IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1199-1211
Journal homepage: http://iieta.org/journals/ts

Skill Spectrum Construction and Task Allocation Optimization Driven by Image Recognition for Digital Organization Management

Jiangyong Liu[1], Wenbo Yu[2], Yu Li[1*]

[1] School of Management, Guangzhou Huashang College, Guangzhou 510000, China
[2] Guangdong Jin Yi Bai Technology Co., Ltd., Foshan 528200, China

Corresponding Author Email: liyu6666660509@163.com

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.430310

Received: 19 December 2025
Revised: 22 May 2026
Accepted: 29 May 2026
Available online: 30 June 2026

Keywords:
image representation learning, spatiotemporal graph convolution, contrastive learning, skill spectrum, task allocation optimization, multi-scale feature fusion

ABSTRACT

In the process of digital transformation, the skill assessment and task scheduling for manual operation roles constitute a critical challenge in human resource management. Conventional approaches, which rely heavily on subjective experience, are inherently limited by coarse-grained matching granularity, static scheduling mechanisms, and a significant lack of objectivity, rendering them inadequate for the demands of flexible production and refined operational control. To address these limitations, an end-to-end image-driven framework for skill spectrum construction and task allocation optimization was proposed, through which the problems of skill modeling and scheduling decision-making in organizational human resource management were reformulated as three visual computing tasks: spatiotemporal image representation, multi-scale feature association, and temporal sequence prediction. Within this framework, an operational skill heatmap trajectory encoding mechanism was first constructed, through which the motion trajectories of human keypoints were mapped to multi-channel spatiotemporal heatmap images. A spatiotemporal graph convolutional network was then employed to achieve unsupervised objective quantification of skill magnitude, from which structured skill spectrum nodes were generated. Subsequently, multi-scale visual features were fused, and latent mappings between skills and tasks were uncovered through a cross-attention mechanism, generating an interpretable skill-task bipartite association graph. Finally, temporal contrastive learning was introduced to construct a skill evolution prediction model, which was integrated with a bi-objective greedy allocation strategy to achieve the joint optimization of task execution efficiency and skill development, thereby establishing a closed-loop management mechanism. Extensive experimental validation, conducted on a self-constructed industrial assembly dataset and a public surgical skill dataset, demonstrated that the proposed methodology consistently outperformed baselines across skill quantification accuracy, association retrieval precision, and comprehensive allocation utility. The results establish a novel technological pathway for the effective deployment of vision techniques within organizational management contexts.

1. INTRODUCTION

The ongoing digital transformation of industry has progressively established flexible production systems and dynamic, multi-scenario operations [1, 2] as mainstream development trajectories within the manufacturing and service sectors. Under these paradigms, the precision of skill assessment and the efficiency of human resource scheduling for operational roles directly determine overall organizational performance. Traditional skill evaluation frameworks, however, remain heavily dependent on on-site supervision and manual scoring protocols [3, 4]. The outcomes of such evaluations are substantially influenced by the differences in rater experience and inconsistent application of assessment standards, resulting in pervasive issues of subjectivity, coarse granularity, and low throughput. These inherent limitations render conventional approaches ill-suited for the fine-grained and dynamic demands of modern workforce deployment. Recent advances in computer vision—particularly in human pose estimation and action quality analysis [5-7]—have rendered the automatic extraction of human behavioral features from routine operational video footage technically viable. This capability provides a foundational substrate for non-intrusive, objective skill modeling of individual employees. However, human resource decision-making within digitally transformed organizations continues to rely predominantly on manual rules and empirical judgments [8, 9], with no end-to-end decision architecture centered on visual computing having been established to date. To bridge this gap, the problems of skill modeling and task allocation within organizational management are reformulated in this work as standard visual computing tasks, and a comprehensive, image-driven management decision framework is constructed. This reformulation serves a dual purpose: it expands the deployable

application domain of image processing technologies, while simultaneously furnishing a novel computational paradigm for human capital management in digitally transformed enterprises.

From the perspective of visual computing, a systematic review of the existing literature reveals three pronounced research gaps that remain unaddressed. At the skill quantification level, existing vision-based action skill assessment methods predominantly produce a single quality score tailored to a specific task, without establishing a generalized, structured skill representation framework [10, 11]. Furthermore, most approaches directly model input keypoint coordinate sequences, without fully exploiting the spatiotemporal image-encoded information embedded within action trajectories [12, 13]. This limitation precludes the precise characterization of fine-grained skill dimensions such as action standardization and operational redundancy. At the association modeling level, the matching mechanisms between skills and tasks continue to rely predominantly on manually predefined skill dimensions and task requirement specifications, lacking the capability for data-driven discovery of latent associations [14, 15]. Existing visual matching methodologies have not yet established automatic mappings between multi-scale operational image features and skill attributes [16, 17], rendering the matching outcomes insufficiently interpretable and precluding the generation of structured skill-task association graphs. At the dynamic optimization level, prevailing task allocation methods are largely confined to optimal matching under static scenarios [18-20], with no consideration given to incorporating the temporal evolutionary characteristics of employee skills into the optimization objective. To date, no investigation has been conducted that transforms the skill evolution process into an image sequence learning task [21, 22]. Consequently, the capacity for forward-looking skill deficit forecasting remains absent, and the closed-loop, bi-objective optimization of both task execution efficiency and employee skill development has yet to be realized.

To address the aforementioned research gaps, a comprehensive image-driven framework for skill spectrum construction and task allocation optimization is proposed. The core technical innovations of this framework are fourfold. First, an operational skill heatmap trajectory encoding method is developed, through which the motion trajectories of human keypoints are transformed into multi-channel spatiotemporal heatmap images. A spatiotemporal graph convolutional network is then employed to achieve unsupervised skill magnitude quantification, and an image gradient-based redundant action metric is defined to construct a structured set of skill spectrum nodes. Second, a multi-scale visual feature fusion method for association modeling is introduced, which integrates three distinct scales of image features. A cross-attention mechanism is leveraged to uncover latent mappings between skills and tasks, and singular value decomposition is subsequently applied to generate an interpretable skill-task bipartite association graph—eliminating the need for manually predefined skill labels and task requirement specifications. Third, a temporal contrastive learning-driven framework for skill evolution prediction is established. The temporal variation of the skill spectrum is formulated as an image sequence learning task, and temporally contrastive constraints are imposed to learn semantically consistent skill embeddings. A gated recurrent unit is then integrated to enable forward-looking estimation of future skill deficits, thereby enhancing the robustness of long-horizon temporal predictions. Fourth, a bi-objective greedy task allocation policy is constructed, which jointly optimizes current skill-task compatibility and future skill development gains as the dual optimization objectives. This policy achieves closed-loop optimization of task scheduling and skill development, and comprehensive multi-dimensional performance validation is conducted across two distinct datasets.

The remainder of this study is organized as follows. Section 2 provides a detailed exposition of the complete image-driven algorithmic framework and the technical specifications of each constituent module, thereby establishing the theoretical foundation for the entire study. Section 3 presents a comprehensive experimental evaluation, through which the proposed methodology is assessed across multiple dimensions—including quantification, association, prediction, allocation, and generalization. Section 4 summarizes the methodological advantages, analyzes the remaining limitations, and identifies promising directions for future research. Section 5 concludes the study with a synthesis of the research contributions and a discussion of the technical value and application prospects.

2. METHODOLOGY

2.1 Overall framework

A comprehensive image-driven framework for skill spectrum construction and task allocation optimization is proposed, targeting the human resource management requirements of operational roles within digitally transforming organizations. With employee operation videos serving as the sole raw input, the framework proceeds sequentially through three progressive stages—skill spectrum node quantification, skill-task association graph generation, and dynamic prediction with allocation optimization—culminating in the final outputs of optimal task allocation schedules and employee skill evolution forecasts. Throughout the entire pipeline, image representation and visual computation serve as the core technical paradigm, eliminating the need for manually predefined skill dimensions and task requirement labels and enabling end-to-end automated computation from raw video data to management decision outputs.

These three stages constitute a fully integrated technical chain with sequential dependencies. In the skill spectrum node quantification stage, discrete human keypoint trajectories are transformed into continuous multi-channel image representations through spatiotemporal heatmap encoding. A spatiotemporal graph convolutional network is then employed to extract fine-grained skill features, from which a structured set of skill spectrum nodes is constructed, thereby achieving unsupervised objective quantification of employee operational skills. In the skill-task association graph generation stage, three distinct scales of visual features are fused to establish a unified task representation. A cross-attention mechanism is leveraged to uncover latent mappings between skill features and task features, and singular value decomposition is subsequently applied to generate an interpretable bipartite association graph, through which the correspondences between skill dimensions and task requirements are automatically established. In the dynamic prediction and allocation optimization stage, temporal skill data are reconstructed as image sequences, and temporal contrastive

learning is used to learn semantically consistent skill evolution representations. A temporal prediction network is then integrated to enable forward-looking estimation of future skill deficits. Finally, a bi-objective greedy policy is employed for task scheduling, thereby establishing a closed-loop optimization mechanism that jointly optimizes task execution efficiency and employee skill development.

2.2 Skill spectrum node quantification via spatiotemporal heatmap encoding

Skill spectrum node quantification serves as the foundational prerequisite for association modeling and task allocation, with the core objective of extracting objective, fine-grained, structured skill representations from unstructured operation videos. Modeling approaches that directly operate on keypoint coordinate sequences are inherently limited in their capacity to fully capture the spatiotemporal distribution characteristics of motion trajectories, and consequently provide only restricted utilization of fine-grained information such as motion rhythm and spatial coverage. The complete skill quantification pipeline based on spatiotemporal heatmap encoding is illustrated in Figure 1. First, a high-resolution pose estimation network is employed to extract two-dimensional coordinates of K human keypoints from each frame of the input video. The continuous motion trajectories are then transformed into standardized two-dimensional heatmap image representations. For the k-th keypoint, the heatmap trajectory image $H_k \in \mathrm{R}^{T\times L}$ is defined, with pixel values computed via a Gaussian kernel function:

$$H_k(t,l)=exp\left(-\frac{(d_{k,t}-\mu_l)^2}{2\sigma^2}\right) \tag{1}$$

where, T denotes the total number of video frames; L denotes the number of discretized bins for spatial positions or joint angles; $d_{k,t}$ denotes the spatial coordinate or computed joint angle value of the k-th keypoint at frame t; μ_l denotes the center value of the l-th bin; and σ denotes the Gaussian kernel width. Through this encoding scheme, the one-dimensional temporal sequence of each individual keypoint is mapped to a two-dimensional spatiotemporal image, in which pixel intensity corresponds to the probability density of the keypoint's presence at the corresponding spatiotemporal location. The complete spatiotemporal distribution pattern of the motion is thereby preserved.

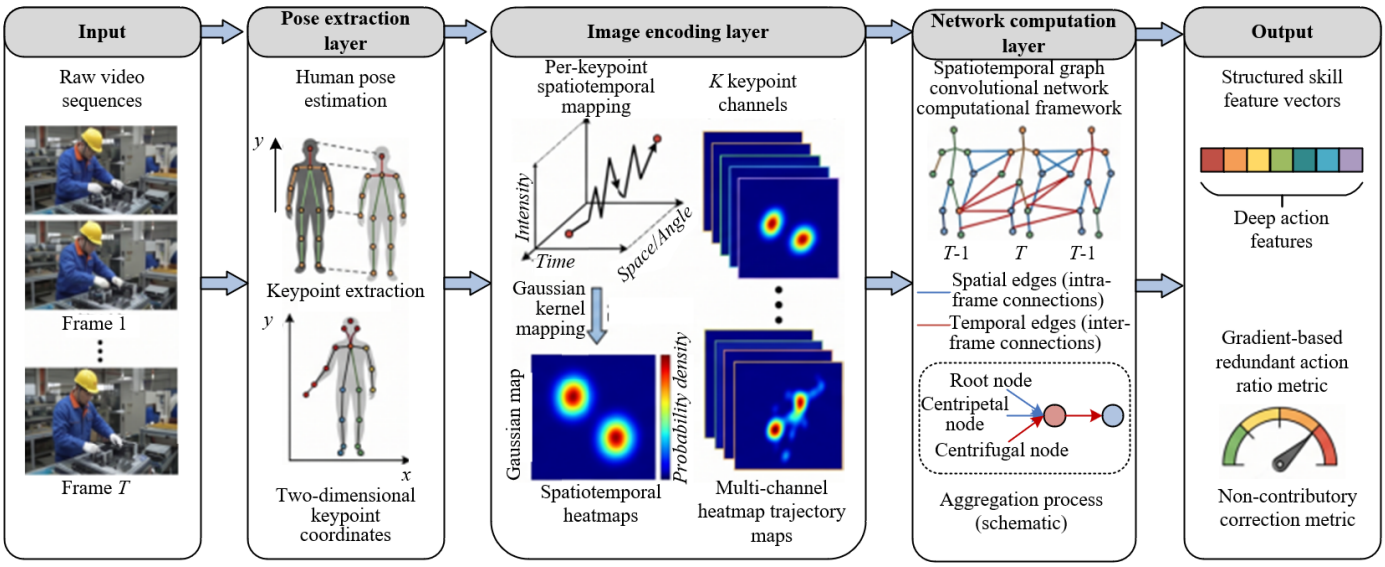


Figure 1. Flowchart of skill quantification based on spatiotemporal heatmap encoding

The heatmap trajectories of all keypoints are stacked along the channel dimension to form a multi-channel heatmap trajectory map $H \in \mathrm{R}^{T\times L\times K}$, which serves as the unified input for subsequent feature extraction. To simultaneously incorporate the topological constraints of the human skeleton and the temporal evolution patterns of the motion, the multi-channel heatmap trajectory map is formulated as a spatiotemporal graph structure. Graph nodes correspond to individual human keypoints, and the edge set is partitioned into two categories: spatial edges and temporal edges. Spatial edges are constructed according to the natural connectivity of the human skeleton, capturing the spatial relationships among keypoints within the same frame. Temporal edges connect corresponding keypoints across adjacent frames, capturing the temporal evolution of each individual keypoint. On this basis, a spatiotemporal graph convolutional network is employed to perform multi-layer feature extraction. The node feature update rule for the l-th convolutional layer is given by:

$$f^{(l+1)}(v)=\sum_{u\in N(v)}\frac{1}{Z_{v,u}}f^{(l)}(u)\cdot W^{(l)}(label(u,v)) \tag{2}$$

where, v denotes the current target node; $N(v)$ denotes the spatiotemporal neighborhood set of node v; $Z_{v,u}$ denotes the neighborhood normalization coefficient; and $W^{(l)}$ denotes the learnable weight matrix corresponding to the neighborhood type. The function $label(u,v)$ is used to partition the neighboring nodes into three categories—root nodes, centripetal nodes, and centrifugal nodes—to accommodate the physical characteristics of human skeletal motion. Through multi-layer spatiotemporal convolutions, spatial topological information and temporal contextual information are progressively aggregated, enabling the extraction of deep action features.

Following multi-layer convolutional operations and global average pooling, a D-dimensional skill feature vector s_i is

generated as the final output, corresponding to the comprehensive skill representation of the i-th employee. The individual dimensions of this vector implicitly correspond to skill attributes such as motion standardization, rhythm fluency, and operational precision. To further characterize the proportion of non-contributory action during the operation, a gradient-based redundant action ratio metric is introduced, through which the density of gradient discontinuities in the heatmap images is used to quantify the extent of corrective adjustments and non-contributory motion. The redundant action ratio is computed as:

$$Redundancy_i = \frac{1}{T \cdot L} \sum_{t,l} \| \nabla H_i(t,l) \|_2 \cdot 1_{\| \nabla H_i(t,l) \|_2 > \tau} \quad (3)$$

where, $\nabla H_i(t,l)$ denotes the gradient operator applied to the heatmap image of the i-th employee at location (t,l); τ denotes the gradient magnitude threshold; and the indicator function is used to select regions where the gradient magnitude exceeds the threshold. A higher value of this metric indicates a greater proportion of non-contributory corrective adjustments and redundant actions during the operation, corresponding to lower overall operational proficiency. The skill feature vectors of all employees, together with their corresponding redundancy metrics, collectively constitute the node set of the skill spectrum, completing the transformation from raw video to structured skill representations and providing a unified input foundation for subsequent skill-task association modeling.

2.3 Skill-task association graph generation via multi-scale cross-attention

Skill-task association modeling serves as the critical bridge connecting skill quantification results to task allocation decisions. Conventional matching approaches, which rely on manually predefined skill dimensions and task requirements, are incapable of capturing the substantial latent association information embedded within operational processes, and incur prohibitively high manual costs when adapting to new tasks. To construct a comprehensive visual representation of tasks, complementary features are extracted at three distinct scales—microscopic, mesoscopic, and macroscopic—enabling multi-dimensional characterization of operational details, semantic patterns, and temporal regularities. The complete multi-scale feature fusion and skill-task association modeling pipeline is illustrated in Figure 2. At the microscopic scale, the focus is placed on local operational details. Local binary patterns and the histogram of oriented gradients are employed to capture fine-grained information such as workpiece surface textures and operational edge trajectories. The local binary pattern descriptor, which encodes local texture structures through grayscale differences between neighboring pixels, is computed as:

$$LBP_{P,R}(x_c,y_c) = \sum_{p=0}^{P-1} sgn(g_p - g_c) 2^p, sgn(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \quad (4)$$

where, g_c denotes the grayscale value of the center pixel; g_p denotes the grayscale value of the p-th sampling point within a circular neighborhood of radius R; and P denotes the total number of neighborhood sampling points. The two types of hand-crafted features are concatenated to form a d_1-dimensional microscopic feature vector, which precisely reflects task attributes related to operational precision and procedural detail.

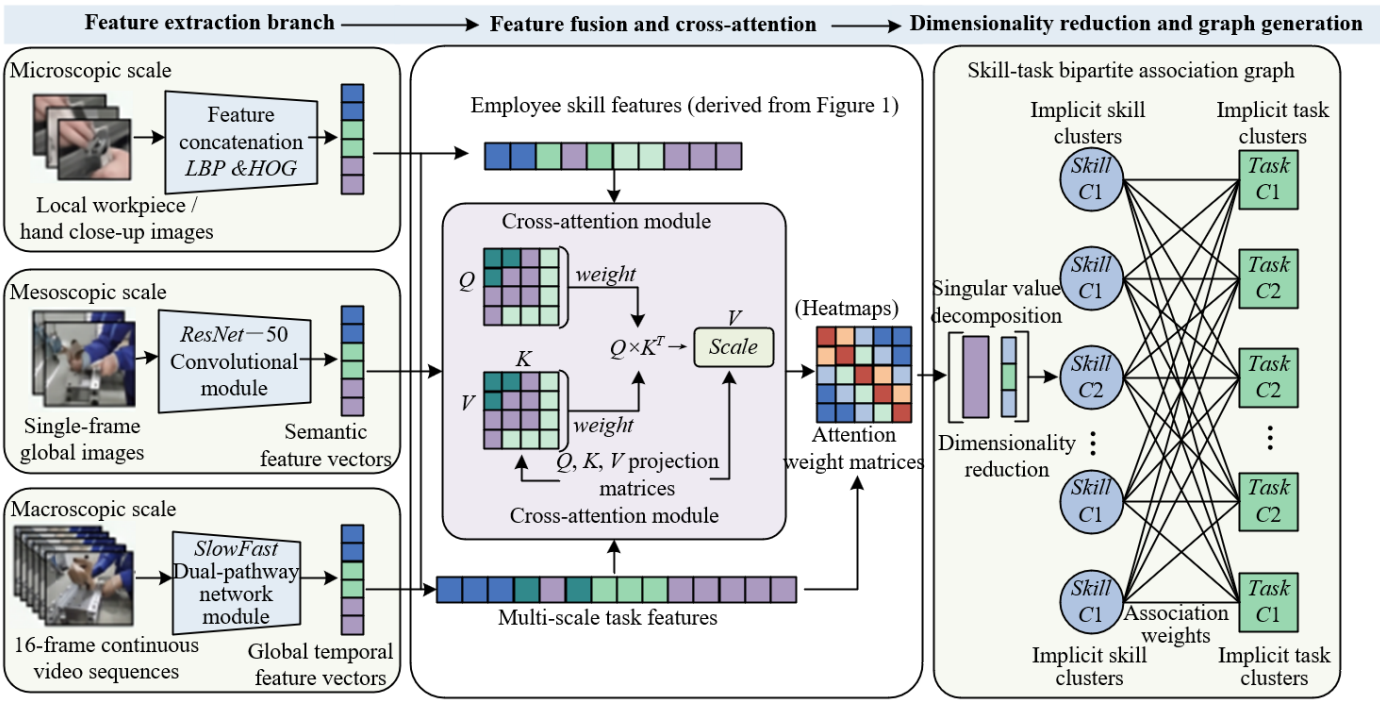


Figure 2. Schematic diagram of multi-scale feature fusion and skill-task association modeling

At the mesoscopic and macroscopic scales, the task representation is supplemented from semantic and temporal perspectives, respectively. At the mesoscopic scale, convolutional layer features are extracted using a pre-trained ResNet-50 network, and global average pooling is applied to obtain a 2048-dimensional semantic feature vector, which characterizes mid-level semantic patterns of operational scenes and action phases, reflecting the task type and core

workflow. At the macroscopic scale, a SlowFast network is employed to process 16-frame continuous video segments, fusing the spatial semantic information from the slow pathway with the motion temporal information from the fast pathway to obtain a 2048-dimensional global temporal feature vector, which characterizes the overall rhythm and motion evolution patterns of the task. For each task, the three types of features extracted from all sampled frames are averaged separately and then concatenated along the dimension to form a multi-scale fused feature vector of dimension d_1+4096 . This representation comprehensively covers the multi-level visual attributes of the task—from local details to global processes—providing a unified feature foundation for subsequent association modeling.

To establish latent mappings between employee skill features and task features, a cross-attention mechanism is employed to compute the association weights between the two feature sets. Let the employee skill matrix be denoted as $S\in \mathrm{R}^{N\times D}$, where N is the total number of employees and D is the skill feature dimension, and let the task feature matrix be denoted as $T\in \mathrm{R}^{J\times(d_1+4096)}$, where J is the total number of tasks. Through learnable projection matrices, the two types of features are mapped into a common latent space, and the cross-attention weight matrix $A\in \mathrm{R}^{N\times J}$ is computed as:

$$A_{i,j}=softmax_j\left(\frac{(s_iW_Q)(t_jW_K)^\top}{\sqrt{d_k}}\right) \tag{5}$$

where, $W_Q\in \mathrm{R}^{D\times d_k}$ and $W_K\in \mathrm{R}^{(d_1+4096)\times d_k}$ denote the query projection matrix and the key projection matrix, respectively; d_k denotes the latent space dimension; and $\sqrt{d_k}$ serves as the scaling factor to stabilize the training process. After row-wise softmax normalization, the value of $A_{i,j}$ corresponds to the skill-task compatibility between the i-th employee and the j-th task. Through the attention mechanism, the contributions of different feature dimensions are automatically weighted, enabling the discovery of latent associations that cannot be captured by manually defined rules.

To enhance the interpretability of the association results and reduce the subsequent computational complexity, singular value decomposition is applied to the cross-attention weight matrix to extract a low-rank latent cluster structure. The decomposition is formulated as:

$$A=U\Sigma V^\top \tag{6}$$

where, $U\in \mathrm{R}^{N\times R}$ denotes the latent skill cluster representation; $V\in \mathrm{R}^{J\times R}$ denotes the latent task cluster representation; R denotes the reduced latent space dimension satisfying $R\ll\min(N,J)$; and Σ denotes the diagonal singular value matrix. Based on these two latent representations, a skill-task bipartite association graph is constructed, in which nodes on the two sides correspond to skill clusters and task clusters, respectively, and edge weights are determined by the corresponding elements of $U^\top V$, reflecting the association strength between the two types of clusters. This graph provides an intuitive visualization of the intrinsic association structure between skill dimensions and task requirements. When combined with t-distributed stochastic neighbor embedding dimensionality reduction for visualization, the graph can further support management decision analysis, while the low-rank representation simultaneously provides an efficient computational foundation for the subsequent task allocation algorithm.

2.4 Skill prediction via temporal contrastive learning and bi-objective allocation optimization

Static task allocation approaches, which perform matching based solely on the current skill state without incorporating the temporal evolution patterns of employee skills, are incapable of supporting the coordinated optimization of human resource planning and skill development from a long-term perspective. The complete framework for temporal skill prediction and bi-objective task allocation is illustrated in Figure 3. To integrate the skill evolution process into a unified visual computing framework, the temporal skill representations are reconstructed into image formats, transforming the problem into a standard image sequence learning task. Let the employee skill matrix at the τ-th time step be denoted as:

$$S^{(\tau)}\in \mathrm{R}^{N\times D} \tag{7}$$

where, N is the total number of employees, and D is the skill feature dimension. For the i-th employee, the skill vector is reshaped according to the correspondence between skill categories and sub-dimensions to form a two-dimensional skill image $G_i^{(\tau)}\in \mathrm{R}^{d_h\times d_w}$, satisfying $d_h\times d_w=D$, where the rows and columns correspond to different hierarchical levels of skill attributes, thereby preserving the topological associations among dimensions. After stacking along the temporal dimension, the skill evolution process of a single employee is represented as the image sequence $\{G_i^{(\tau)}\}_{\tau=1}^{T}$, where T is the total number of time steps. This representation format is directly compatible with established image sequence prediction and self-supervised learning methods.

Based on the skill image sequences described above, temporal contrastive learning is employed to pre-train the skill embeddings, thereby enhancing the representational consistency and robustness of long-horizon temporal predictions. The training process involves two learnable networks: an encoder f_θ and a predictor g_ϕ. For the same employee, skill images from two different time steps are selected to form positive sample pairs, while skill images from different employees are selected to form negative sample pairs. Through the contrastive loss, the temporal embeddings of the same employee are constrained to remain proximate in the feature space. The contrastive loss function is defined as:

$$L_{contrast}=-\log\frac{exp\left(sim(z_i^{(\tau_1)},z_i^{(\tau_2)})/\eta\right)}{\sum_{i'\neq i}exp\left(sim(z_i^{(\tau_1)},z_{i'}^{(\tau_2)})/\eta\right)} \tag{8}$$

where, $z_i^{(\tau)}=f_\theta(G_i^{(\tau)})$ denotes the skill image embedding output by the encoder, $sim(\ ,\ )$ denotes the cosine similarity function, and η denotes the temperature hyperparameter used to adjust the concentration of the feature distribution. After pre-training, the encoder parameters are fixed, and a prediction network consisting of a single-layer gated recurrent unit and a decoder is constructed. The historical skill embeddings from L consecutive time steps are input to the network, and the predicted embedding for the Δ-th future time step is output. The predicted skill image $\widehat{G}_i^{(\tau+\Delta)}$ is then reconstructed through the decoder. This computation process is formulated as:

$$\hat{z}_i^{(\tau+\Delta)}=g_\phi(GRU(\{z_i^{(\tau-L+1)},\ldots,z_i^{(\tau)}\})) \quad (9)$$

After the predicted skill image is mapped back to vector form, it is subtracted element-wise from the target skill requirements corresponding to the task, yielding the skill deficit vector δ_i for the employee, where positive values indicate skill dimensions in which the employee currently exhibits deficiencies.

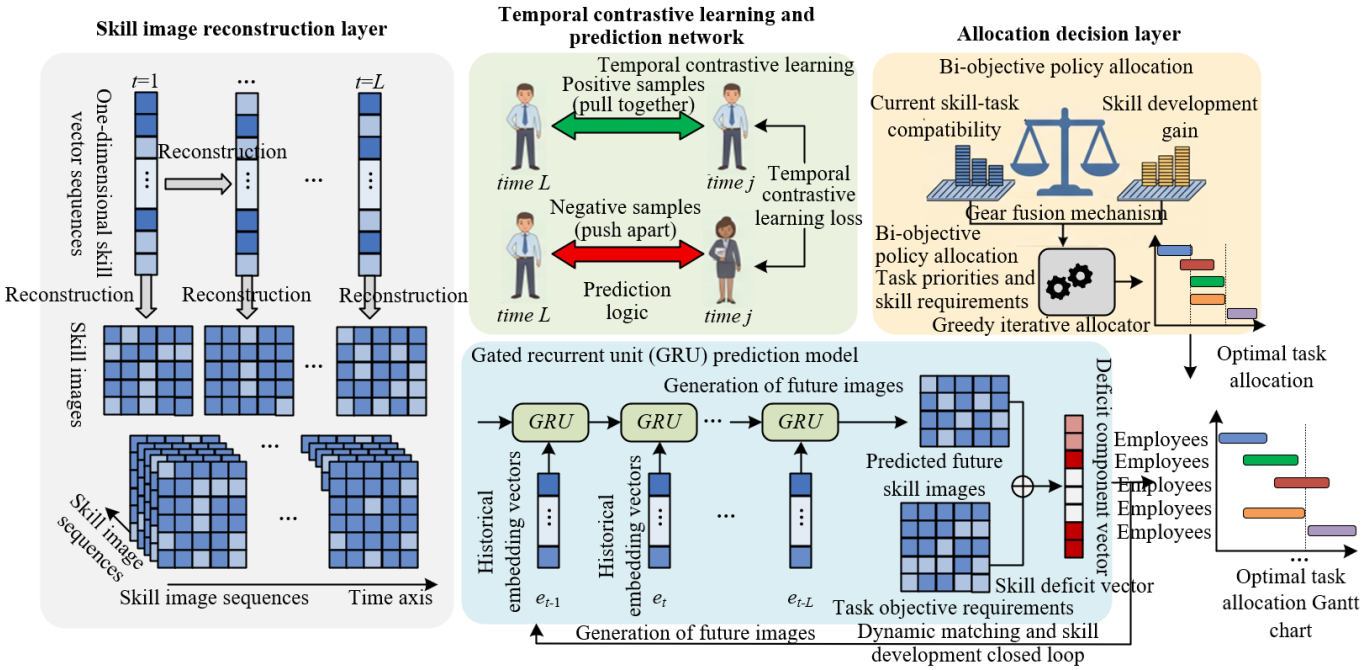


Figure 3. Framework diagram of temporal skill prediction and bi-objective task allocation

Based on the current skill state and the future skill deficit predictions, a bi-objective greedy task allocation strategy is constructed to achieve the joint optimization of task execution efficiency and employee skill development. Let the set of pending tasks be denoted as $J_{pending}$, and let the skill requirement vector r_j for the j-th task be obtained through mapping from the previously constructed skill-task association graph. The comprehensive matching score between the employee and the task is defined as:

$$Score(i,j)=s_i^\top r_j+\lambda \sum_{d:\delta_{i,d}>0} \delta_{i,d} \cdot r_{j,d} \quad (10)$$

where, the first term denotes the current skill-task compatibility, computed as the inner product between the employee's current skill vector and the task requirement, reflecting the immediate efficiency of task execution. The second term denotes the skill development gain, which is accumulated only over skill dimensions in which the employee exhibits deficits, reflecting the developmental value of task execution for the employee's deficient skill dimensions. λ denotes a balancing hyperparameter used to adjust the weight between immediate efficiency and long-term skill development. The allocation process employs a greedy iterative strategy, in which the unassigned employee-task pair with the highest comprehensive score is selected for matching in each round, until all pending tasks have been assigned. This strategy simultaneously ensures task delivery efficiency while directionally reinforcing employees' skill deficits. Through this mechanism, a closed-loop management system is established—encompassing data acquisition, skill modeling, predictive allocation, and capability development—that satisfies the long-term human resource operational requirements of digitally transforming organizations.

3. EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

Five progressive experiments were conducted to systematically validate the various performance aspects of the proposed framework. These experiments sequentially covered five evaluation dimensions: skill quantification accuracy, association mining effectiveness, evolution prediction capability, allocation optimization gain, and cross-domain generalization. For all experiments, a validation strategy was adopted in which each experiment was repeated with multiple runs, and the results were averaged to ensure statistical reliability.

3.1 Experimental setup

Two datasets were employed for performance validation. The first dataset was a self-constructed industrial assembly skill dataset, comprising 32 frontline employees with varying skill levels performing eight typical assembly tasks, including bolt fastening, component insertion, and cable routing. Each operation video ranged from 30 to 120 seconds in duration, captured at 30 frames per second with both top-down global views and first-person perspectives simultaneously recorded. Annotation was performed independently by five senior process engineers, encompassing overall skill scores, sub-dimensional scores across three dimensions (operational standardization, fluency, and precision), task completion times, and skill dimension labels. The mean values across multiple annotators were adopted as the ground truth, which served as the benchmark for core performance validation. The second dataset is the publicly available JIGSAWS surgical skill dataset, containing laparoscopic surgical operation videos

from eight operators across three procedure types, along with corresponding expert-assigned comprehensive skill scores. This dataset was employed for cross-domain generalization validation to assess the adaptability of the proposed method to non-industrial operational scenarios.

For each module, mainstream methods in the respective fields were selected as comparative baselines. For the skill quantification module, four baselines were selected: keypoint coordinates combined with a support vector machine, an inflated 3D video understanding network, a standard graph convolutional network, and a spatiotemporal graph convolutional network without heatmap encoding. For the association modeling module, three baselines were selected: single-scale features combined with cosine similarity, non-negative matrix factorization, and standard self-attention. For the prediction module, four baselines were selected: a standalone gated recurrent unit, a long short-term memory network, a Transformer, and a gated recurrent unit without contrastive learning pre-training. For the allocation module, three baselines were selected: the Hungarian algorithm, a pure compatibility-based greedy allocation, and random allocation.

All experiments were implemented using the PyTorch framework and executed on a single NVIDIA A100 graphics processing unit. Pose estimation was performed using a publicly available pre-trained high-resolution pose estimation network. The spatiotemporal graph convolutional network was configured with 8 layers of spatiotemporal convolutions, and the output skill feature dimension was set to 128. For contrastive learning pre-training, the Adam optimizer was employed with an initial learning rate of 1e-4, trained for 50 epochs, with the temperature parameter set to 0.07. For the task allocation experiments, 10 iterative rounds were simulated to represent long-term operational scenarios, with 20 tasks allocated and skill states updated in each round.

3.2 Skill quantification performance validation

In this experiment, the skill quantification accuracy and fine-grained representation capability of the proposed heatmap trajectory encoding combined with the spatiotemporal graph convolutional network scheme were validated. Comparisons were conducted against various baseline methods on the industrial assembly skill dataset, and ablation experiments were performed to verify the contribution of each core component. The skill quantification performance comparison across different methods is presented in Table 1.

Table 1. Skill quantification performance comparison across different methods on the industrial assembly skill dataset

Method	Overall Pearson Correlation Coefficient	Overall Spearman Rank Order Correlation Coefficient	Overall Mean Absolute Error	Pearson Correlation Coefficient of Standardization	Pearson Correlation Coefficient of Fluency	Pearson Correlation Coefficient of Precision
Keypoint coordinates + support vector machine	0.712	0.685	0.093	0.734	0.701	0.658
Inflated 3D	0.786	0.759	0.078	0.802	0.775	0.741
Standard graph convolutional network	0.803	0.781	0.072	0.815	0.794	0.767
Spatiotemporal graph convolutional network without heatmap encoding	0.859	0.842	0.057	0.867	0.851	0.833
Proposed method	0.921	0.908	0.047	0.913	0.905	0.927

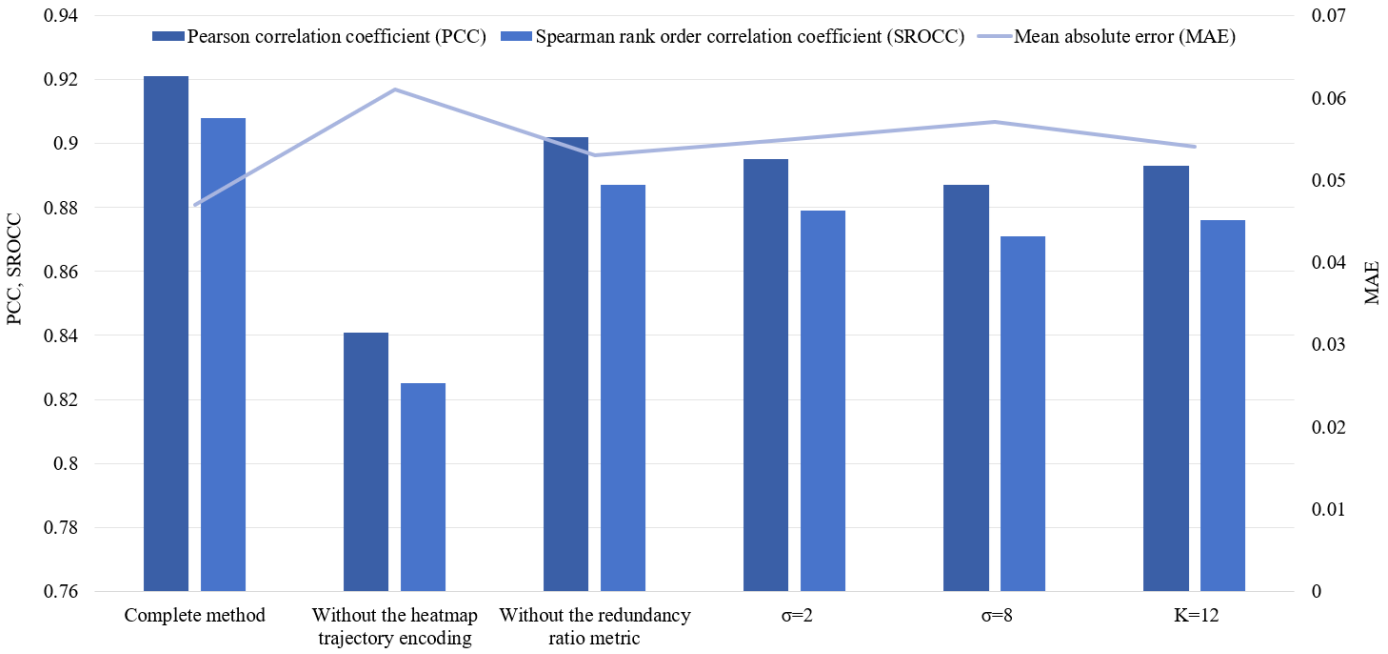


Figure 4. Ablation study results for the skill quantification module

From the results, it can be observed that the proposed method achieves optimal performance across both overall and all sub-dimensional metrics. The overall Pearson correlation coefficient reaches 0.921, the Spearman rank order correlation coefficient reaches 0.908, and the mean absolute error is reduced to 0.047. Compared with the sub-optimal baseline—

the spatiotemporal graph convolutional network without heatmap encoding—the correlation coefficient is improved by 7.2%, while the absolute error is reduced by 18.1%. Among the sub-dimensional metrics, the most significant improvement is observed in the precision dimension, indicating that the heatmap trajectory encoding provides stronger characterization of spatial position details and enables more accurate capture of skill features related to operational precision. In comparison with coordinate-based methods, the image-based encoding fully preserves the spatiotemporal distribution information of the motion, effectively enhancing the quantification accuracy of fine-grained skill dimensions.

To further verify the contribution of each component, ablation experiments were conducted, with results presented in Figure 4. The most significant performance degradation was observed when the heatmap trajectory encoding was removed, with the Pearson correlation coefficient decreasing by 8.7%, validating the core contribution of image-based encoding to skill representation. When the redundancy ratio metric was removed, an increase in the mean absolute error was observed, indicating that this metric provides supplementary information regarding non-contributory motion dimensions, thereby enhancing the comprehensiveness of the skill representation. Results from experiments varying the Gaussian kernel width and the number of keypoints indicate that the model performance remains within the optimal range under the default parameter configuration. Both excessively large and excessively small kernel widths were found to compromise the precision of trajectory distribution characterization, while a reduction in the number of keypoints resulted in the loss of limb-level detail information. Visualization analysis revealed distinct patterns between skill levels. For high-skill employees, the heatmap trajectories exhibited smooth and continuous distribution patterns with sparse gradient discontinuity regions. In contrast, for low-skill employees, the heatmap trajectories contained numerous discrete peaks and recurrent trajectory patterns, corresponding to a greater number of corrective adjustments and redundant motions. These observations are consistent with the quantitative metric results.

3.3 Skill-task association graph performance validation

In this experiment, the latent association mining effectiveness of the multi-scale feature fusion and cross-attention mechanism was validated. The retrieval performance and interpretability of the association graph were evaluated. The skill-task association performance comparison across different methods is presented in Table 2.

The results demonstrate that the proposed method achieves superior performance over all comparative baselines across all retrieval metrics. The mAP@10 reaches 0.874, and the Top-3 recall reaches 0.912, representing an 11.3% improvement in mAP@10 over the sub-optimal standard self-attention baseline. The performance of single-scale feature methods is generally lower, with the method relying exclusively on microscopic features yielding the lowest recall, indicating that single-scale features are insufficient to comprehensively cover the multi-level attributes of tasks, thereby failing to establish accurate association mappings. The non-negative matrix factorization method, which relies on a linear mapping assumption, exhibits limited capacity for capturing complex latent associations, and its performance is inferior to that of the attention-based method.

Table 2. Skill-task association performance comparison across different methods

Method	mAP@5	mAP@10	Top-1 Recall	Top-3 Recall	Top-5 Recall
Single-scale microscopic features + cosine similarity	0.523	0.561	0.425	0.637	0.721
Single-scale mesoscopic features + cosine similarity	0.617	0.654	0.512	0.719	0.803
Non-negative matrix factorization	0.682	0.715	0.588	0.784	0.856
Standard self-attention	0.746	0.785	0.653	0.841	0.902
Proposed method	0.839	0.874	0.756	0.912	0.953

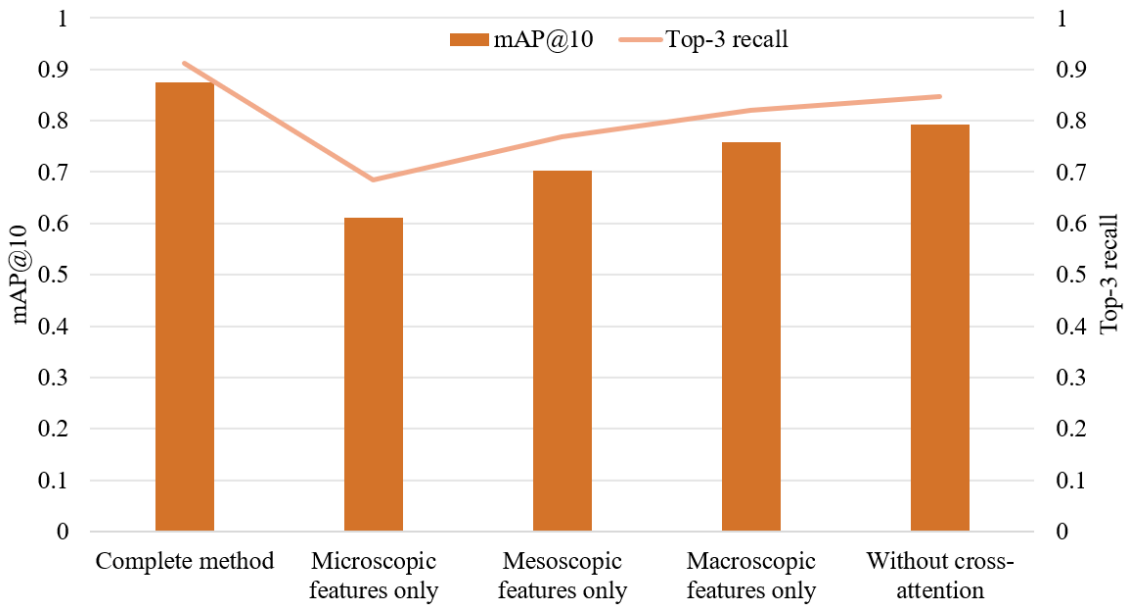


Figure 5. Ablation study results for multi-scale features and attention mechanism

The ablation study results for multi-scale features are presented in Figure 5. Removing any single scale of features resulted in performance degradation, with the largest decrease observed when the macroscopic temporal features were removed, indicating that the temporal motion patterns of tasks serve as the core basis for association modeling. After the removal of the cross-attention mechanism, the mAP@10 decreased by 9.4%, validating the critical role of cross-attention in heterogeneous feature alignment and latent association mining. Through the bipartite association graph obtained via singular value decomposition, three core skill clusters and their corresponding task clusters were clearly identified. Among these, the fine-manipulation skill cluster was found to correspond to high-precision assembly tasks, while the gross-motor skill cluster corresponded to heavy component assembly tasks. These findings are highly consistent with practical business domain knowledge, demonstrating strong interpretability of the association graph.

3.4 Skill deficit prediction performance validation

In this experiment, the improvement effect of temporal contrastive learning on skill evolution prediction was validated. The prediction performance and parameter sensitivity across different prediction horizons were evaluated. The prediction performance comparison across different methods and horizons is presented in Table 3.

Across all prediction horizons, the proposed method consistently achieves significantly superior prediction accuracy compared to all baseline methods. For 1-step prediction, the Pearson correlation coefficient reaches 0.895 with a mean absolute error of 0.052. For 5-step prediction, the correlation coefficient remains at 0.783, representing a 14.6% improvement in the Pearson correlation coefficient and a 21.3% reduction in mean absolute error compared with the gated recurrent unit without contrastive learning pre-training. The performance of pure temporal models exhibits substantial degradation under long-horizon predictions. In contrast, the embeddings obtained through contrastive learning pre-training exhibit stronger semantic consistency and temporal stability, effectively mitigating the error accumulation problem inherent in long-sequence prediction. The Transformer model achieves performance comparable to the proposed method at short horizons, but demonstrates insufficient generalization capability at longer horizons, while also incurring higher computational overhead.

The parameter sensitivity analysis results for the temporal window length are presented in Figure 6. Optimal prediction performance was achieved when the window length was set to 7. When the window was too short, the complete skill evolution cycle could not be adequately covered; conversely, when the window was excessively long, redundant historical information was introduced, which interfered with the prediction results. The optimal window length for both 3-step and 5-step predictions was found to be consistent, indicating that this parameter configuration exhibits good generalizability. The skill evolution curves of representative employees demonstrate that the predicted trajectories generated by the proposed method are highly consistent with the actual evolution trends. Skill deficits in dimensions such as precision were identified in advance, providing data-driven support for forward-looking human resource planning.

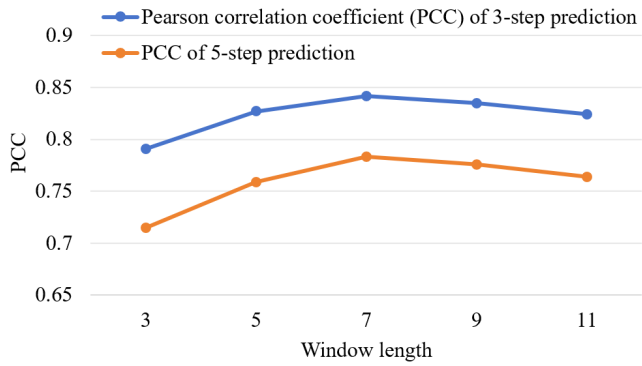


Figure 6. Performance impact of the temporal window length

3.5 Bi-objective task allocation performance validation

In this experiment, the comprehensive advantages of the bi-objective greedy allocation strategy in terms of execution efficiency and skill development were validated. The long-term operational performance across different allocation strategies was compared. The comprehensive performance comparison after 10 iterative rounds is presented in Table 4.

The results demonstrate that the Hungarian algorithm and the pure compatibility-based greedy allocation achieve the shortest average task completion times; however, their skill improvement rates are significantly lower, amounting to only approximately 42% of that achieved by the proposed method. The proposed bi-objective allocation strategy achieves an average task completion time of 187.2 seconds, which is only 3.8% slower than the efficiency-optimal Hungarian algorithm, while attaining an average skill improvement rate of 12.7% and achieving the highest comprehensive utility among all strategies. Random allocation yields the worst performance across all metrics. These findings indicate that the bi-objective strategy can significantly accelerate employee skill development with virtually no loss in execution efficiency, thereby achieving a balanced trade-off between efficiency and development.

Table 3. Skill prediction performance comparison across different methods and prediction horizons

Method	1-Step Prediction		3-Step Prediction		5-Step Prediction	
	Pearson correlation coefficient	Mean absolute error	Pearson correlation coefficient	Mean absolute error	Pearson correlation coefficient	Mean absolute error
Standalone gated recurrent unit	0.782	0.071	0.654	0.093	0.561	0.112
Long short-term memory	0.795	0.068	0.673	0.089	0.584	0.107
Transformer	0.831	0.062	0.712	0.082	0.617	0.101
Gated recurrent unit without contrastive learning	0.817	0.065	0.695	0.085	0.683	0.094
Proposed method	0.895	0.052	0.842	0.063	0.783	0.074

Table 4. Comprehensive performance comparison across different allocation strategies (after 10 iterative rounds)

Allocation Strategy	Average Task Completion Time (s)	Average Skill Improvement Rate (%)	Allocation Balance	Comprehensive Utility
Random allocation	241.5	3.2	0.521	0.317
Hungarian algorithm	180.3	5.3	0.684	0.582
Pure compatibility-based greedy allocation	183.7	5.7	0.672	0.596
Proposed bi-objective allocation	187.2	12.7	0.795	0.813

The sensitivity analysis results for the balancing parameter are presented in Table 5. As the parameter value increases, the average task completion time gradually rises, while the skill improvement rate initially increases and then decreases. The comprehensive utility reaches its peak when the parameter is 0.5. When the parameter is 0, the strategy degenerates into pure compatibility-based allocation, yielding only marginal skill improvement. When the parameter value is excessively large, the allocation becomes overly biased toward skill development, resulting in excessive loss of task efficiency and a consequent decline in overall utility. In practical applications, this parameter can be adjusted according to the organization's stage-specific objectives to accommodate different management requirements. Tests conducted across varying task scales demonstrate that the time complexity of the proposed method scales linearly with the number of tasks. Even in scenarios involving 100 tasks, allocation can still be completed within 0.1 seconds, demonstrating strong scalability and promising potential for practical engineering deployment.

Table 5. Performance impact of the balancing parameter λ

λ Value	Average Completion Time (s)	Average Skill Improvement Rate (%)	Comprehensive Utility
0	183.7	5.7	0.596
0.2	185.1	8.9	0.704
0.5	187.2	12.7	0.813
0.8	192.4	13.1	0.789
1	198.6	12.5	0.742

3.6 Cross-dataset generalization validation

In this experiment, the generalizability of the proposed method across different operational scenarios was validated. The core experiments for skill quantification and association modeling were reproduced on the JIGSAWS dataset, with the task and skill definitions adapted to the surgical operation context. The results are presented in Table 6.

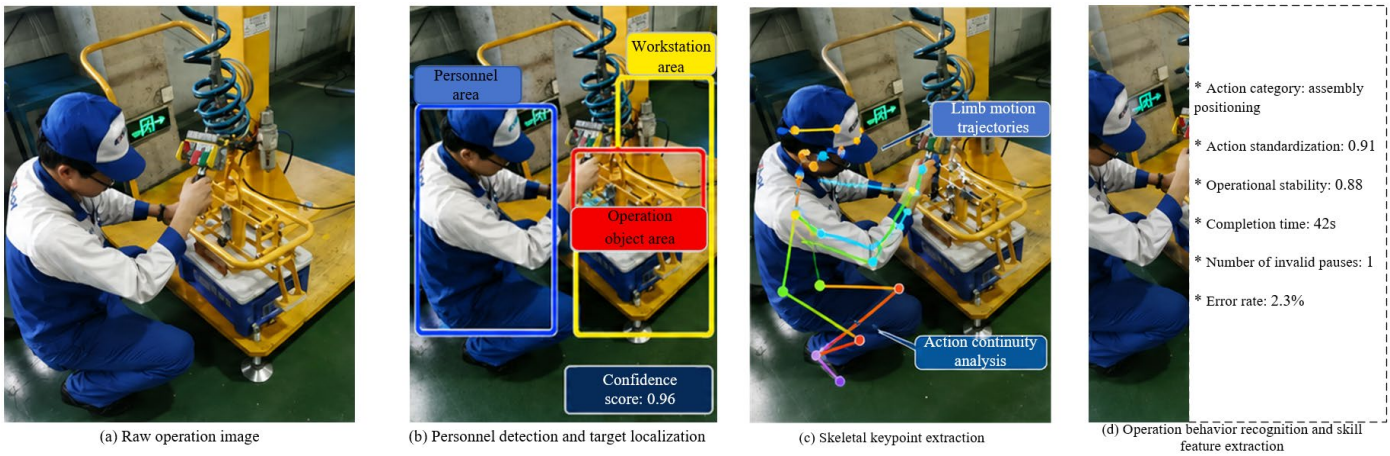
The cross-domain experimental results demonstrate that the proposed method maintains optimal performance on the JIGSAWS dataset, achieving a Pearson correlation coefficient of 0.863 for skill quantification and a task retrieval mAP@10 of 0.802, representing improvements of 6.5% and 9.7%, respectively, over the sub-optimal baseline. Compared with the results on the industrial assembly dataset, a modest performance degradation was observed, which is primarily attributable to differences in viewing angles and the higher precision requirements inherent to surgical operations. Nevertheless, the performance reduction was found to be within a reasonable range. These findings indicate that the proposed heatmap trajectory encoding and association modeling framework exhibits strong domain generalizability, and can be effectively transferred to skill management scenarios in other operational domains, such as healthcare and service industries.

Table 6. Cross-domain performance comparison on the JIGSAWS dataset

Method	Pearson Correlation Coefficient of Skill Quantification	Spearman Rank Order Correlation Coefficient of Skill Quantification	Task Retrieval mAP@10
Keypoint coordinates + support vector machine	0.674	0.648	0.527
Inflated 3D	0.752	0.726	0.634
Spatiotemporal graph convolutional network without heatmap encoding	0.81	0.793	0.731
Proposed method	0.863	0.847	0.802

To examine the perceptibility, interpretability, and translatability of the image-based recognition method to management in real-world operational scenarios, a visualization analysis of the method's implementation was conducted using typical assembly operation images. As illustrated in Figure 7, the original operation images were first used to preserve the authentic relationships among employees, workstation equipment, and operation objects coexisting within the same natural scene, demonstrating that the data sources are capable of reflecting actual operational states in unstructured production environments. Subsequently, through the localization of personnel regions, workstation regions, and operation object regions, the complex on-site images were transformed into computable operational spatial units, providing stable target boundaries for subsequent motion analysis. At the skeletal keypoint extraction stage, the system further identified the key motion nodes of the employees' upper limbs, torso, and lower limbs, with the body coordination relationships during the operation being represented through connecting lines and trajectory annotations. This indicates that the method does not merely remain at the level of personnel identification, but is capable of deeply capturing the action structures directly related to skill performance. Furthermore, through the operation behavior recognition results, the assembly positioning process was transformed into quantifiable metrics—including action category, action standardization, operational stability, completion time, pause frequency, and error rate—thereby achieving the mapping from visual features to skill features. From these results, it can be concluded that the image recognition technology enables the continuous extraction of key features reflecting operational standardization, stability,

efficiency, and error control capability, all without interrupting the employees' workflow. This provides an objective data foundation for the construction of employee skill spectrums and furnishes interpretable decision-making evidence for subsequent task allocation optimization.



(a) Raw operation image (b) Personnel detection and target localization (c) Skeletal keypoint extraction (d) Operation behavior recognition and skill feature extraction

Figure 7. Visualization of employee skill recognition, spectrum generation, and task allocation results based on operation videos

4. DISCUSSION

Synthesizing the experimental results, the core advantages of the proposed framework can be identified at three levels: image-based representation design, end-to-end automation, and closed-loop optimization mechanisms. At the image-based representation level, the operational skill heatmap trajectory encoding transforms discrete keypoint motion sequences into structured spatiotemporal images, fully preserving the spatiotemporal distribution patterns and fine-grained trajectory features of the motion. This provides a standardized input for subsequent convolution operations. Compared with modeling approaches that directly operate on coordinate sequences, the proposed encoding enables more precise characterization of fine-grained skill attributes—such as motion standardization and operational redundancy—thereby providing a unified visual representation foundation for general-purpose skill spectrum construction. At the automation capability level, the entire pipeline—from skill feature extraction to skill-task association generation—operates with raw video data as the sole input, eliminating the need for manually predefined skill dimension labels and task matching rules. The modeling process is completed through latent association mining in the feature space, substantially reducing the adaptation costs for new tasks and new scenarios. This automation capability renders the framework well-suited for the rapidly iterative operational requirements of flexible production environments. At the application value level, the skill evolution prediction and bi-objective allocation strategy form a mutually reinforcing closed-loop mechanism. Task allocation is no longer confined to immediate efficiency optimization; rather, the skill development gain is incorporated into the optimization objective. Through this approach, employee skill development can be significantly accelerated with only controlled efficiency loss, providing a practically viable technological pathway for long-term human resource planning in digitally transforming organizations.

At the same time, three limitations of the current method should be objectively acknowledged. First, the method's performance is significantly influenced by camera viewpoints and scene occlusions. The current experiments were conducted based on clear, fixed-viewpoint images. However, real-world industrial scenarios frequently involve equipment occlusion, personnel movement, and viewpoint shifts, which can degrade pose estimation accuracy and consequently propagate errors to the skill quantification stage. The robustness of the proposed method in extreme scenarios therefore requires further improvement. Second, the multi-scale feature extraction and spatiotemporal graph convolutional network entail substantial computational overhead. The end-to-end inference of the complete pipeline relies on graphics processing unit computing power, making it difficult to deploy directly on resource-constrained edge devices. Consequently, the deployment cost for full-coverage scenarios on large-scale production lines remains high. Third, performance stability under small-sample conditions is insufficient. For newly hired employees or newly introduced task types, when the amount of historical operation data is limited, the effectiveness of contrastive learning pre-training and association modeling exhibits notable fluctuations. The adaptation capability during the cold-start phase therefore requires further enhancement.

To address the aforementioned limitations, future research can be extended along multiple directions. First, a multi-view visual fusion mechanism could be introduced, through which information from multiple camera angles could be leveraged to mitigate the impact of occlusions and viewpoint deviations, thereby enhancing the method's robustness in complex industrial scenarios. Second, the scope of skill modeling could be expanded from individual operations to team collaboration scenarios, with spatiotemporal association representation methods for multi-person interactive actions being investigated to construct team-level skill spectrums and collaborative task allocation models, thereby adapting to a broader range of team-based operational scenarios. Third, multi-modal data fusion schemes could be explored, integrating force perception data, equipment operational parameters, and production logs with visual data to further enhance the comprehensiveness of skill assessment and the accuracy of association modeling. Fourth, model lightweighting and edge deployment research could be pursued, through which model size and computational

complexity could be reduced via knowledge distillation and structural pruning, thereby facilitating the deployment of the proposed method on production line edge devices and extending the applicable boundaries of the technology.

5. CONCLUSION

To address the core challenges in digitally transforming organizations—namely, the subjectivization of skill assessment for operational roles, the reliance on task matching experience, and the static nature of workforce scheduling—a comprehensive image-driven framework for skill spectrum construction and task allocation optimization was proposed. Through this framework, organizational human resource management problems were reformulated as three standard visual computing tasks: spatiotemporal image representation, multi-scale feature association, and temporal sequence prediction. End-to-end automated modeling from raw operation videos to management decision outputs was thereby achieved. Within this framework, the motion sequences of human keypoints were transformed into multi-channel spatiotemporal heatmap representations through operational skill heatmap trajectory image encoding. Fine-grained skill features were extracted via a spatiotemporal graph convolutional network, from which structured skill spectrum nodes were constructed. Multi-scale visual features were fused, and latent mappings between skills and tasks were uncovered through a cross-attention mechanism, generating an interpretable skill-task bipartite association graph. Temporal contrastive learning was introduced to construct a skill evolution prediction model, which was integrated with a bi-objective greedy allocation strategy to achieve the joint optimization of task execution efficiency and employee skill development, thereby establishing a complete closed-loop management mechanism. Extensive comparative experiments and ablation studies conducted on a self-constructed industrial assembly dataset and the publicly available surgical skill dataset demonstrated that the proposed method consistently outperformed all baseline methods across skill quantification accuracy, association retrieval performance, and comprehensive allocation utility. Strong cross-domain generalizability was also exhibited, validating the effectiveness and versatility of the proposed technical approach.

This study establishes a deep connection between visual computing technology and organizational human resource management, transcending the inherent limitations of traditional management methods that rely heavily on expert experience. The deployable application boundaries of image representation learning have been extended, while a novel technological paradigm for refined human resource operations in digitally transforming organizations has been provided. As complementary technologies—such as multi-view information fusion and model lightweighting for edge deployment—continue to mature, the proposed framework can be further adapted to diverse scenarios, including complex industrial manufacturing, surgical skill training, and service sector competency assessment. Practical technical support for the digital management upgrade of various operation-intensive organizations can thus be furnished, with broad application prospects and significant room for future development.

ACKNOWLEDGMENT

This work was supported by the Mentorship Program of Guangzhou Huashang College (Grant No.: 2020HSDS19); and the Key Discipline of Business Administration at Guangzhou Huashang College (Grant No.: 910108005).

REFERENCES

[1] Zhang, X., Ming, X., Bao, Y. (2022). A flexible smart manufacturing system in mass personalization manufacturing model based on multi-module-platform, multi-virtual-unit, and multi-production-line. Computers & Industrial Engineering, 171: 108379. https://doi.org/10.1016/j.cie.2022.108379

[2] Pannok, M., Lier, S. (2023). Operation of modular intralogistic systems in chemical industry production networks. Chemical Engineering & Technology, 46(9): 1915-1923. https://doi.org/10.1002/ceat.202200604

[3] Pedrett, R., Mascagni, P., Beldi, G., Padoy, N., Lavanchy, J.L. (2023). Technical skill assessment in minimally invasive surgery using artificial intelligence: A systematic review. Surgical Endoscopy, 37(10): 7412-7424. https://doi.org/10.1007/s00464-023-10335-z

[4] Igaki, T., Kitaguchi, D., Matsuzaki, H., Nakajima, K., Kojima, S., Hasegawa, H., Takeshita, N., Kinugasa, Y., Ito, M. (2023). Automatic surgical skill assessment system based on concordance of standardized surgical field development using artificial intelligence. JAMA Surgery, 158(8): e231131-e231131. https://doi.org/10.1001/jamasurg.2023.1131

[5] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M. (2023). Deep learning-based human pose estimation: A survey. ACM Computing Surveys, 56(1): 1-37. https://doi.org/10.1145/3603618

[6] Liu, W., Bao, Q., Sun, Y., Mei, T. (2022). Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. ACM Computing Surveys, 55(4): 1-41. https://doi.org/10.1145/3524497

[7] Jain, H., Harit, G., Sharma, A. (2020). Action quality assessment using siamese network-based deep metric learning. IEEE Transactions on Circuits and Systems for Video Technology, 31(6): 2260-2273. https://doi.org/10.1109/tcsvt.2020.3017727

[8] Jiang, L., Qin, X., Yam, K.C., Dong, X., Liao, W., Chen, C. (2023). Who should be first? How and when AI-human order influences procedural justice in a multistage decision-making process. PLOS ONE, 18(7): e0284840. https://doi.org/10.1371/journal.pone.0284840

[9] Bankins, S., Formosa, P., Griep, Y., Richards, D. (2022). AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. Information Systems Frontiers, 24(3): 857-875. https://doi.org/10.1007/s10796-021-10223-8

[10] Wang, T., Jin, M., Li, M. (2021). Towards accurate and interpretable surgical skill assessment: A video-based method for skill score prediction and guiding feedback generation. International Journal of Computer Assisted Radiology and Surgery, 16(9): 1595-1605. https://doi.org/10.1007/s11548-021-02448-4

[11] Pan, J., Gao, J., Zheng, W. (2021). Adaptive action assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8779-8795. https://doi.org/10.1109/tpami.2021.3126534
[12] Lei, Q., Li, H., Zhang, H., Du, J., Gao, S. (2023). Multi-skeleton structures graph convolutional network for action quality assessment in long videos. Applied Intelligence, 53(19): 21692-21705. https://doi.org/10.1007/s10489-023-04613-5
[13] Li, H., Lei, Q., Zhang, H., Du, J., Gao, S. (2022). Skeleton-based deep pose feature learning for action quality assessment on figure skating videos. Journal of Visual Communication and Image Representation, 89: 103625. https://doi.org/10.1016/j.jvcir.2022.103625
[14] Afshar-Nadjafi, B. (2021). Multi-skilling in scheduling problems: A review on models, methods and applications. Computers & Industrial Engineering, 151: 107004. https://doi.org/10.1016/j.cie.2020.107004
[15] Rinaldi, M., Fera, M., Bottani, E., Grosse, E.H. (2022). Workforce scheduling incorporating worker skills and ergonomic constraints. Computers & Industrial Engineering, 168: 108107-108107. https://doi.org/10.1016/j.cie.2022.108107
[16] Du, Z., He, D., Wang, X., Wang, Q. (2023). Learning semantics-guided representations for scoring figure skating. IEEE Transactions on Multimedia, 26: 4987-4997. https://doi.org/10.1109/tmm.2023.3328180
[17] Gedamu, K., Ji, Y., Yang, Y., Shao, J., Shen, H.T. (2024). Self-supervised sub-action parsing network for semi-supervised action quality assessment. IEEE Transactions on Image Processing, 33: 6057-6070. https://doi.org/10.1109/tip.2024.3468870
[18] Chen, J.C., Chen, Y., Chen, T., Lin, Y. (2022). Multi-project scheduling with multi-skilled workforce assignment considering uncertainty and learning effect for large-scale equipment manufacturer. Computers & Industrial Engineering, 169: 108240. https://doi.org/10.1016/j.cie.2022.108240
[19] Borgonjon, T., Maenhout, B. (2023). The impact of dynamic learning and training on the personnel staffing decision. Computers & Industrial Engineering, 187: 109784. https://doi.org/10.1016/j.cie.2023.109784
[20] Guan, Z., Yang, J., Yang, Y., Zhu, H., Li, W., Xiong, H. (2024). JobFormer: Skill-aware job recommendation with semantic-enhanced transformer. ACM Transactions on Knowledge Discovery from Data, 19(1): 1-20. https://doi.org/10.1145/3701735
[21] Xu, H., Wu, H., Ke, X., Li, Y., Xu, R., Guo, W. (2025). Quality-guided vision-language learning for long-term action quality assessment. IEEE Transactions on Multimedia, 27: 7326-7339. https://doi.org/10.1109/tmm.2025.3599078
[22] Zhou, K., Shum, H.P.H., Li, F.W.B., Zhang, X., Liang, X. (2025). PHI: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. IEEE Transactions on Image Processing, 34: 3718-3732. https://doi.org/10.1109/tip.2025.3574938