Journal of Engineering, Project, and Production Management
2026, 16(3), 2026-199

Assessing Vocational Students' Gemstone Operation Using Internet of Things and ST-GCN Graph Modeling

Zheng Zhou
Lecturer, Anhui Technical College of Industry and Economy, Hefei, 230051, China, E-Mail: zhouzhengzzz22@163.com

Project Management
Received February 7, 2026; revised March 20, 2026; accepted April 18, 2026
Available online April 25, 2026

Abstract: A skeleton spatiotemporal graph-based modeling and training skill evaluation management system for the industrial Internet of Things (IoT) is proposed to address subjectivity and the lack of quantitative analysis in skill evaluation during vocational jade processing training. Firstly, a perceptual environment that integrates vision and inertia is constructed, and an improved High Resolution Network (HRNet) model that combines L-Basicblock and Convolutional Block Attention Module (CBAM) is proposed to address hand occlusion and extract skeletons. Subsequently, a skill assessment model is designed that combines the improved Spatial Temporal Graph Convolutional Network (ST-GCN) with the Gated Recurrent Unit (GRU). The system uses joint spatial position deviation in motion prediction and the ergonomic angle error of the hand as evaluation criteria, forming a quantifiable skill-scoring system. The results show that the Average Precision (AP) of the improved HRNet model reaches 92.8%, achieving a balance between accuracy and efficiency. The average joint position error of the evaluation model in the 400ms prediction time domain is as low as 28.1mm, which is significantly better than the 38.6mm of the Simple Multi-Layer Perceptron (SiMLPe) model. In addition, the median scores for the model in experts and novices are 95.5 and 54.0, respectively, with clear hierarchical discrimination. In actual teaching verification, the system sets a score decision threshold, allowing teachers to intuitively locate student's specific deduction points, effectively reducing teaching communication costs and consumables consumption. The system proposed by the research achieves objective quantification of gemstone operation behavior, which not only provides effective support for standardized skill certification and teaching decision optimization of precision manufacturing majors in higher vocational education, but also provides scalable management tools for the production preparation of high-quality labor under the background of Industry 4.0.

Keywords: Gemstone operation, ST-GCN, GRU, HRNet, higher vocational talents, wireless perception, skill assessment, workforce training, vocational education management, digital manufacturing training.

DOI 10.32738/JEPPM-2026-199

1. Introduction

Gemstone processing is a skill that integrates artistic aesthetics and complex coordination, and it places extremely high demands on the operator's skill level (Huang et al., 2024; Lei et al., 2023). For a long time, gemstone talent training in higher vocational colleges has followed the traditional master-apprentice model, and evaluation has been based on teachers subjective observations. This model lacks quantitative analysis of the operating process and is difficult to meet the talent training needs of modern vocational education (Jiang et al., 2024). With the advancement of digital transformation in vocational education, the use of perception technology to identify skill actions has become increasingly popular (Hu et al., 2024; Kathirvel et al., 2024). Early research relied mostly on traditional computer vision techniques. These methods have simple computational logic and are difficult to manage complex environments (Tsoy et al., 2024; An et al., 2024). Some methods simplify skeleton data into vector sequences and ignore the topological connection structure of human joints (Zhen et al., 2023; Kumar et al., 2024). Based on the characteristics of gemstone operations, there is an urgent demand for a more comprehensive intelligent evaluation method.

HRNet is widely used in research across various fields of society. Zhao and Yan (2024) proposed context-based C-HRNet, which enhances semantic representation by adding a stage to the original HRNet architecture. The results showed that this method outperformed the original HRNet in quantification. Ren et al. (2023) introduced a landmark point-based HRNet approach that uses a local perceptual loss function to optimize generation quality. The results showed that this

method effectively maintained consistency in background light and shadow. Zhang et al. (2024) proposed an improved HRNet model based on uniform-sized modules to address the limited reasoning of existing networks on dedicated equipment. The results showed that the model ran 13 times faster on the neural processor. Li et al. (2023) introduced a local window self-attention mechanism to enhance HRNet's capture of semantic information. The results showed that this method reduced memory consumption and accelerated sample identification.

The Spatial-Temporal Graph Convolutional Network (ST-GCN) has attracted much attention for its ability to effectively manage skeleton topology. Wu et al. (2023) reused data across consistent spatiotemporal dimensions to improve the network's capacity. The results showed that this method saved energy by an average of 103.88 times. Li et al. (2023) proposed an ST-GCN with part-level refinement to improve information utilization. This method used physical adjacency matrices to refine deep features. Results showed that this architecture achieved the highest recognition accuracy on mainstream benchmarks. To address the tendency of skeletal action recognition methods to ignore inter-frame motion representation, Zhuang et al. (2023) proposed a temporal refinement module that introduced a temporal correlation matrix to split the graph convolution operation. The findings showed that the recognition accuracy of this framework reached 96.8%. Ling et al. (2024) proposed an ST-GCN that combined multimodal features to address the computational complexity of existing pedestrian crossing-intention prediction models. The findings demonstrated that the accuracy and inference speed of this model across multiple data sets were better than those of existing innovative approaches.

Previous research has achieved fruitful results in its respective fields, but directly applying it to the evaluation of gemstone training in higher vocational colleges still faces significant challenges. First, there is an occlusion problem in gemstone processing. Meanwhile, the existing high-precision model has many parameters, making it difficult to train on edge devices in the training workshop. Secondly, existing ST-GCN research mostly focuses on determining action categories, but it is difficult to measure the quality of actions. To address this problem, the study introduces spatiotemporal diagram modeling and a skill assessment method for gemstone operation behavior skeletons based on wireless sensing in the industrial Internet of Things (IoT). Under this framework, industrial IoT multimodal perception and ST-GCN graph modeling are not isolated technical goals, but serve as underlying data-driven engines. The system captures the spatiotemporal characteristics of complex hand-eye coordination movements in precision manufacturing, transforming abstract experience into measurable management indicators, providing objective employee skill benchmarks for vocational colleges and manufacturing enterprises. The innovation of the research lies in the construction of a multi-modal perception environment for the industrial IoT that connects humans, machines and things, and in proposing an improved HRNet model that integrates the Convolutional Block Attention Module (CBAM) and L-Basicblock, which effectively tackles the problem of capturing key points in occluding precise hand operations. On this basis, the study proposes an Attention-based Multi-scale Residual Spatial-Temporal Graph Convolutional Network (AMR-ST-GCN) network that integrates spatiotemporal key point attention, residual connections and a multi-scale Time Convolutional Network (TCN) that combines the Gated Recurrent Unit (GRU) to build an action prediction model that achieves objective scoring of students operational level.

The research's true value extends beyond merely enhancing technical precision. It also offers standardized decision-support tools tailored for vocational education management within the precision manufacturing sector. Firstly, by changing the training design and improving teaching intervention measures, the system transforms the traditional, results- and experience-based unified teaching mode into a data-driven, dynamic process management. Teachers can make real-time interventions for students based on the score decision threshold and specific joint error data, thereby improving teaching efficiency and reducing consumable consumption. Secondly, in terms of supporting ability development tracking, the system continuously records learners dynamic evaluation scores at various stages of operation, establishing an objective digital skill profile for them and achieving a full-cycle performance benchmark test from novice to expert. Finally, in terms of integration with manufacturing education management, this model transforms abstract skills into quantifiable data indicators, which is beneficial for curriculum design in the manufacturing industry.

2. Methods and Materials

2.1. Industrial IoT Wireless Sensing and Gemstone Operation Skeleton Diagram Construction

Gem and jade processing is a high-precision, hand-eye coordination operation that places extremely high demands on the operator's hand stability, force application angle, and movement consistency. The core process includes five steps: cutting, forming, plating, engraving, and polishing. Among them, the engraving and polishing processes require extremely high demands on the operator's force stability, spatial angle perception, and hand-eye coordination ability. The layout of multi-view cameras and edge nodes in the jade processing station is shown in Fig. 1.

In Fig. 1, the overall layout of the workstation shows three high-frame-rate industrial cameras deployed around the control panel, with Cam 1 capturing the operator's overall frontal operational posture. Cam 2 records the lifting angle and lateral feed of the operating arm. Cam 3 monitors the spatial interaction between the hand and the grinding table surface from a vertical perspective. The operator wears an Inertial Measurement Unit (IMU) at the wrist to collect high-frequency six-axis dynamic data, and micro-sensing nodes are deployed at the index and thumb joints. Based on the multimodal data, this study defines the accuracy of motion capture in terms of the Percentage of Correct Keypoints@0.5 (The percentage of key points correctly detected at a similarity threshold of 0.5 (i.e., 50% overlap)) (PCK@0.5) for spatial single-frame localization and the Mean Per Joint Position Error (MPJPE) for continuous temporal trajectories, thereby quantifying the extent to which the extracted virtual skeleton model reconstructs the actual physical movements. The layout of the real gemstone processing training classroom and the camera hardware used are shown in Fig. 2.

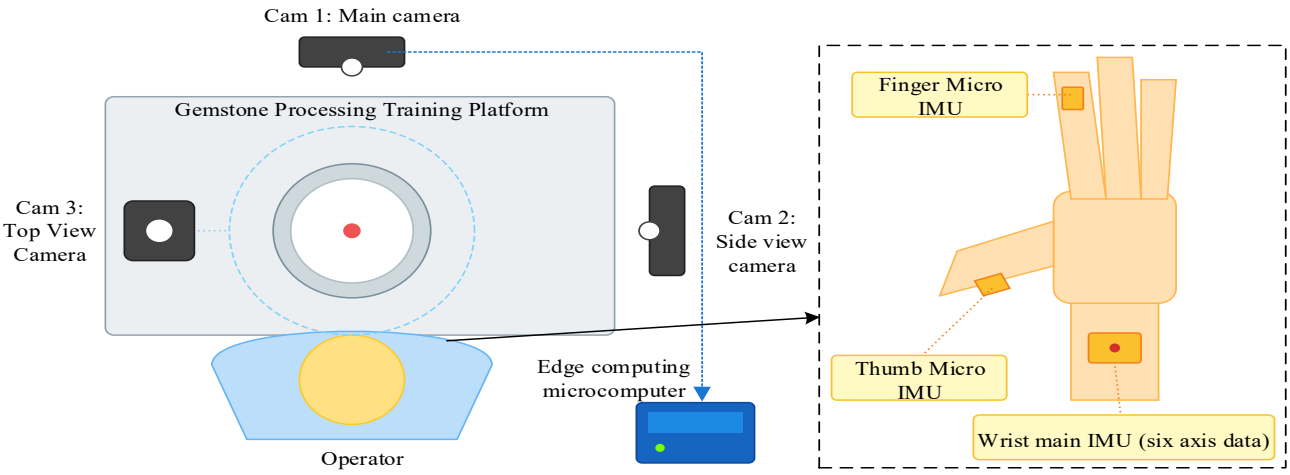


Fig. 1. Multi-angle camera and hand IMU node for gemstone processing station

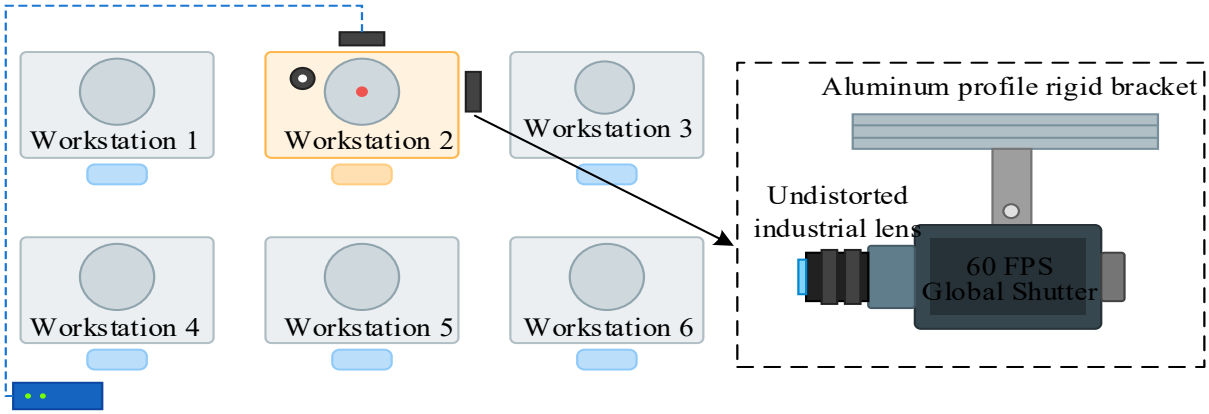


Fig. 2. Layout of training classroom and hardware of cameras used

As shown in Fig. 2, multiple standard processing stations are arranged in a matrix in the training room, each station is independently equipped with a multi-view visual acquisition system. The research selected a high frame rate global shutter industrial camera with a resolution of 1920 × 1080, an acquisition frame rate of 60 fps, and equipped with distortion-free industrial lenses. In the deployment of actual training environments, this sensing infrastructure has high economic efficiency and scalability. Compared with expensive traditional optical motion capture systems, the high frame rate industrial camera and miniature wireless IMU array used in the study have reduced the hardware cost of a single workstation by about 80%. The system adopts a cloud-edge end-to-end collaborative architecture. The training workshop only needs to configure edge computing nodes that can simultaneously process data flows from 5-8 stations through the wireless Local Area Network (LAN), to fulfill the requirements of large-scale teaching and concurrent deployment in higher vocational colleges. Under this architecture, end-side sensors collect and synchronize multimodal data via wireless protocols at high frequencies. Edge nodes perform low-latency inference with lightweight models to provide millisecond-level teaching feedback, and cloud platforms coordinate long-term storage of a digital skill archive.

To extract joint-point information from the collected video streams for the gemstone operator, the study introduced an improved HRNet model. Although the original HRNet achieves excellent performance in preserving high-resolution features, its parameter count is too high, making it difficult to run in real time on edge devices in the training workshop with limited computing power (Chen et al., 2023). In response to the urgent need for real-time feedback in higher vocational training, the study conducted targeted model optimization. First, study the use of a lightweight L-Basicblock module to replace the original Basicblock residual module in the backbone network. L-Basicblock uses 1×1 point convolution for channel compression and combines 3×3 group convolution for feature extraction. It greatly reduces the model's parameter count and Floating-Point Operations (FLOPs) while ensuring feature expression. Secondly, during gemstone processing, frequent hand-to-tool interaction makes occlusion more likely. To this end, the study integrated the CBAM module in the skeleton extraction network. The structure of the improved HRNet human pose estimation network that combines the CBAM attention mechanism and L-Basicblock is shown in Fig. 3.

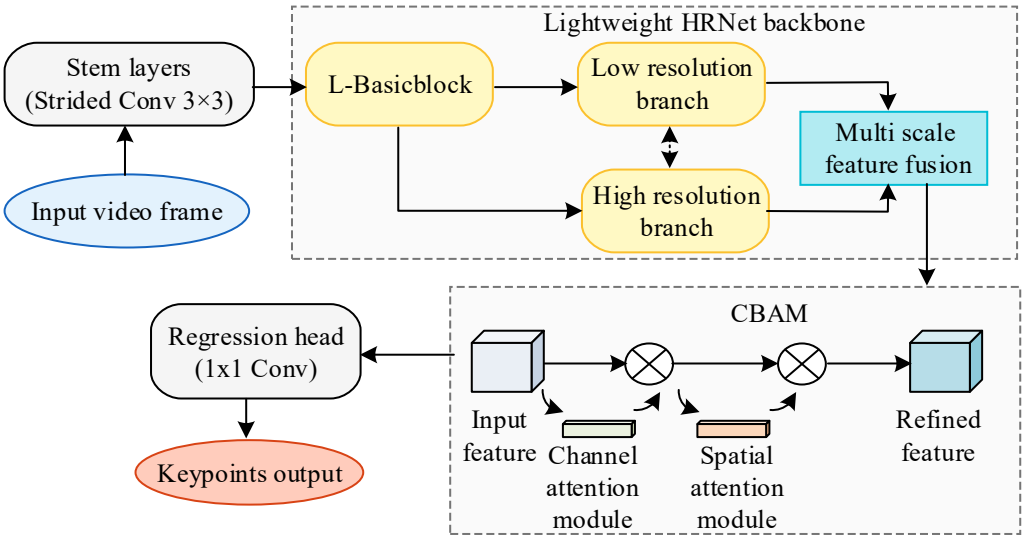


Fig. 3. Improved HRNet network architecture

In Fig. 3, CBAM uses a series connection of the channel attention module and the spatial attention module to enable the network to adaptively focus on key areas, such as hand joints and tool ends, to suppress background noise. For the input feature map (FM) F, the channel attention module first aggregates spatial information through maximum pooling and average pooling to generate a channel FM F_c, calculated as shown in Eq. (1).

$$F_c = \sigma(MLP(MaxPool(F)) \oplus MLP(AvgPool(F))) \tag{1}$$

In Eq. (1), σ is the Sigmoid activation function (AF), and $\oplus$ represents element-level addition. Subsequently, the spatial attention module focuses on the positioning of the operating area and generates a spatial FM F_s as shown in Eq. (2) by aggregating channel features.

$$F_s = \sigma(f^{7\times7}(MaxPool(F \otimes F_c); AvgPool(F \otimes F_c))) \tag{2}$$

Finally, the network outputs k heat maps containing key point information. For gemstone operations, the study defined the key point set as V including shoulder, elbow, wrist, and finger joints, and specially added tool end nodes, totaling twenty-one nodes. By calculating the mathematical expectation of the pixel value in the heat map, the coordinate (x_i, y_i) of each key point v_i is returned, thereby achieving high-precision gesture capture. After obtaining the single-frame skeleton, it needs to be converted into a spatiotemporal graph structure suitable for graph neural network processing to analyze the continuity and coordination of actions. The research defines the gemstone operation sequence as an undirected space-time graph $G = (V, E)$. The schematic diagram of the skeleton spatiotemporal model for gemstone processing operation behavior is shown in Fig. 4.

In Fig. 4, the node set contains all N key points in T frames in the video sequence. The edge set consists of two subsets, space and time, and is used to capture the spatiotemporal characteristics of actions. The spatial edges construct node connections within a single frame based on the anatomical structure of the human hand and human-computer interaction logic. Time edges connect the same joint nodes in adjacent frames to describe the evolution trajectory of actions over time. Through this graph-based modeling strategy, the complex cutting and grinding actions in gemstone processing are abstracted into signal flows on topological structures, thereby providing a standardized data input format for the skill assessment model.

2.2. Operational Skill Assessment Model based on Improved ST-GCN+GRU

After obtaining the skeleton space-time sequence of the gemstone operation, the operator's skill level needs to be evaluated. Traditional action recognition methods focus on determining action categories, whereas skill assessment requires quantitative analysis of the quality of action execution. To address this, the study constructs an action prediction and evaluation model with encoding and decoding architecture. The model consists of two parts. The front-end uses AMR-ST-GCN as the encoder, which extracts spatiotemporal skill features with high discriminability. The back-end uses GRU to build a predictive decoder and learn the temporal logic of expert data. By comparing the prediction error between the ideal expert trajectory generated by the model and the student's actual operating trajectory, skill scores are achieved. Gemstone processing actions feature keyframes and key joints. During the faceting process, the exact moment when the cutting tool makes contact with the grinding disc is crucial. Additionally, the delicate control force applied by the index finger and thumb significantly determines the quality of the final product. In contrast, arm support movements play a minor role at other times. Standard ST-GCN gives the same weight to all nodes and frames when processing spatiotemporal graphs, making it difficult to capture these nuances. To this end, the study introduces a spatiotemporal key point attention

mechanism within the ST-GCN unit. This mechanism is a lightweight module that can adaptively learn spatiotemporal weights. The module architecture is shown in Fig. 5.

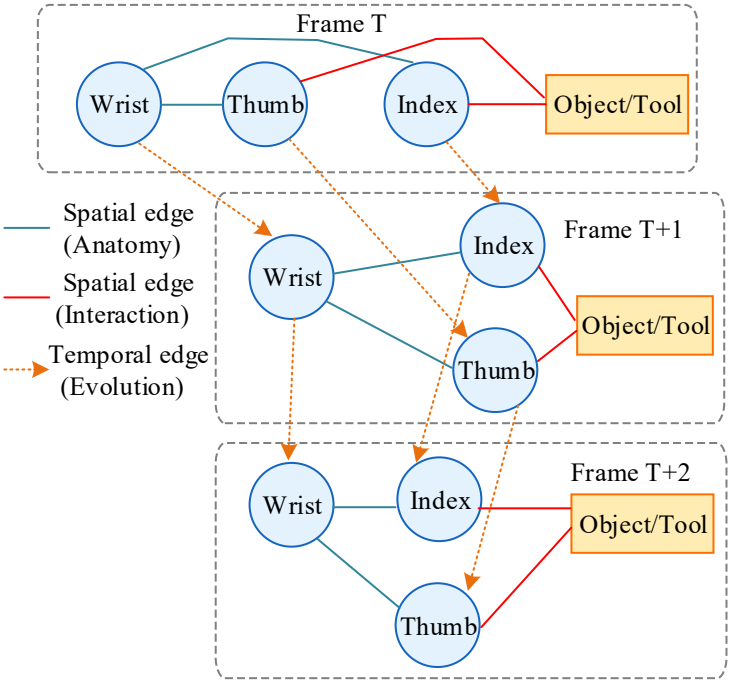


Fig. 4. Schematic diagram of skeleton spatiotemporal diagram construction

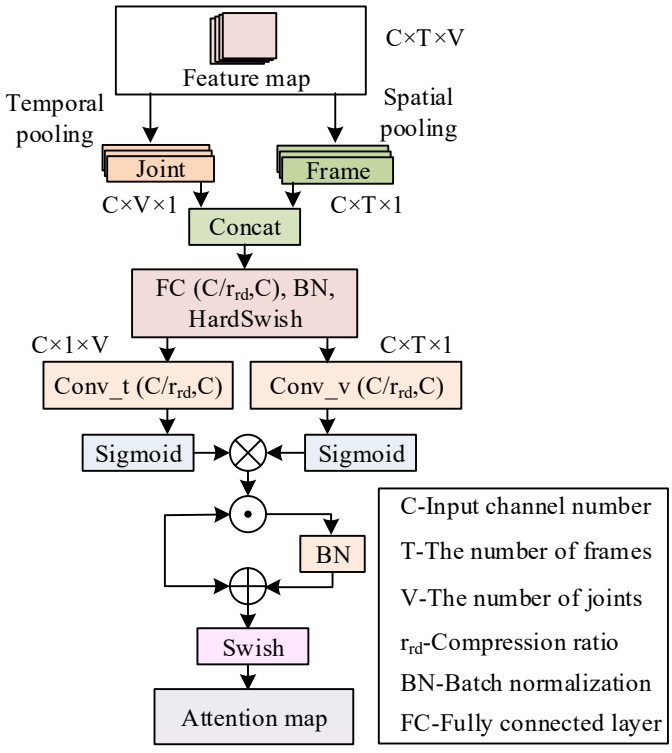


Fig. 5. Architecture of spatiotemporal key point attention mechanism module

In Fig. 5, for the input feature tensor f_{in}, the study first performs average pooling from the time and space dimensions to aggregate global information. Then the temporal attention mask M_t and the spatial attention mask M_v are generated through the fully connected layer and the Sigmoid AF. The final output feature f_{out} is obtained by weighing the attention mask and the original feature, and the calculation process is shown in Eq. (3).

$$f_{out} = f_{in} \square (\sigma(f_{mid} \cdot W_t) \otimes \sigma(f_{mid} \cdot W_v)) \tag{3}$$

In Eq. (3), $\odot$ represents element-wise multiplication and $\otimes$ represents the outer product operation. Through this mechanism, the model can automatically focus on delicate finger movements that reflect high-level skills and ignore non-critical jitters caused by tension, thereby extracting more discriminative skill features. To capture long-term dependencies in gemstone operations, it is necessary to increase the number of network layers. However, as the number of GCN layers increases, gradient vanishing and over-smoothing problems will affect the training effect (Liu et al., 2024; Zhang et al., 2024). To this end, the study introduced residual connections in the graph convolutional unit, which directly transmit the input features to the output end through identity mapping and add them to the convolution result. In addition, single-scale convolutional kernels are difficult to adapt to the problem of different student's different speeds of movement. Inspired by multi-scale decoders, the study designs multi-scale convolutional kernels in TCN. The AMR-ST-GCN topology diagram is shown in Fig. 6.

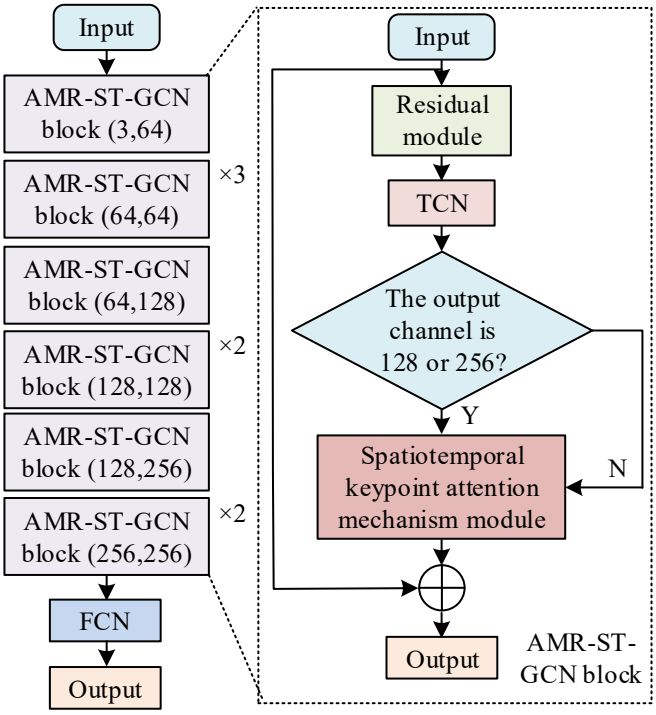


Fig. 6. Topology diagram of AMR-ST-GCN

In Fig. 6, the AMR-ST-GCN consists of multiple AMR-ST-GCN units stacked in series. To capture long-term dependencies, the network has a deep number of layers, with the number of feature channels gradually increasing from 64 to 256 as the layers deepen. In the AMR-ST-GCN unit, the input features first pass through the residual module and the TCN. The residual connection alleviates the gradient vanishing problem in deep networks, while the multi-scale TCN adapts to different action rhythms. A judgment logic is included within the unit, activating the spatiotemporal key point attention mechanism module only in key layers with high feature dimensions, ensuring that the model captures subtle fingertip micro-adjustments in deep features while maintaining computational efficiency in shallow layers. Unlike traditional classification tasks, skill assessment is essentially a metric learning problem. This study proposes the following hypothesis: if a student's operation conforms to the norms, the model trained on expert data should accurately predict their next-frame action; conversely, the greater the prediction deviation, the less standardized the operation. The flowchart of skill quantification assessment based on action prediction error is shown in Fig. 7.

In Fig. 7, during the assessment phase, the student's operation sequence $X_{student}$ is input into an expert prediction model that integrates AMR-ST-GCN and GRU to generate a prediction sequence $\hat{X}_{pred}$. The skill score S is computed grounded in the difference between the prediction sequence and the student's actual subsequent actions, using a variant of the average joint position error, as shown in Eq. (4).

$$S = 100 - \alpha \cdot \frac{1}{T \cdot V} \sum_{t=1}^{T} \sum_{v=1}^{V} || \hat{x}_{t,v} - x_{t,v}^{std} ||_2 \tag{4}$$

In Eq. (4), $x_{t,v}^{std}$ is the ideal expert trajectory derived from the current state, and α is the normalization coefficient. Furthermore, the study introduced joint angle error as an auxiliary indicator, focusing on assessing whether the bending angles of the wrist and fingers are within the optimal ergonomic range. Using the above mathematical model, the study transforms abstract tactile sensations into quantifiable indicators. The system's final output includes the total skill score and specific deduction points, which are fed back to the student in real time through an industrial IoT platform, achieving data-driven, personalized teaching optimization.

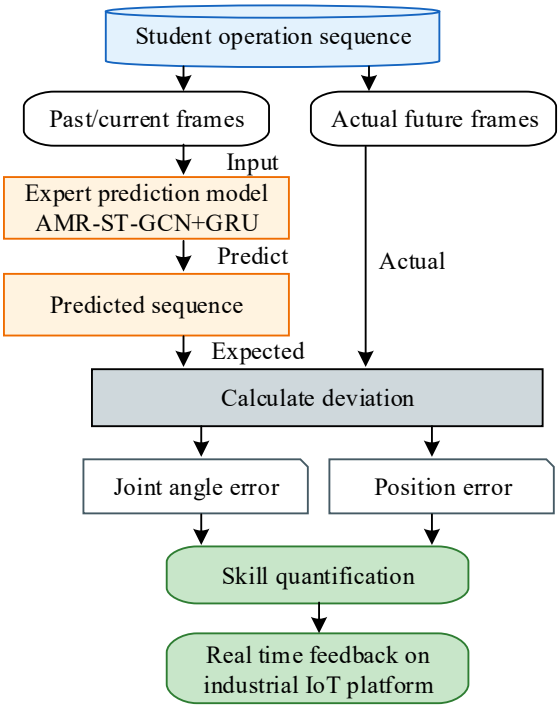


Fig. 7. Flow chart of skill quantitative evaluation

3. Results

3.1. Human Pose Estimation and Evaluation based on an Improved HRNet Model

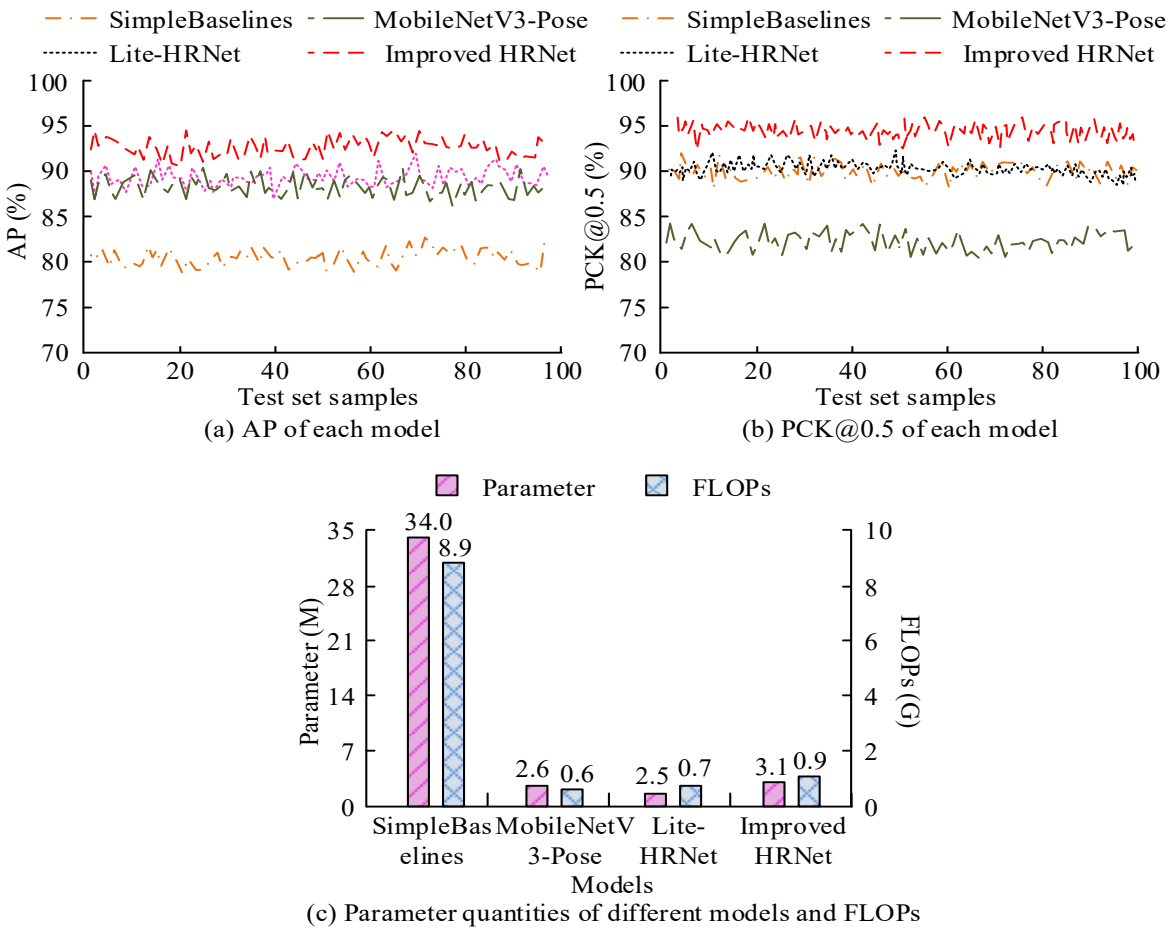
The Gemstone Processing Multimodal Dataset, comprising video streams from three industrial camera perspectives and inertial data from a wireless IMU, has been granted usage rights. The dataset includes practical training data from fifty vocational college students and 5 expert teachers, covering processes such as faceting, polishing, and passing inspection, totaling approximately 1.5 million frames. The dataset was divided into training, validation, and test sets in a 7:2:1 ratio. The experimental environment configuration is presented in Table 1.

To confirm the efficacy of the improved lightweight HRNet for human pose estimation, this study compared the model with Simple Baselines (Ahlmann-Eltze et al., 2025), MobileNetV3-Pose (Cao et al., 2025), and Lite-HRNet (Gao et al., 2024). Evaluation metrics included parameter count, FLOPs, average precision (AP), and percentage of correct key points (PCK@0.5). Fig. 8(a) and 8(b) show that the proposed improved HRNet achieved AP and PCK@0.5 of 92.8% and 94.3%, respectively, on the test set, both of which are superior to those of other state-of-the-art models. In contrast, Simple Baselines and MobileNetV3-Pose performed poorly, with the former achieving only 81.2% AP and the latter only 82.6% PCK@0.5, failing to meet high-accuracy requirements. Fig. 8(c) shows that Lite-HRNet performed best in terms of model complexity, with parameter count and FLOPs as low as 2.5M and 0.7G, respectively. The improved HRNet model had 3.1M parameters and 0.9G FLOPs, which were not significantly different from Lite-HRNet. This indicated that the improved HRNet model achieved the optimal balance between accuracy and efficiency.

Table 1. Experimental environment configuration

Configuration Item	Parameter/Version
Operating System	Ubuntu 20.04 LTS (64-bit)
CPU	Intel Core i9-12900K@ 3.20GHz
GPU	NVIDIA GeForce RTX 3090 (24GB) × 2
RAM	64 GB DDR4 3200MHz
Language	Python 3.8.10
Framework	PyTorch 1.12.1
Acceleration Library	CUDA 11.3, cuDNN 8.2.1

Computer vision lib	OpenCV-Python 4.5.5



(a) AP of each model (b) PCK@0.5 of each model

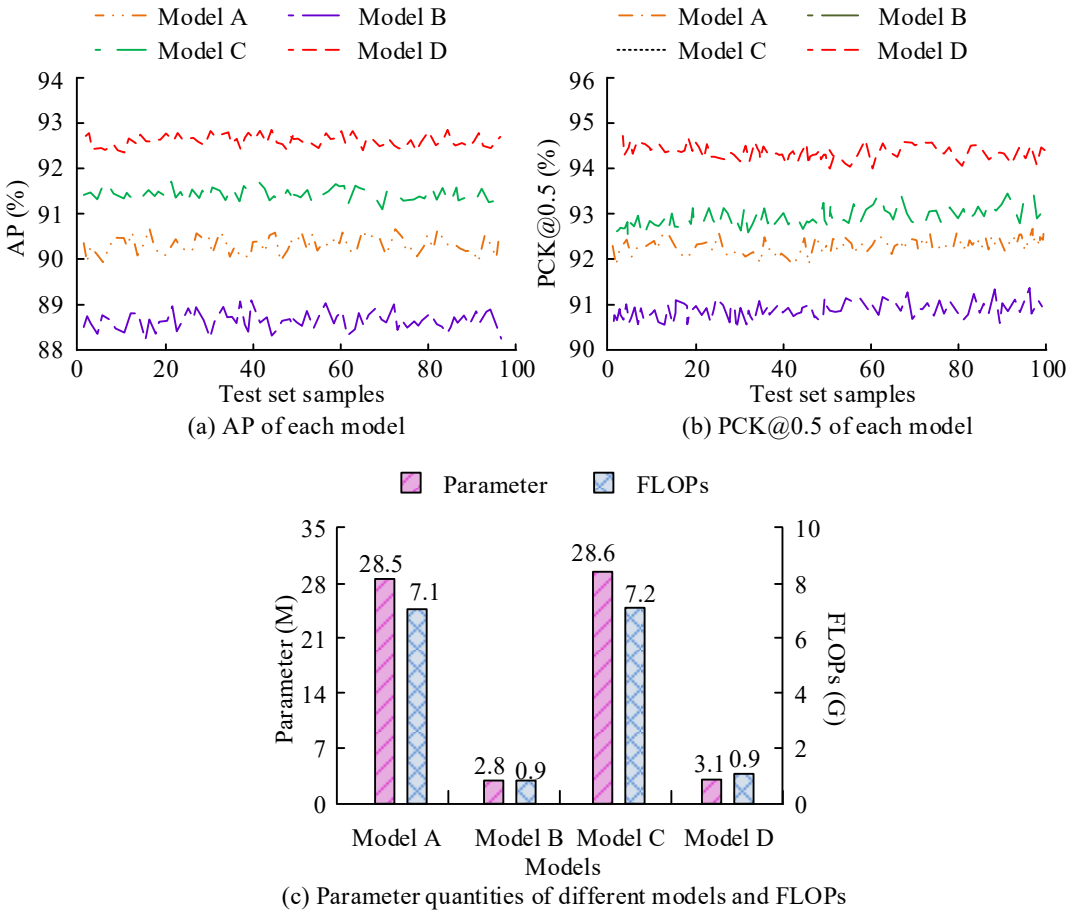
(c) Parameter quantities of different models and FLOPs

Fig. 8. Performance evaluation results of different attitude estimation models

The study then conducted ablation experiments to verify the specific contributions of the introduced CBAM and L-Basicblock module to solving the occlusion problem and reducing model weight. Model A was the original HRNet, Model B was HRNet+L-Basicblock without CBAM, Model C was HRNet+CBAM without L-Basicblock, and Model D was HRNet+L-Basicblock+CBAM, i.e., the proposed model. The performance evaluation results of each ablation model are presented in Fig. 9. As shown in Figs. 9(a) and 9(b), Model B, which only introduced the lightweight module, experienced a slight decrease in detection accuracy compared to the original benchmark Model A due to the significant reduction in network capacity, with an AP of only 88.8%. The AP values of Models C and Model D, which introduced the CBAM attention mechanism, were both improved, with Model D's AP stabilizing at around 92.5% and approaching 94.5% at PCK@0.5. This indicated that CBAM effectively compensated for the feature loss caused by reducing the weight and further improved the recognition accuracy. As shown in Fig. 9(c), both Model A and Model C had 28.5M parameters and over 7G FLOPs, resulting in a heavy computational burden. In contrast, Model D's resource consumption was on par with the extremely lightweight Model B. In summary, Model D achieved an effective balance between accuracy and efficiency by integrating L-Basicblock and CBAM.

3.2. Motion Prediction Performance and Skill Evaluation of AMR-ST-GCN+GRU Model

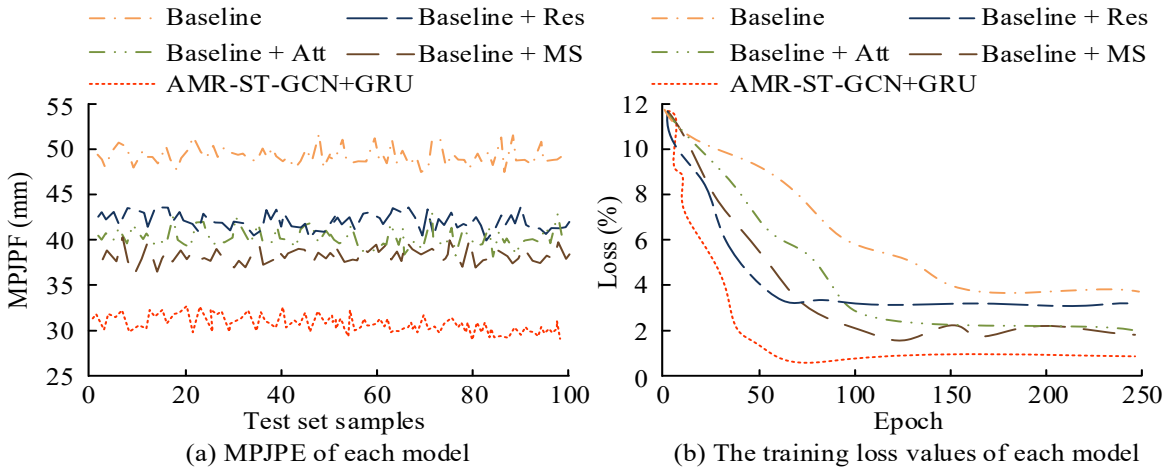
The study first verified the contribution of each improved module in AMR-ST-GCN+GRU to gemstone motion capture, through training and testing on a self-built gemstone processing expert dataset. Five variant models were set up: the original ST-GCN+GRU (Baseline), Baseline+Res with a residual module, Baseline+Att with a spatiotemporal key point attention mechanism, Baseline+MS with a multi-scale TCN, and the proposed AMR-ST-GCN+GRU. The mean joint position error (MPJPE) and loss curves of each model are presented in Fig. 10. As shown in Fig. 10(a), the original Baseline model had the highest MPJPE error, with a mean of 48.5 mm. With the introduction of the residual module, the attention mechanism, and the multi-scale TCN, the errors of the variant models decreased in a stepwise manner, reaching 42.2 mm, 39.8 mm, and 38.6 mm, respectively. The proposed AMR-ST-GCN+GRU model had an MPJPE of only 31.3 mm, indicating that the multi-module fusion strategy could effectively improve the accuracy of skeleton spatial location reconstruction. As shown in Fig. 10(b), the Baseline model reached a final loss of 4.5%, indicating low training efficiency. In contrast, the AMR-ST-GCN+GRU model's loss curve converged to 1.2% after only 70 iterations, demonstrating that each improved module could effectively improve the model's convergence speed and fitting stability.



(a) AP of each model (b) PCK@0.5 of each model

(c) Parameter quantities of different models and FLOPs

Fig. 9. Performance evaluation results of various ablation models



(a) MPJPE of each model (b) The training loss values of each model

Fig. 10. MPJPE and loss values for each model

The study then compared AMR-ST-GCN+GRU with mainstream prediction models, including Simple Multi-Layer Perceptron for Motion Prediction (SiMLPe), Equivariant Motion Prediction (EqMotion), and the MotionMixer model. MPJPE and Mean Angle Error (MAE) were used as evaluation metrics to compare prediction performance in short-term (400ms) and long-term (1000ms) predictions. The MPJPE values of each model in different prediction time domains are shown in Fig. 11. As shown in Fig. 11(a), in the short-term prediction time domain of 400ms, the SiMLPe model had the highest error, with an MPJPE value as high as 38.6mm. The MPJPE values of EqMotion and MotionMixer were 34.1mm and 32.8mm, respectively. The AMR-ST-GCN+GRU model exhibited the lowest MPJPE, with a stable MPJPE value of around 28.1mm, significantly outperforming other models. As shown in Fig. 11(b), the MPJPE of SiMLPe surged to 78.20mm as the prediction time increased to 1000ms. In contrast, the AMR-ST-GCN+GRU model, benefiting from the GRU units ability to remember temporal dependencies and the spatiotemporal key point attention mechanism's ability to

lock onto keyframes, maintained a long-term MPJPE of 52.6mm. This demonstrated the model's excellent robustness in both short-term and long-term tasks.

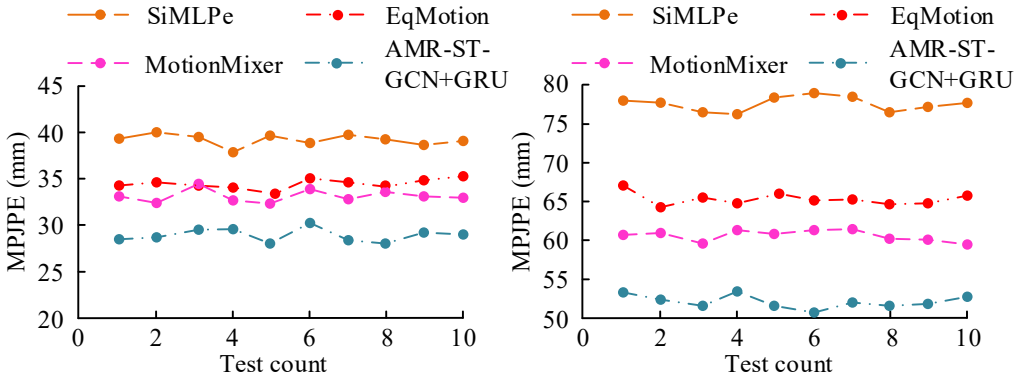


Fig. 11. MPJPE values under different prediction time domains

Fig. 12 shows the MAE values of each model under different prediction time domains. As shown in Fig. 12(a), in the short-term prediction stage of 400ms, the SiMLPe model had the highest MAE value, fluctuating between 12° and 13°. The errors of the EqMotion and MotionMixer models were controlled between 10.5° and 11.5°. The MAE of the AMR-ST-GCN+GRU model was only 9.2°. As shown in Fig. 12(b), when the prediction time was extended to 1000ms, the prediction errors of all models increased. Among them, the MAE of the SiMLPe model increased to around 25°, and the errors of the other two comparative models also increased to over 20°. The MAE of the AMR-ST-GCN+GRU model remained stable at around 16.5°, far lower than the other models. This indicated that the algorithm could effectively support the quantitative evaluation of the standardization of hand movements in gemstone processing under different time domains.

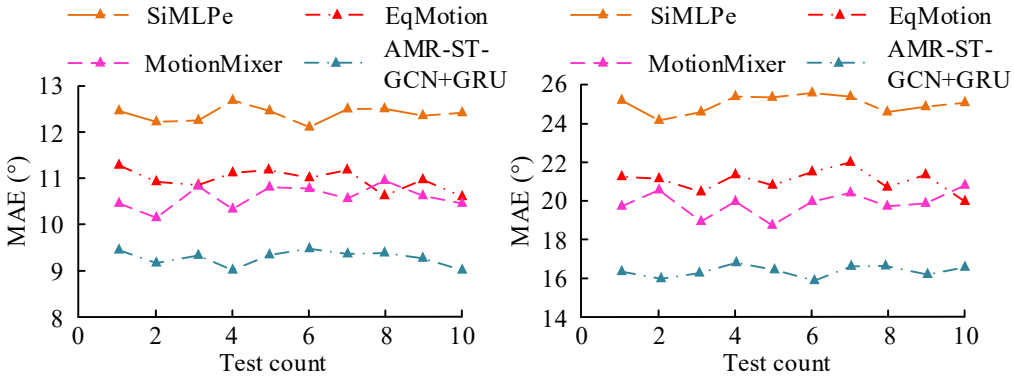
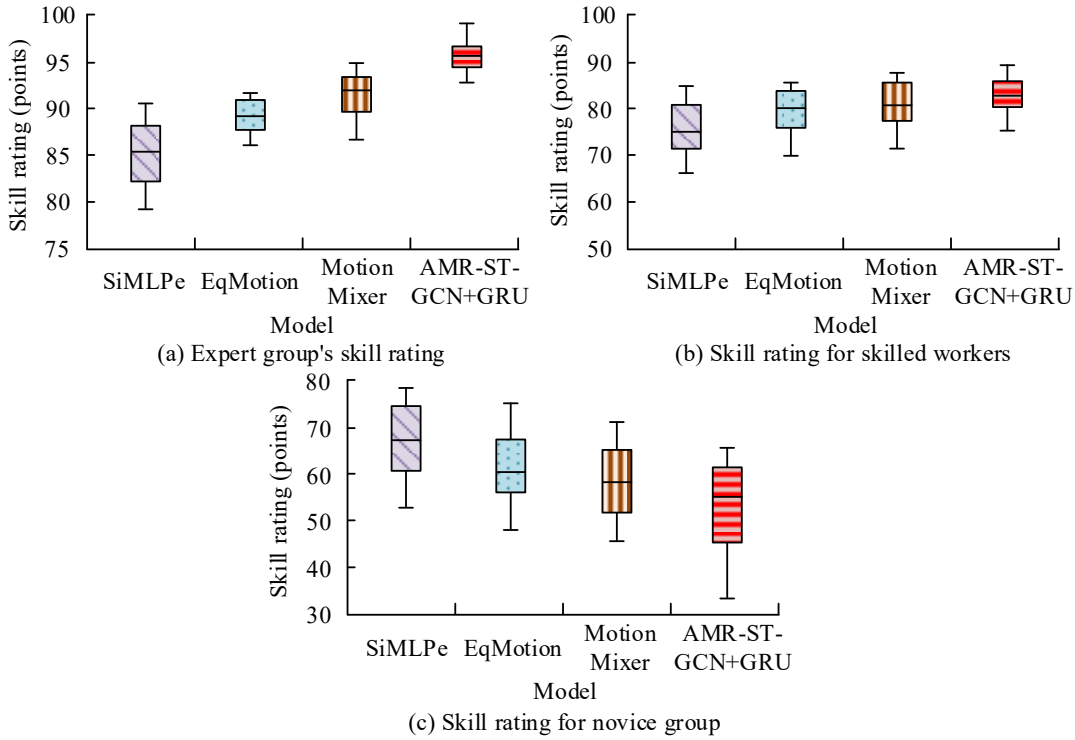


Fig. 12. MAE of each model in different prediction time domains

The study verified whether the scores generated by the AMR-ST-GCN+GRU model could distinguish between experts, skilled workers, and novices. Three groups of participants were invited to perform the same faceted task: 5 experts, 20 skilled workers, and 20 novices. The data from the three groups were input into each model, and box plots of the skill scores for each group were generated, as shown in Fig. 13. Fig. 13(a) shows that the AMR-ST-GCN+GRU model achieved a median score of 95.5, with the flattest box height. The SiMLPe and EqMotion models, however, had lower scores and more dispersed distributions, failing to fully reflect the technical advantages of the experts. Fig. 13(b) shows the scores for the skilled worker group, and Fig. 13(c) shows the scores for the novice group. As skill level decreased, the scores of all models showed a downward trend. The AMR-ST-GCN+GRU model was sensitive to the non-standard movements of novices, with scores dropping significantly to around 54.0 and the box lengthening, reflecting the uneven skill levels among novices. In summary, the AMR-ST-GCN+GRU model has the best hierarchical discrimination, verifying its effectiveness as an automatic assessment tool for vocational college practical training.

To quantify the management value of implementing deployment mechanisms in actual operations, a 4-week comparative verification experiment was conducted in a jade processing training course at a vocational college. The experiment selected 40 novice students with similar initial skill levels and randomly divided them into two groups. The control group (20 people) used a traditional teaching mode that combined teacher inspections, verbal feedback, and direct practical operation. The experimental group (20 people) adopted an IoT system deployed with the AMR-ST-GCN evaluation model. The comparison results for various operational performance indicators between two groups of student's after completing the same number of training hours are shown in Table 2. From the data in Table 2, in terms of teaching efficiency and resource consumption, the blind inspection time of teachers and the number of manual interventions per

student have been significantly reduced. Meanwhile, the economic cost of training consumables decreased from 450 yuan per student to 120 yuan, and the actual scrap rate of raw stones also decreased to 3.50%. In addition, in terms of training quality, the excellent rate of the final assessment of the experimental group students was as high as 85.00%, and the one-time pass rate of the industry certificate was as high as 95.00%, both far exceeding that of the control group. This system released the management pressure on teachers and improved the quality and standardized output ability of vocational jade processing talents.



(a) Expert group's skill rating

(b) Skill rating for skilled workers

(c) Skill rating for novice group

Fig. 13. Skill scoring box diagram for three groups of personnel

4. Conclusion

In vocational college gemstone processing training, challenges often arise, including subjective skill assessment, difficulty handling occlusion during precision operations, and challenges with edge deployment. To address these issues, this study constructed an industrial IoT sensing environment that integrates human-machine-object interconnection and proposed a skeleton spatiotemporal graph modeling and skill assessment method that integrates an improved HRNet and AMR-ST-GCN + GRU. Results showed that, in terms of posture capture, the improved HRNet model, integrating the CBAM attention mechanism and lightweight L- Basicblock, achieved an average accuracy of 92. 8% and a PCK@ 0. 5 of 94. 3% on the test set, outperforming other mainstream models. This was because the CBAM module effectively overcame frequent mutual occlusion interference in precision processing through the concatenation of channel and spatial attention. Simultaneously, L- Basicblock significantly reduced computational redundancy. In terms of skill assessment, the AMR-ST- GCN + GRU model achieved an average MAE of only 9. 9.2° in the 400ms short-time prediction domain, lower than comparative models such as SiMLPe and EqMotion. The model's spatiotemporal key point attention mechanism, in conjunction with a multi-scale TCN, adaptively focused on key action frames, effectively capturing differences in operational rhythm among students of varying skill levels. Furthermore, the model showed a clear distribution of ratings for experts, skilled workers, and novices. In summary, the proposed method achieved high-precision spatiotemporal modeling and quantitative evaluation of gemstone handling behavior.

The application value of the research lies in the deep integration of vocational training systems with industrial production capabilities. In terms of labor quality control in precision manufacturing, this evaluation system converts small motion deviations during gemstone processing into traceable quantitative data, making it possible to conduct pre-screening of labor quality. Meanwhile, the objective indicators output by the system have broken down the evaluation barriers caused by subjective experience differences, effectively promoting the standardization of skills among different training institutions. By aligning threshold parameters, such as action fault tolerance, with the actual industry competency framework, vocational colleges can reverse optimize course standards based on the actual job requirements of the enterprise. In addition, in the context of digital transformation in the manufacturing industry, allowing students to learn in a highly interconnected industrial IoT perception environment can effectively enhance their operational skills and strengthen their digital literacy, providing a high-quality talent reserve for labor preparation in the industry 4.0 era. Additionally, the skeleton spatiotemporal graph modeling and prediction error evaluation framework based on multimodal perception has high cross task applicability. In practical applications, by redefining the graph nodes of specific tools and introducing expert data from the corresponding field for model fine-tuning, the system can be migrated to vocational precision manufacturing training scenarios.

Table 2. Comparison results of various operational performance indicators

Evaluation Dimension	Management Metric	Control Group	Experimental Group
Efficiency and Intervention	Teacher's blind inspection time	120 min	25 min
	Manual interventions per student	18.5 times	4.2 times
Time Cost	Average time to proficiency	45.5 hours	28.0 hours
Material Cost	Scrap rate of real gemstones	22.40%	3.50%
	Average material cost per student	450 yuan	120 yuan
Training Quality	Excellent rate in final exam	45%	85.00%
	First-time certification pass rate	70%	95.00%

However, the data collected in this study were collected in a controlled training environment, and robustness verification under extreme lighting variations and complex dust interference in a workshop remains insufficient. Meanwhile, the system's motion capture accuracy still has certain limitations. On the one hand, when the hand and tool self-obstruct, it will affect the capture accuracy. On the other hand, wearable IMUs are difficult to completely eliminate integral drift errors during long-term continuous operation. Future research will expand the training dataset to include multiple scenarios to verify the model's generalization performance.

Funding

Anhui Provincial Department of Education Humanities and Social Sciences Key Research Project: Mechanism Innovation and Systematic Pathways for the Cultivation of Craftsman Talents in Higher Vocational Education from the Perspective of Industry–Education Integration within Provincial High-Level Gemstone and Jewelry Professional Clusters (2025AHGXSK30549). Education and Teaching Research Planning Project of the Anhui Vocational and Adult Education Association: An Empirical Study on the Relationships among Learning Motivation, Academic Self-Efficacy, and Academic Achievement of Students in Higher Vocational Colleges in Anhui Province (Azcj2022086). Young and Middle-Aged Teacher Development Program of the Anhui Provincial Department of Education: Domestic Visiting Scholar and Professional Development Funding Program for Young Backbone Teachers (JNFX2023154).

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence AI Tools

The author used Grammarly Premium solely for language editing and readability improvement. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

Reference

Ahlmann-Eltze, C., Huber, W., and Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8), 1657-1661. https://doi.org/10.1038/s41592-025-02772-6

An X., Zhao, L., Gong, C., Wang, N., Wang, D., and Yang, J. (2024). Sharpose: Sparse high-resolution representation for human pose estimation. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2), 691-699. https://doi.org/10.1609/aaai.v38i2.27826

Cao, Z., Li, J., Fang, L., Li, Z., Yang, H., and Dong, G. (2025). Research on efficient classification algorithm for coal and gangue based on improved MobilenetV3-small. *International Journal of Coal Preparation and Utilization*, 45(2), 437-462. https://doi.org/10.1080/19392699.2024.2353128

Chen, C., Wang, C., Liu, B., He, C., Cong, L., and Wan, S. (2023). Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 24(11), 13023-13034. https://doi.org/10.1109/TITS.2022.3232153

Gao, Z., Wang, S., Yang, Z., Peng, G., Li, Y., Fang, X., and Li, S. (2024). Attention mechanism and lightweight network fusion HRNet: a lightweight remote sensing road extraction algorithm integrating attention mechanisms. *Journal of Electronic Imaging*, 33(6), 063015-063015. https://doi.org/10.1117/1.JEI.33.6.063015

Huang, X., Liu, Q., Chen, H., Li, Y., Huang, C., Zhu, X., and Yin, W., Y. (2024). A new hybrid equivalent modeling method of low-frequency radiation source based on GS and JADE algorithms and phaseless near-field data. *IEEE Transactions on Electromagnetic Compatibility*, 66(3), 917-927. https://doi.org/10.1109/TEMC.2024.3359259

Hu, Y., Jia, Q., Yao, Y., Lee, Y., Lee, M., Wang, C., and Yu, F., R. (2024). Industrial internet of things intelligence empowering smart manufacturing: A literature review. *IEEE Internet of Things Journal*, 11(11), 19143-19167. https://doi.org/10.1109/JIOT.2024.3367692

Jiang, X., Xu, J., and Xu, X. (2024). An overview of domestic and international applications of digital technology in teaching in vocational education: Systematic literature mapping. *Education and Information Technologie*s, 29(13), 16867-16899. https://doi.org/10.1007/s10639-024-12528-y

Kathirvel, N., Bharat, S., Kathirvel, A., and Maheswaran, C., P. (2024). Artificial General–Internet of Things (AG-IoT) for robotics of automation. *Systemic Analytics*, 2(1), 59-76. https://doi.org/10.31181/sa21202417

Kumar, M., Kim, C., Son, Y., Singh, S., K., and Kim, S. (2024). Empowering cyberattack identification in IoHT networks with neighborhood-component-based improvised long short-term memory. *IEEE Internet of Things Journal*, 11(9), 16638-16646. https://doi.org/10.1109/JIOT.2024.3354988

Lei, Z., Qu, Q., Chen, H., Zhang, Z., Dou, G., and Wang, Y. (2023). Mainlobe jamming suppression with space–time multichannel via blind source separation. *IEEE Sensors Journal*, 23(15), 17042-17053. https://doi.org/10.1109/JSEN.2023.3278709

Li, C., Mao, Y., Huang, Q., Zhu, X., and Wu, J. (2023). Scale-aware graph convolutional network with part-level refinement for skeleton-based human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6), 4311-4324. https://doi.org/10.1109/TCSVT.2023.3334872

Li, J., Bi, Y., Wang, S., and Li, Q. (2023). CFRLA-Net: A context-aware feature representation learning anchor-free network for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9), 4948–4961. https://doi.org/10.1109/TCSVT.2023.3245613

Ling, Y., Ma, Z., Zhang, Q., Xie, B., and Weng, X. (2024). PedAST-GCN: Fast pedestrian crossing intention prediction using spatial–temporal attention graph convolution networks. *IEEE Transactions on Intelligent Transportation Systems*, 25(10), 13277-13290. https://doi.org/10.1109/TITS.2024.3398252

Liu, S., L., Ding, Y., N., Zhang, J., R., Liu, K., Y., Zhang, S., F., Wang, F., L., and Huang, G. (2024). Multidimensional refinement graph convolutional network with robust decouple loss for fine-grained skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4), 7615-7626. https://doi.org/10.1109/TNNLS.2024.3384770

Ren, Q., Lu, Z., Wu, H., Zhang, J., and Dong, Z. (2023). HR-Net: A landmark-based high realistic face reenactment network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11), 6347–6359. https://doi.org/10.1109/TCSVT.2023.3268062

Tsoy, A., Liu, Z., Zhang, H., Zhou, M., Yang, W., Geng, H., and Geng, Z. (2024). Image-free single-pixel keypoint detection for privacy-preserving human pose estimation. *Optics Letters*, 49(3), 546-549. https://doi.org/10.1364/OL.514213

Wu, W., Tu, F., Niu, M., Yue, Z., Liu, L., Wei, S., and Yin, S. (2023). STAR: An STGCN architecture for skeleton-based human action recognition. *IEEE Transactions on Circuits and Systems I: Regular Paper*s, 70(6), 2370-2383. https://doi.org/10.1109/TCSI.2023.3254610

Zhang, H., Dun, Y., Pei, Y., Lai, S., Liu, C., Zhang, K., and Qian, X. (2024). HF-HRNet: A simple hardware-friendly high-resolution network. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8), 7699-7711. https://doi.org/10.1109/TCSVT.2024.3377365

Zhang, M., Quan, Z., Wang, W., Chen, Z., Guo, X., and Li, Y. (2024). ASMGCN: Attention-based semantic-guided multistream graph convolution network for skeleton action recognition. *IEEE Sensors Journal*, 24(12), 20064-20075. https://doi.org/10.1109/JSEN.2024.3388154

Zhao, W., Yan, J., Jin, D., and Ling, J. (2024). C-HRNet: High-resolution network based on contexts for single-frame phase unwrapping. *IEEE Photonics Journal*, 16(2), 1-10. https://doi.org/10.1109/JPHOT.2024.3381544

Zhen, P., Yan, X., Wang, W., and Chen, H., B. (2023). A highly compressed accelerator with temporal optical flow feature fusion and tensorized LSTM for video action recognition on terminal device. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(10), 3129-3142. https://doi.org/10.1109/TCAD.2023.3241113

Zhuang, T., Qin, Z., Ding, Y., Deng, F., Chen, L., Qin, Z., and Choo, K., K., R. (2023). Temporal refinement graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Artificial Intelligence*, 5(4), 1586-1598. https://doi.org/10.1109/TAI.2023.3329799

Zhou Zheng obtained her PhD in Education (2024) from Infrastructure University Kuala Lumpur, Malaysia and her master's degree in teaching Chinese to Speakers of Other Languages (2015) from Nanjing University. She was a visiting scholar at the University of Science and Technology of China and Troy University, USA. She is currently working as a lecturer at Anhui Technical College of Industry and Economy. Dr. Zhou Zheng has been invited to deliver academic lectures and teaching-related talks in the fields of language education and pedagogy. She has published multiple peer-reviewed articles and an academic monograph. Her strengths lie in linguistics, education, and teaching Chinese to Speakers of Other Languages, with a focus on learning motivation, curriculum development, and instructional innovation.