robot collaborative assembly (HRCA) is vital for achieving high-level flexible automation for mass personalization in today's smart factories. However, existing works in both industry and academia mainly focus on the adaptive robot planning, while seldom consider human operator's intentions in advance. Hence, it hinders the HRCA transition towards a proactive manner. To overcome the bottleneck, this research proposes a multimodal transfer learning-enabled action prediction approach, serving as the prerequisite to ensure the Proactive HRCA. Firstly, a multimodal intelligence-based action recognition approach is proposed to predict ongoing human actions by leveraging the visual stream and skeleton stream with short-time input frames. Secondly, a transfer learning-enabled model is adapted to transfer learnt knowledge from daily activities to industrial assembly operations rapidly for online operator intention analysis. Thirdly, a dynamic decision-making mechanism including robotic decision and motion control is described to allow mobile robots to assist operators in a proactive manner. Lastly, an aircraft bracket assembly task is demonstrated in the lab environment, and the comparative study result shows that the proposed approach outperforms other state-of-the-art ones for efficient action prediction.**

***Index Terms*—Human-robot collaboration, multimodal intelligence, action recognition, transfer learning.**

I. Introduction

In today's industrial transformation towards smart manufacturing, modern factories are striving for an ever higher degree of flexible production with mass efficiency, which is known as mass personalization [1]. To achieve it, human-robot collaboration becomes a prevailing strategy, which combines high accuracy, strength, and repeatability of robots with high flexibility and adaptability of human operators to realize optimal overall productivity [2]. In this paradigm, human-robot collaborative assembly (HRCA) has been adopted, which allows seamless communication and cooperation between human operators and robots to fulfill manufacturing assembly tasks readily. By monitoring changes in a shared workspace, a collaborative robot can reactively update its motion to ensure the safety of operators in a sophisticated assembly side-by-side [3]. Nevertheless, existing HRCA systems normally co-work in a pre-defined or adaptive manner by fixed robot arms, which lack certain intelligence and flexibility in collaborations. Motivated by this, it is naturally for us to consider a new paradigm of Proactive HRCA, where mobile/fixed collaborative robots can understand "what will the worker do next?" [4] and act subsequent plannings to assist workers in a proactive manner.

To realize it, human action recognition, as the prerequisite of robotic dynamic decision-making, is of paramount importance. One of the prevailing research areas is to recognize human actions from RGB videos [5]. For instance, a 3D convolutional neural network (3D CNN) was introduced to recognize operators' actions during the assembly of a visual controller [6]. Another main pillar aims to deriving action representations based on the topology of human body, i.e., skeleton joints [7]. Accordingly, in an HRCA system, assembly contextual information can be inferred from key points of the human body [8]. However, the aforementioned efforts can only analyze human action intentions from one single modality in industrial settings, either by 1) RGB video-based action recognition or 2) skeleton-based one. The former is effective in capturing fine-grained details, but too rigid to associate visual patterns of one same action from different views, let alone the high computing cost for long timespan videos. For the latter one, skeleton-based action recognition is computationally efficient, but lacks much low-level detailed information [9]. Hence, in real workplace settings towards proactive HRCA, single modality based methods fail to high-reliably predict operators' on-going operations due to the following challenges: 1) subtle and similar motion patterns of human action in a short timespan, such as wedging pins and screwing bolts while installing a mechanical part; 2) diverse visual patterns of the same action caused by different camera views; and 3) insufficient labeled data of assembly operations in varying time lengths.

Aiming to fill this research gap towards the Proactive HRCA, a multimodal intelligence-based transfer learning-

Manuscript received May 26, 2021; major revised April 19, 2021; minor revised July 05, 2021; accepted August 7, 2021. This research work was partially funded by the Laboratory for Artificial Intelligence in Design (Project Code: RP2-1), Hong Kong Special Administrative Region, and Research Committee of The Hong Kong Polytechnic University under Departmental General Research Fund (G-UAHH). This research project has also been approved by the Human Subjects Ethics Sub-committee (HSESC), at the Hong Kong Polytechnic University (No. HSEARS20201110002).

S. F. Li, J. M. Fan and P. Zheng are with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, 999077, HKSAR (e-mail: shufei.li@connect.polyu.hk, junming.fan@connect.polyu.hk. *Corresponding author phone: +852-2766-5633, email: pai.zheng@polyu.edu.hk).

L.H. Wang is with the Department of Production Engineering, KTH Royal Institute of Technology, Sweden (e-mail: lihuiw@kth.se).

enabled human action prediction approach is proposed, which allows higher cooperation flexibility and efficiency between operators and fixed/mobile robots in a global workspace. The rest of this paper is organized as follows. Section II reviews recent related works in the HRCA system field. Aiming towards Proactive HRCA, a three-step based approach is proposed in Section III, which is achieved by multimodal action pattern extraction and transfer learning-based industrial deployment. Section IV further conducts a typical Proactive HRCA task for aircraft bracket assembly with comparative studies. At last, conclusions and future works are summarized in SectionV.

II. Related work

In this section, the recent development of HRCA is firstly recapped, followed by a comprehensive review of advanced human recognition methodologies, to reveal the research gaps and motivations towards Proactive HRCA.

A. Human-robot collaborative assembly

Instead of the strictly separated tasks between pre-programmed industrial robots and shop-floor operators, HRCA enables collaborative robots to assist human operators side by side in the shared assembly works [10]. For example, Wang et al. [11] reviewed human-robot collaborative systems for welding process, where robots can dynamically modify their pre-planned tasks to collaborate with human welders. For seamless cooperation, collaborative robots in current HRCA systems interact with human operators via user-friendly approaches, such as Augmented Reality (AR)-based interface [12], intuitive human body pose [13] or hand gestures [14]. Nevertheless, these collaborative robots are normally fixed in one monitored zone and controlled by reactive instructions, thus far from efficient integration of robotic automation and human cognition.

To overcome this challenge, recent works began to allow a mobile robot to learn about operators' next intention and conduct assistant planning in advance. In this context, a high-level teamwork intelligence of the Proactive HRCA is enabled by two preconditions, including contextual awareness of the industrial scenario and dynamic robot decision-making in the execution loop. Context-awareness perception ensures basic characteristics of HRCA, such as active collision avoidance and coordinated task allocation [15]. For instance, a recurrent neural network (RNN) was utilized to predict a human operator's future motion trajectory for proactive assistance, while avoiding the collision [16]. Meanwhile, dynamic decision-making allows mobile robots to make proactive assembly planning for common goals in a holistic view. In task-oriented HRCA, based on observable human actions, robots can infer semantic knowledge of operators' execution and conduct reasonable assistant plans [17]. For both aspects, accurate human action recognition and intention analysis are critical.

B. Industrial operation recognition and intention analysis

Industrial operator's actions represent operation intention during production, of which the analysis procedure contains both offline recognition during the model training process, and online prediction in real implementations. Hence, accurate online action prediction serves as the prerequisite for Proactive HRCA, which allows robots to reason about operators' intentions and assist them in a long-term adaptive or even proactive manner [16]. As analysed in the comprehensive survey of human action recognition methods [18], current deep learning techniques [19] can directly predict human activities and human-object interaction from RGB videos, depth, and skeleton data.

In today's HRCA field, ongoing activity prediction [20] and 3D action estimation [21] are of particular interests in the human operation recognition tasks. The former one allows robots to rapidly response and assistance during cooperation. In this field, a two-layer graph network was utilized to explore human-object interaction in a video scene, so as to predict human activities along the time as early as possible [22]. Meanwhile, the latter one extracts features of 3D human activities [23] and builds a bridge between the accurateness of robotic control and strict safety requirements for collision avoidance. Nevertheless, scarcely any work considers action patterns from more than one modality in HRCA, which can support ongoing intention prediction with fewer frames.

C. Multimodal intelligence and knowledge transferring

Multimodal perception increases the utilization of digital resources [24] and creates new information capital towards Proactive HRCA [10]. With expanded multimodal information, e.g., visual and kinematic knowledge, it is realistic to recognize ongoing human actions ahead of schedule. Meanwhile, some other works provide a procedure of a multimodal intelligence-based network, which can learn human action patterns from skeleton sequences and RGB videos [25].

Another critical issue lies in a lack of annotations among the captured data of operators' assembly motion in real industrial settings. These data also suffer from huge distribution discrepancies caused by different human body characteristics [26]. To overcome it, transfer learning is an essential procedure to extract invariant features and refining shared action representations across data. A preliminary research of transferable CNN network (i.e., finetune strategy) showed its capability of transferring knowledge from non-manufacturing specific human activities to engine block assembly actions [4]. Also, some cross-modal similarity metrics were utilized to transfer action pattern knowledge among images and videos [27].

From the literature, one can find that most existing methods fail to provide semantic knowledge for HRCA in advance, let alone to handle the robot planning proactively.

III. Methodology

To realize Proactive HRCA, three critical steps should be undertaken. The first step is human ongoing action inference, which is achieved by decreasing the time dimension of input data while increasing information modalities, i.e., multimodal intelligence. Secondly, transfer learning-based online prediction from the real-case data stream is a prerequisite, which allows quick deployment of the inference model with few

TABLE I
ARCHITECTURE OF INFLATED RSENET50

Layer	Operator	Parameter size
Input	Frame sampling	B, C, T, W, H
Video2Img	Reshaping	$B \times T, C, W, H$
2D ConvNet	2D Convolution	$7 \times 7, 64, /2$
	2D Max Pooling	$3 \times 3, 64, /2$
Img2Video	Reshaping	B, C, T, W, H
Res2 block	3D Convolution	$\begin{bmatrix} 1\times1\times1 \\ 3\times3\times3 \\ 1\times1\times1 \end{bmatrix} \times 3, 256, /1$
Res3 block	3D Convolution	$\begin{bmatrix} 1\times1\times1 \\ 3\times3\times3 \\ 1\times1\times1 \end{bmatrix} \times 3, 512, /2$
Res4 block	3D Convolution	$\begin{bmatrix} 1\times1\times1 \\ 3\times3\times3 \\ 1\times1\times1 \end{bmatrix} \times 3, 1024, /2$
Res5 block	3D Convolution	$\begin{bmatrix} 1\times1\times1 \\ 3\times3\times3 \\ 1\times1\times1 \end{bmatrix} \times 3, 2048, /2$
Avg-Pool	3D Average Pooling	$t \times w \times h$
Dense	Linear Regression	K

annotation data in various industrial scenarios. Lastly, based on the intention prediction ahead of schedule in today's Industrial Internet-of-Things (IIoT) environment of modern factories, dynamic decision-making can be made to instruct the mobile robot for proactive collaborations.

A. Multimodal intelligence-enabled human ongoing action recognition

A multimodal intelligence-enabled method is proposed to reliably classify ongoing human actions from RGB videos and skeleton joints, as shown in Fig. 1. The multimodal fusion network mainly contains three parts. The first one is implemented by efficient inflated ResNet [5], which can extract subtle visual action patterns. Based on spatial-temporal graph convolutional networks (ST-GCN) [7], the second skeleton part element-wise maximizes action representations of human body topology. The fusion part adopts intermediate attention to explore cross-channel relationships of visual and skeleton modalities and to improve the late fusion [28].

1) Efficient inflated ResNet for video representation classification: Visual action patterns are recognized by an efficient video partitioning strategy and an inflated ResNet50 module.

In the visual stream, an input video is firstly split into N equal sequences, $S_1, S_2, \ldots, S_n$. Then, one frame is randomly sampled from each split sequence. The partitioning strategy enables the visual sub-network to against instance variations of an action. Meanwhile, this mechanism can allow a lite network architecture, including either long-term videos or high-frame-rate clips with redundant information.

Inflated ResNet50 is utilized to extract spatio-temporal features of human activities from sampled frames. As presented in Table I, the module is composed of 2D ConvNet and inflated 3D ones. Parameters of the network, i.e., B, C, T, W and H denote the batch size, channels, temporal length, width and height of input data, respectively. 2D ConvNet is utilized to enhance the spatial features of each RGB frame along the time dimension. Then, inflated 3D ConvNet directly explores spatial-temporal representations of these features. The 3D ConvNet can be achieved by inflating all square filters $N \times N$ to cubic ones ($N \times N \times N$). Therefore, convolutional and pooling filters are endowed with temporal modeling characteristics. These 3D kernels can be initialized via repeating pre-trained weights of 2D filters N times along the time dimension and are divided by N for normalization, instead of training from scratch. In this way, a 3D ConvNet can be trained without increasing data scale, despite increased huge training parameters. Finally, a dense layer is stacked upon the 3D ConvNet to predict K action classes.

2) Element-wise ST-GCN for human body topology extraction: Nine ST-GCN layers and an element-wise maximization mechanism are major components of the skeleton stream.

To provide input for ST-GCN layers, human skeleton sequences in 2D or 3D coordinates are mapped to an undirected spatial-temporal graph $G = (V, E)$. In detail, node set $V = v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N$ represents N body joints across T frames in the time dimension. These nodes are connected based on skeleton sequences in the spatial graph, while the temporal graph links body joints between contiguous frames.

Then, graph convolution filters of an ST-GCN layer slide over the skeleton graph to extract latent features of human actions. The spatial-temporal graph convolution is defined as,

$$f_{out} = \Lambda^{-\frac{1}{2}}(A + I)\Lambda^{-\frac{1}{2}} f_{in}(P(v_{ti})) \cdot w(l_{st}(v_{ti})) \quad (1)$$

where $\lambda^{ii} = \sum_j A^{ij} + I^{ij}$. Adjacency matrix A and identity matrix I denote nodes' connection in graph G. The rest parts are feature mapping f_{in}, sampling function P, weighting function w and partition operation l_{st} . Similarly to convolution for images, sampled graph nodes are mapped to dimension c via $f_{in} : V^2 \to \mathbb{R}^c$, while function w generates a weight vector of the same dimension for computing their inner product.

For a node v_{ti}, input neighbors for graph convolutional filter are enumerated via sampling function P

$$P(v_{ti}) = v_{qj} | d(v_{tj}, v_{ti}) \leq D, |q - t| \leq \Gamma \quad (2)$$

where $d(v_{tj}, v_{ti})$ denotes the distance between these two nodes for a graph in frame t, while Γ is the time length across two spatial graphs. D and Γ are set to 1 in our experiment. For each node v_{tj} in a spatial graph, its neighbor nodes are then divided into three subsets via partition strategy l. These With v_c denoting the gravity center of one spatial graph, the partition strategy can be achieved by

$$l_{ti}(v_{tj}) = \begin{cases} 0, & \text{if } d(v_{tj}, v_c) = d(v_{ti}, v_c) \\ 1, & \text{if } d(v_{tj}, v_c) < d(v_{ti}, v_c) \\ 2, & \text{if } d(v_{tj}, v_c) > d(v_{ti}, v_c) \end{cases} \quad (3)$$

Similarly, the neighbor node v_{qj} of v_{ti} follows strategy l_{st} in the spatial temporal graph. The graph convolution filters slide over the graph with weighting different values to these nodes, extracting their spatial and temporal relationships.

$$l_{st}(v_{tj}) = l_{qi}(v_{qj}) + (q - t + \Gamma) \times 3 \quad (4)$$

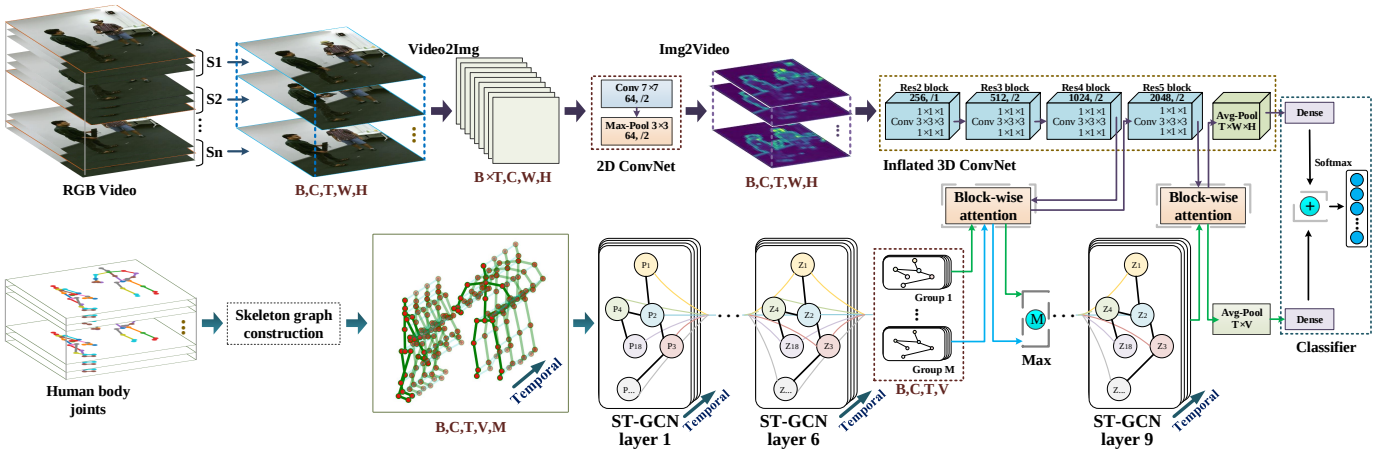

Fig. 1. The architecture of our multimodal fusion network, including a visual stream, a skeleton stream, and fusion modules.

In this way, six ST-GCN layers are stacked on graph G and to distill low-level information (e.g., motion of a joint) of human actions. Then, the extracted feature maps are split into M (i.e., the number of input persons) groups. The element-wise maximization mechanism compares every element in feature maps across these groups for retrieving the maximum value, which is fed to the extra three ST-GCN layers to extract high-level skeleton patterns of human actions. Pooling operation and linear regression are utilized as the final classifier.

3) Intermediate attention and fusion for multi-modalities: The visual and skeleton modalities are integrated via intermediate attention and late fusion. As presented in Fig. 2, the intermediate attention can generate optimized feature maps $\hat{M}_1, \ldots, \hat{M}_n$ by exploring inter-context relationships of input modalities $M_1, \ldots, M_n$. This can be achieved by three steps,

(a) *Segmentation*. From modality M_1 to M_n, features maps are split into equal blocks $S_i, i \in 1, \ldots, n$. The number of blocks in modality M_i is calculated by $|S_i| = \lceil C_i/C_s \rceil$, where C_i is the number of channels of data in M_i and $C_s = \min[S_1, ..., S_n]/2$. The last block in S_i can be padded with zeros if C_i is not a multiple of C_s. Then, blocks in one modality M_i are related via element-wise summing over $S_1, \ldots, S_i$.

(b) *Connection*. The output of the above part is denoted as D_i in modality M_i, where $D_i \in \mathbb{R}^{N_1 \times \cdots \times N_k \times C_s}$. After global average pooling P_i on D_i, as denoted in (5), multimodal contextual information is connected together by summing $P_1, \ldots, P_n$. Then, a 1×1 CNN layer is introduced to learn cross-channel relationships of these multimodal features.

(c) *Activation*. The connection part outputs global shared representation G. Attention weight W_i^j of the j-th block of the i-th modality is generated from sequential executions of linear transformation and SoftMax activation on G (see (6)), where $i \in 1, \ldots, n$ and $j \in 1, \ldots, |S_i|$. Then, a optimized feature block $\hat{S}_i^j$ is obtained via (7). λ is set to 0.5 in our paper. Finally, optimized $\hat{M}_i$ is output after concatenation over $\hat{S}_i^j$, i.e., $\hat{M}_i = \left[\hat{S}_i^1, \ldots, \hat{S}_i^{|S_i|}\right]$.

$$P_i(C_s) = \frac{1}{\prod_{j=1}^{k}} \sum_{(N_1,\ldots,N_k)} D_i(N_1, \ldots, N_k, C_s) \tag{5}$$

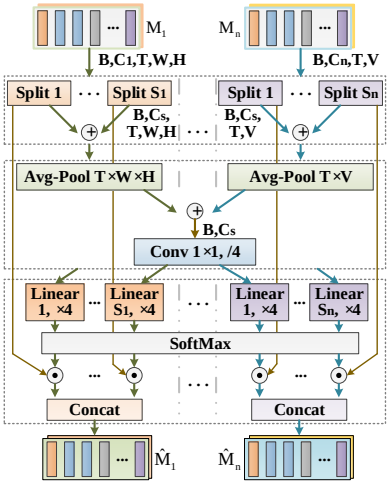

Fig. 2. The architecture of the intermediate attention module.

$$W_i^j = \frac{exp(h_i^j G + b_i^j)}{\sum_i^n \sum_j^{|S_i|} exp(h_i^j G + b_i^j)} \tag{6}$$

$$\hat{S}_i^j = \left[\lambda + (1-\lambda) \times W_i^j\right] \times S_i^j \tag{7}$$

As illustrated in Fig. 1, the first intermediate attention is inserted between the fourth layer of inflated ResNet50 and person-wise groups of the sixth ST-GCN layer. Another one is positioned between the fifth layer of the visual stream and the ninth ST-GCN layer. They can relate the block-wise correlation of visual and skeleton representations, from shallow feature recalibration to high-level contextual awareness. Then, the output of the visual part and the skeleton one are element-wise added for the late fusion, followed by a SoftMax layer for action pattern classification. In the training procedure, the cross-entropy loss is utilized to fit the weights of our model.

B. Transfer learning-based online operator intention analysis for proactive HRCA

Based on ongoing human action recognition via multimodal fusion, a transfer learning-based online operator intention analysis method is further proposed to allow proactive HRCA, as presented in Fig. 3. The method mainly tackles three crucial

issues: 1) knowledge transfer with limited annotation data, 2) online operation intention analysis, and 3) dynamic decision-making mechanism, which are elaborated below.

1) Operator action recognition with knowledge transferring: There are two major challenges in the industrial operation recognition tasks. First, it is unrealistic to build a universal HRCA model covering all potential assembly sequences. Second, it is expensive and time consuming for data annotation in various scenarios under massive sensor networks in modern factories. Hence, a semi-supervised transfer learning procedure is presented. Finetune strategy is firstly adopted to enable the extractor G to learn action patterns from daily activities to industrial actions. Then, a domain adaptation module θ_m is connected to the G. In terms of visual and skeleton streams, θ_m^v and θ_m^s are utilized to reduce the distribution discrepancy of source and target data, respectively. For each action pattern among two domains, this allows G to align learned action representations of few labeled data to the remaining unlabelled ones. Specifically, a fully connected layer is stacked on the G to map features of the source domain δ_s and the target domain δ_t. The distribution distance between these two domains is calculated by the maximum mean discrepancy (MMD) metrics,

$$D = \|\delta_s - \delta_t\|_H = \|\langle \delta_s, \delta_t \rangle_H\| \tag{8}$$

where H is the reproducing kernel Hilbert space (RKHS). The unbiased estimation value of $\langle \delta_s, \delta_t \rangle_H$ is calculated by the Gaussian radial basis function, i.e., $\hat{D} = exp(\|\delta_s - \delta_t\|^2 / 2\sigma)$.

The training strategy contains three stages. The first one is to fine-tune weights of the extractor G and the classifier θ_c with a portion of annotation assembly action data, as they are pre-trained from other domains. Then, MMD metrics θ_m^v and θ_m^s are introduced to train the extractor G to learn latent sharing representations of assembly actions between labeled and unlabeled data. This semi-supervised training stage keeps weights of the classifier θ_c frozen. The final stage is to further fine-tune the classifier θ_c with annotation data while freezing the extractor G and removing MMD metrics.

2) Online operation intention analysis in industrial settings: The procedure of online operation intention analysis is divided into five steps, (a) RGB-D video acquisition, (b) color-depth camera calibration, (c) visual and skeleton output, (d) feature extraction, and (e) pattern classification, as shown in Fig. 3. In workshops, Azure Kinect can be adopted to record RGB-D videos and output 3D skeleton joints. However, human subject only accounts for a small part of videos. To remove background interference, pre-processing strategies are applied.

To acquire 2D pixel coordinates from 3D joints, the color camera and the depth camera in Kinect are calibrated first. Based on camera model $P_{uv} = TP_c = KTP_w$, the transform matrix from depth camera to the color image is calculated by,

$$P_{uv-c} = K_c T_c T_d^{-1} P_d \tag{9}$$

where c denotes the parameters of the color camera and d is the depth one, while K is camera intrinsics and $T = [R|t]$ is camera extrinsics. In this way, 2D body joints can be acquired. The width w_b and height h_b of a bounding box of this person can be obtained as well. Then, a crop mask $(w_m, \alpha \times w_m)$ in a uniform aspect ratio α is introduced to crop the human part from images. The width w_m is calculated by $w_m = \max\{\lfloor w_b, h_b/2 \rfloor\} + d_h$, where d_h is the distance from pelvis to the middle of spine. α is set to 2 in this work. For skeleton data, we apply a normalization processing similar to [29]. With the visual and skeleton output, a pre-trained multimodal model estimates human operation intentions.

Meanwhile, for live data streams, the detailed procedures of online intention analysis are presented in algorithm 1. A queue Q which can hold N frames is utilized to record the incoming video stream. Besides, a memory queue M stores T temporal-video segments, each of which is in the same size of Q. N frames stand for a timestamp and it is the number of input frames for the action recognition model. When the live video stream starts, all frames will fill into M uniformly with N frames until overflow. Then, half frames of Q are queued to M in the case of passing N frames. Input container I_d samples frames from these T temporal segments of M for action prediction. Assuming $T = 3$, I_d includes 25% samples of Q at the time step M_0, 25% samples of Q at M_1 and 50% samples of Q at the last timestamp M_2. Long-range information of the incoming stream is recorded in this online mode while recent frames are given more importance. Hence, the online mode can foresee operator operations of varying time lengths.

Algorithm 1: Online operation intention recognition

input: RGB-D live video stream (V_d)
Number of visual-stream input (N)
Pre-processing module (Δ)
Pre-trained action recognition model (Φ)
Output: Action predictions
Calculate the number of timestamps $T = \lceil \log_{0.5} \frac{1}{N} \rceil$;
Initialize video queue Q for N coming frames;
Initialize memory queue M for $N \times T$ frames;
Mark M with timestamps, i.e., $M = \{M_0, \ldots, M_T\}$;
Initialize input container I_d for N RGB-D frames;
while *new frames available from* V_d **do**
  Add RGB-D frame f_i from V_d to queue Q;
  if $i \% N$ **then**
    if $i < N \times T$ **then**
      Add N frames from Q to queue M;
      $I_d := \{0.5^{i/N} M_0\} \cup_{t=1}^{i/N} \{0.5^{i/N - t + 1} M_t\}$;
    else
      Add 50% frames from Q to M and update;
      $I_d := \{0.5^{T} M_0\} \cup_{t=1}^{T} \{0.5^{T - t + 1} M_t\}$;
    end
    Feed I_d to Δ and Φ to obtain prediction;
  end
end

3) Dynamic decision-making mechanism towards proactive HRCA: Based on human ongoing operation analysis ahead of schedule, a dynamic decision-making mechanism is adopted to make proactive robotic planning, which allows a mobile robot to assist the operator intelligently. The mechanism includes robotic decision and adaptive control in the world coordinates.

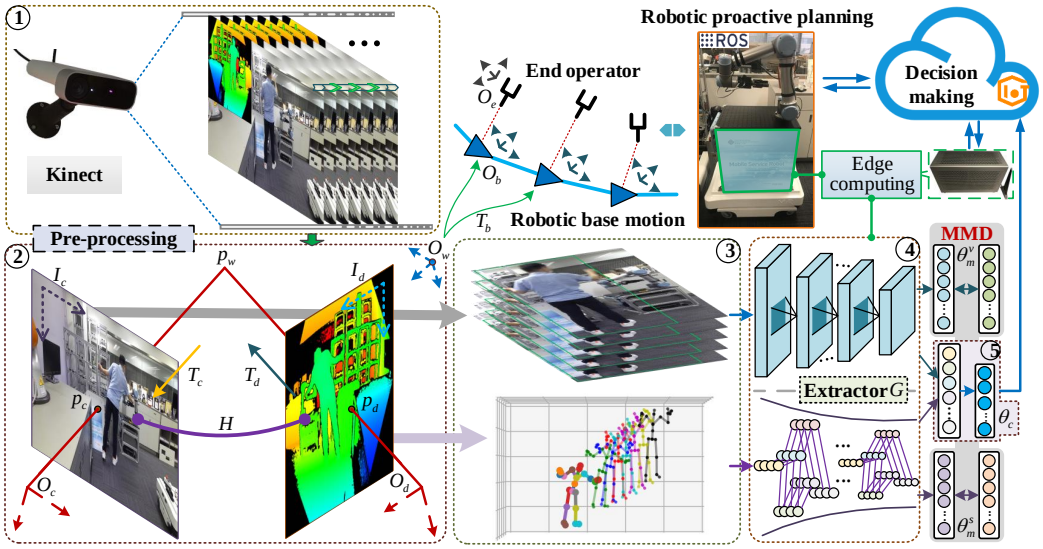

Fig. 3. The proposed flowchart towards proactive HRCA in the IIoT environment.

TABLE II
SAMPLES OF HUMAN ACTIONS AND ROBOT PLANNING

Production	Communication	Robotic planning
Part picking/fixing	Robotic guiding	Obstacle avoidance
Station changing	Robotic leaving	Vision inspection
Screwing	Part selection	Toolbox picking/holding/placing
Taping	Tool selection	Motion following/pausing/leaving

The robotic motion decisions is made for Proactive HRCA considering operators' intentions, which are computed in real-time by edge computing infrastructure. Typical samples of human actions and robotic planning are listed in Table II. Human workers may perform these actions in the production process or during human-robot communication. The collaborative robot can conduct corresponding planning to assist human operators in an industrial task. All of these are encapsulated as retrievable knowledge in the IIoT environment. Moreover, proactive robotic planning is executed in the world coordinates. Similarly to the color-depth camera calibration, robot hand-eye calibration is conducted to connect visual sensors and robot kinematics. Robot base motion can be described as $O_b^i = T_b^i O_w, i = p_1, \ldots, p_n$, where i denotes different locations of the robot base. Similarly, the trajectory of the robot end operator is denoted by $\{O_{e_j}^{p_1}, \ldots, O_{e_j}^{p_n}\}$, where $j = a_1, \ldots, a_n$ is robot controllers' action. Programming with these real-time updated robot states, the robot controller dynamically adjusts the relating contact force and robot position. The calibration ensures the accuracy of robot control and safety of co-work with human operators in close proximity.

IV. CASE STUDY AND EXPERIMENTAL RESULTS

In this section, a demonstrative case study of Proactive HRCA for bracket assembly task in aircraft cabins is carried out in the lab environment, to further evaluate the performance of our proposed approach.

In current aviation industry, the interior assembly tasks largely rely on manual operations with domain expertise, due to the narrow workspace in aircraft cabins. During the assembly process, an operator has to constantly pass between an assembly area and a tool storage area for toolbox change with heavy workload, which is tedious with low efficiency. To overcome this challenge, a mobile robot is leveraged to conduct pick-and-place work in the human-centered assembly. As demonstrated in Fig. 4, the Proactive HRCA for bracket assembly is achieved by online operator intention analysis, robotic dynamic decision and adaptive control stepwisely.

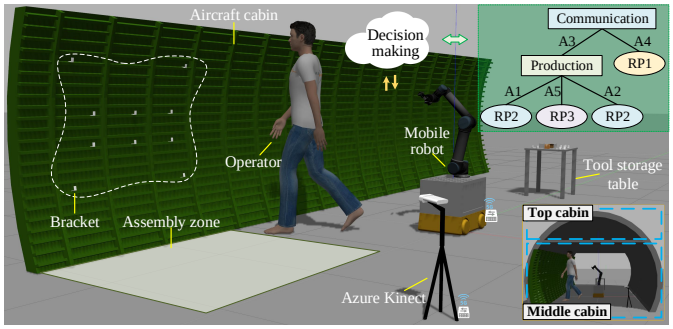

Fig. 4. Demonstration of proactive HRCA in aircraft bracket assembly.

A. Operator action recognition and intention analysis

Operator action recognition and intention analysis is performed by our proposed multimodal transfer learning-based network, which is the prerequisite to ensure Proactive HRCA. In our settings, a large-scale human action dataset is firstly leveraged to evaluate the performance of the proposed multimodal fusion network, including accuracy, efficiency, and universality. Then, an assembly action dataset (AAD) is developed to demonstrate that the proposed action recognition model can estimate ongoing human operations with knowledge transferring.

1) Evaluation of numerous action recognition: NTU-RGB+D is an open human action dataset, which covers 60 daily action classes [29]. Two benchmarks of this dataset, cross-subject (X-sub) and cross-view (X-view) [7], are used to verify the top-1 recognition accuracy of our model.

TABLE III
ACTION RECOGNITION PERFORMANCE ON THE NTU-RGB+D DATASET

Model	X-Sub	X-View
ST-GCN (Skeleton only) [7]	81.50%	88.30%
Glimpse Clouds (RGB only) [30]	89.60%	93.20%
RGB-Skeleton fusion [9]	85.40%	91.60%
IR-Skeleton fusion [31]	91.80%	94.90%
Ours (RGB-Skeleton)	91.10%	96.00%

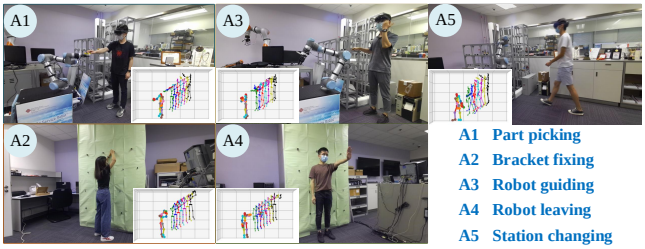

Fig. 5. Samples of industrial operator' assembly action.

TABLE IV
ACTION RECOGNITION PERFORMANCE ON THE AAD DATASET

μ	A1	A2	A3	A4	A5	mAP
0.05	91.11%	98.55%	92.59%	99.26%	94.23%	95.15%
0.10	97.04%	100.00%	100.00%	100.00%	98.08%	99.02%
0.20	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

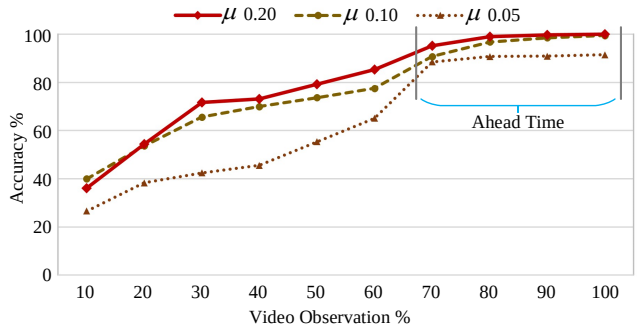

Fig. 6. Early Action Prediction Results for Assembly Operation.

Training. For visual stream, the efficient video partitioning strategy samples 15 RGB frames, while the skeleton stream initializes 300 frames. The multimodal fusion model is optimized by stochastic gradient descent with an initial learning rate of 0.001, which is multiplied by 0.1 after every 10 epochs.

Result. As shown in Table III, compared with single modality, e.g., RGB or skeleton, our proposed multimodal fusion method presents an obvious improvement in the recognition accuracy. Further, the proposed model enables a better fusion result in comparison with another two feature fusion methods. Similar to the definition of frames per second (fps), the speed of our approach can be measured by videos per second (vps). On a Tesla V100 GPU (16G), our model runs at 17 vps, which definitely meet the real-time requirement for online action prediction.

2) Evaluation of action recognition with fewer frames: As presented in Fig. 5, the AAD contains operators' operation for the bracket assembly task. In this case, operators may carry tools and wear smart equipment (e.g. AR glasses) to collaborate with a mobile robot. The dataset consists of 256 RGB-D videos captured by Azure Kinect, including five different operations: A1) part picking, A2) bracket fixing, A3) robot guiding, A4) robot leaving, and A5) station changing. It is noted that the time consumption that an operator executes different operations is normally uncertain. Therefore, these actions last for two seconds to five seconds in unequal time lengths. There are around three to five action groups in one clip on average. Hence, up to 939 action samples are available in this dataset. The RGB-D video comprises visual frames (640×576 resolutions) and 3D skeleton poses (25 body joints).

Experimental setting. The AAD is divided into a training set (467 samples) and the testing one (472 samples). For the training dataset, we control the percentage μ of annotation data to simulate the real cases in today's factories, where data acquired by massive sensor networks are unlabeled. Meanwhile, settings of optimizer remain the same as the training procedure of the multimodal fusion model. During the training process, our semi-supervised model transfers knowledge from the NTU dataset to the assembly pattern, as well as to extract the shared and domain-invariant features between labeled target domain and unlabeled source domain.

Results. As shown in Table IV, the mean average precision (mAP) is utilized to measure the accuracy performance of our model under different thresholds μ. The results evaluate the effectiveness of the knowledge transferring ability of the approach, despite the discrepancy of motion sequences and visual patterns of surrounding scenarios. Since the transfer learning-based model aims to infer ongoing operator actions from visual and skeleton topology, input frames are largely reduced. The visual stream uses 15 RGB frames as input, while the skeleton subnet samples 50 frames. Besides, the model can predict operators' actions ahead of time with partial observations of videos. As illustrated in Fig. 6, the proposed model can output early action prediction results ahead of 30% timestamps for action video streams. The prediction manner is practically significant in real industrial scenarios. Different workers spend inconsistent time for one same operation execution. But the content of an industrial operation normally contains several exact action sequences. In this case, despite uncertain time consumption for one operation, our model can predict the operator's action ahead of time after seeing a short part of these exactly essential sequences.

B. Towards proactive HRCA in aircraft bracket assembly

On the basis of operation prediction and intention analysis, Proactive HRCA can be achieved via robotic dynamic decision and adaptive control in advance.

1) Robotic dynamic decision-making: In the human-centered aircraft bracket assembly task, the mobile robot can proactively assist the operator in these plannings. Table V lists six different cooperation samples. In response to human actions during the production and communication process, there are three assisted robot plannings, namely: a) moving away from the operator to pick toolboxes from storage areas (RP1), b) moving towards the operator to give him toolboxes (RP2), and c) following operators' motion to help him

TABLE V
ROBOT PLANNING IN RESPONSE TO OPERATOR ACTIONS

Num.	Production	Communication	Robot planing
1	Part picking (A1)	Robot leaving (A4)	RP1
2	Part picking (A1)	Robot guiding (A3)	RP2
3	Bracket fixing (A3)	Robot guiding (A3)	RP2
4	Bracket fixing (A3)	Robot leaving (A4)	RP1
5	Station changing (A5)	Robot guiding (A3)	RP3
6	Station changing (A5)	Robot leaving (A4)	RP1

hold tools or parts (RP3). To learn these semantic links of cooperation, Iterative Dichotomiser 3 (ID3) can be utilized to generate three potential decisions for robot planning. ID3 create a multi-way tree via the largest decision information, i.e., $Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2^{p_k}$, where $|y|$ is the number of decision categories and p_k is the proportion of the k-the decision.

2) System deployment toward proactive HRCA: Under the IIoT environment, the demonstrative system towards proactive HRCA for bracket assembly is deployed as Fig. 4. Firstly, Azure Kinect is employed in the middle cabin to capture the living video stream at 30 fps. In the online mode, our action recognition model can predict operators' intentions ahead of the schedule. This inference process is conducted on the edge service and predictions are uploaded to the cloud server in time. Based on the predictions, the lite decision-making model (ID3) can generate robot decisions dynamically in the cloud. The mobile robot, which consists of a Universal Robots UR5 and an AGV base, can obtain the generated robotic control instruction (e.g., programming position and speed) in advance. In this way, the robot can proactively assist the operator in considering his coming operation goals. The proactive HRCA allows for high-level collaboration and productivity in the bracket assembly task.

V. DISCUSSIONS

Proactive HRCA allows a mobile robot to learn about operators' intentions and to conduct corresponding assistance in advance. From the experimental results, it can be found that by leveraging the proposed multimodal fusion network, ongoing human action can be predicted with fewer frames, owing to the ample information integration in a short timespan. Meanwhile, the transfer learning-based online prediction algorithm can estimate operators' ongoing intentions ahead of time with contextual information of the incoming data stream, no matter with short actions (e.g., robot leaving) or long ones (e.g., bracket fixing). It allows universal feasibility for all potential industrial scenarios with knowledge transferring.

This proactive HRCA paradigm is expected to lead to yet unattained efficiency for flexible production in real industrial cases, regardless of the complex structures or narrow spaces in the workspace. Despite these advantages, the computing efficiency in the inference process of our propopsed model can be further accelerated by more explorations, for example, by taking fewer frames (i.e. ten RGB images) as input.

VI. CONCLUSIONS

With the prevailing implementation of IIoT and robot learning, it is foreseen that proactive HRCA will become dominant in the next generation smart manufacturing paradigm, which can largely facilitate flexible production for mass personalization. As the prerequisite, human ongoing action prediction can be achieved by feeding short-time frames but multi-modalities information. Then, with knowledge transferring from daily activities to assembly patterns, the online operation intention analysis can be performed in advance. Based on the predictions, the dynamic decision can be further made in advance to allow a mobile robot to assist the human operators in a proactive manner. To summarize, the main scientific contributions of this research lies in two aspects:

1) Proposed a multi-modal fusion network to recognize ongoing human actions with fewer incoming frames. This overcomes a persistent challenge in industrial action recognition, of which many fine-grained operations lie in, such as wedging pins and screwing bolts.
2) Introduced an online prediction procedure based on knowledge transferring to predict operators' ongoing operations ahead of time even in a new industrial scenario. This work allows quick model deployment with less supervised information, which can also be applied to many other manufacturing scenarios, in lack of annotations from massive sensor data.

Moreover, based on the comparative results by testing on both daily activity dataset and assembly operations, our proposed action recognition model achieves competitive performance than other existing ones. Apart from these achievements, several potential future research directions are also highlighted here, including 1) developing an adaptive robot control program (precision location and tracking) considering human safety, availability of resources, and the required time of operation holistically, 2) accelerating the training process of the model by leveraging all the available distributed computing resources in the IIoT environment, i.e., federated learning, and 3) augmenting human intelligence, e.g., AR glasses, to well equip operators for collaborating with robots in a more cognitive manner.

REFERENCES

[1] P. Zheng, Z. Wang, C.-H. Chen, and L. P. Khoo, "A survey of smart product-service systems: Key aspects, challenges and future perspectives," *Advanced engineering informatics*, vol. 42, p. 100973, 2019.
[2] S. Liu, L. Wang, and X. V. Wang, "Symbiotic human-robot collaboration: multimodal control using function blocks," *Procedia CIRP*, vol. 93, pp. 1188–1193, 2020.
[3] A. Hietanen, R. Pieters, M. Lanz, J. Latokartano, and J.-K. Kämäräinen, "Ar-based interaction for human-robot collaborative manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 63, p. 101891, 2020.
[4] Q. Xiong, J. Zhang, P. Wang, D. Liu, and R. X. Gao, "Transferable two-stream convolutional neural network for human action recognition," *Journal of Manufacturing Systems*, 2020.
[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.
[6] X. Wen, H. Chen, and Q. Hong, "Human assembly task recognition in human-robot collaboration based on 3d cnn," in *2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 1230–1234. IEEE, 2019.
[7] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[8] H. Liu and L. Wang, "Collision-free human-robot collaboration based on context awareness," *Robotics and Computer-Integrated Manufacturing*, vol. 67, p. 101997, 2021.

[9] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, and P. Liu, "Action recognition based on 3d skeleton and rgb frame fusion," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 258–264. IEEE, 2019.

[10] L. Wang, R. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris, and G. Chryssolouris, "Symbiotic human-robot collaborative assembly," *CIRP annals*, vol. 68, no. 2, pp. 701–726, 2019.

[11] B. Wang, S. J. Hu, L. Sun, and T. Freiheit, "Intelligent welding system technologies: State-of-the-art review and perspectives," *Journal of Manufacturing Systems*, vol. 56, pp. 373–391, 2020.

[12] X. V. Wang, L. Wang, M. Lei, and Y. Zhao, "Closed-loop augmented reality towards accurate human-robot collaboration," *CIRP Annals*, vol. 69, no. 1, pp. 425–428, 2020.

[13] Q. Gao, J. Liu, Z. Ju, and X. Zhang, "Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9663–9672, 2019.

[14] M. A. Simao, O. Gibaru, and P. Neto, "Online recognition of incomplete gesture data to interface collaborative robots," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9372–9382, 2019.

[15] L. Johannsmeier and S. Haddadin, "A hierarchical human-robot interaction-planning framework for task allocation in collaborative industrial assembly processes," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 41–48, 2016.

[16] J. Zhang, H. Liu, Q. Chang, L. Wang, and R. X. Gao, "Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly," *CIRP annals*, vol. 69, no. 1, pp. 9–12, 2020.

[17] Y. Cheng, L. Sun, C. Liu, and M. Tomizuka, "Towards efficient human-robot collaboration with robust plan recognition and trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2602–2609, 2020.

[18] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[19] S. Li, P. Zheng, and L. Zheng, "An ar-assisted deep learning-based approach for automatic inspection of aviation connectors," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1721–1731, 2020.

[20] A. A. Hammam, M. M. Soliman, and A. E. Hassanien, "Real-time multiple spatiotemporal action localization and prediction approach using deep learning," *Neural Networks*, vol. 128, pp. 331–344, 2020.

[21] D. Luvizon, D. Picard, and H. Tabia, "Multi-task deep learning for real-time 3d human pose estimation and action recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[22] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," *arXiv preprint arXiv:2103.02758*, 2021.

[23] Y. Dang, F. Yang, and J. Yin, "Dwnet: Deep-wide network for 3d action recognition," *Robotics and Autonomous Systems*, vol. 126, p. 103441, 2020.

[24] Q. Liu, Z. Liu, W. Xu, Q. Tang, Z. Zhou, and D. T. Pham, "Human-robot collaboration in disassembly for sustainable manufacturing," *International Journal of Production Research*, vol. 57, no. 12, pp. 4027–4044, 2019.

[25] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in *European Conference on Computer Vision*, pp. 72–90. Springer, 2020.

[26] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2018.

[27] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3168–3182, 2019.

[28] L. Su, C. Hu, G. Li, and D. Cao, "Msaf: Multimodal split attention fusion," *arXiv preprint arXiv:2012.07175*, 2020.

[30] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 469–478, 2018.

[29] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

[31] A. M. De Boissiere and R. Noumeir, "Infrared and 3d skeleton feature fusion for rgb-d action recognition," *IEEE Access*, vol. 8, pp. 168 297–168 308, 2020.

Shufei Li is currently working towards the Ph.D. degree in the Department of Industrial and Systems Engineering at the Hong Kong Polytechnic University. He received his M.S. degree in the Department of Industrial and Manufacturing Systems Engineering from Beihang University in 2020, and the B.E. degree in the Mechatronic Engineering from Shandong Jianzhu University, China, in 2017. His research interests include proactive HRC, AR-assisted smart manufacturing, deep learning and computer vision.

Pai Zheng (M'17) is currently an Assistant Professor in the Department of Industrial and Systems Engineering, at the Hong Kong Polytechnic University. He received his Ph.D. degree in Mechanical Engineering from the University of Auckland in 2017, Master's Degree in Mechanical Engineering from Beihang University in 2013, and Dual Bachelor's Degrees in Engineering from Huazhong University of Science and Technology, in 2010. His research interest includes smart product-service systems, engineering informatics, and manufacturing servitization. He is a member of IEEE, CMES, and ASME, and serves as the Associate Editor of IET Collaborative Intelligent Manufacturing, and Editorial Board Member for the journal of Advanced Engineering Informatics.

Junming Fan is currently working towards the Ph.D. degree in the Department of Industrial and Systems Engineering, at the Hong Kong Polytechnic University, China. He received his Bachelor's degree and Master's degree in electronic engineering from University of Electronic Science and Technology of China in 2015 and 2018 respectively. His research interest includes computer vision, deep learning, and human-robot collaboration.

Lihui Wang is a Chair Professor at KTH Royal Institute of Technology, Sweden. His research interests are focused on cyber-physical systems, real-time monitoring and control, human-robot collaborations, and adaptive manufacturing systems. Professor Wang is actively engaged in various professional activities. He is the Editor-in-Chief of International Journal of Manufacturing Research, Journal of Manufacturing Systems, and Robotics and Computer-Integrated Manufacturing. He has published 10 books and authored in excess of 500 scientific publications. Professor Wang is a Fellow of Canadian Academy of Engineering, CIRP, SME and ASME. He is also a Professional Engineer in Canada, the President of North American Manufacturing Research Institution of SME, and the Chairman (2018-2020) of Swedish Production Academy.