

Interactive Teaching of Robotics Based on Multimodal Human Pose Estimation

Xianhua Li¹, Minhui Wang¹, Zisen Hua², Junjie Mi¹, Long Li², Zhipeng Yu²

¹ School of Mechatronics Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China

² School of Artificial Intelligence, Anhui University of Science and Technology, Huainan 232001, Anhui, China

DOI: <https://doi.org/10.47772/IJRISS.2026.100900089>

Received: 20 September 2026; Accepted: 25 September 2026; Published: 30 September 2026

ABSTRACT

Traditional robotics courses often provide limited opportunities for engineering practice, making it difficult for students to apply theoretical knowledge to integrated engineering tasks. To address this issue, this study proposes a practical robotics teaching approach integrating multimodal vision and human pose estimation. An RGB-IR multimodal human pose dataset is constructed using an Intel RealSense D435 camera, and an RGB-IR dual-branch pose estimation network with a Cross-Modal Adaptive Gated Fusion (CM-AGF) mechanism is developed based on YOLOv8-Pose. The detected two-dimensional keypoints are combined with depth information for three-dimensional reconstruction and joint-angle calculation, and the resulting human motion parameters are mapped to robot control commands. Experimental results demonstrate the feasibility of integrating multimodal perception, pose estimation, three-dimensional motion reconstruction, and robot control into a unified practical teaching platform. The proposed approach provides students with opportunities to integrate computer vision, artificial intelligence, and robotics in engineering practice.

Keywords: robotics; multimodal vision; human pose estimation; teaching practice; human-robot interaction.

INTRODUCTION

Robotics courses traditionally focus on mechanism modeling, kinematic analysis, dynamic calculation, and trajectory planning [1]. In recent years, human pose estimation, depth vision, virtual simulation, and other intelligent perception technologies have gradually been introduced into robotics education. The deep integration of robotics with artificial intelligence, computer vision, and intelligent perception has introduced new requirements for the knowledge structure, practical content, and engineering competence development of robotics courses [2]. Yang et al. reconstructed experimental teaching content based on real robot interaction tasks and adopted problem-oriented and project-driven teaching methods to promote the transformation of theoretical knowledge into capabilities in algorithm design, system integration, and engineering debugging [3]. Xu and Wu detected human keypoints, calculated joint angles, and controlled robotic-arm motion, improving the intuitiveness and interactivity of robot-related applications [4]. Upadhyay et al. used RGB-D sensing to acquire three-dimensional human motion parameters, providing a non-contact source of motion information for robot following control [5]. These studies provide a foundation for integrating intelligent perception technologies into robotics education. Although existing studies can help establish a relatively systematic theoretical and practical knowledge framework, many existing approaches rely on a single RGB image for pose estimation or use depth information mainly for three-dimensional reconstruction following two-dimensional keypoint detection [6,7]. Comparatively limited attention has been given to the complementary characteristics of RGB and infrared modalities and to the design of multimodal feature fusion networks in robotics teaching applications [8–10].

In this study, an Intel RealSense D435 camera is used as the data acquisition device to construct an RGB-IR bimodal human pose dataset. Based on YOLOv8-Pose, an RGB-IR dual-branch feature extraction network incorporating a Cross-Modal Adaptive Gated Fusion (CM-AGF) mechanism is designed. Two-dimensional

human keypoints are first detected, after which depth information is used for three-dimensional keypoint reconstruction and joint-angle calculation. Human motion parameters are then mapped to robot control commands, forming an interactive teaching process of “multimodal acquisition–feature fusion–pose estimation–motion calculation–robot execution.” The proposed approach integrates multimodal vision, deep learning, robot control, and system development into a unified robotics practice and provides a reference for developing students’ comprehensive practical capabilities.

Multimodal Vision-Based Human Pose Estimation

A. Multimodal Data Acquisition

RGB-IR multimodal data acquisition is the basis for human pose estimation and robot interaction control. High-quality and synchronized multimodal images provide reliable data for the training and testing of pose estimation networks and help students understand the imaging characteristics and complementary information provided by different visual modalities. RGB images contain rich information on human texture, color, and surrounding scenes, but their quality can deteriorate under low-light, backlight, and complex-background conditions. Infrared images are less affected by visible-light variations and can retain relatively clear human contours under low-illumination conditions. Therefore, synchronized RGB and infrared images provide complementary visual information for subsequent cross-modal feature extraction and fusion. In this study, an Intel RealSense D435 camera is used to construct the multimodal visual data acquisition system. The camera simultaneously outputs RGB, infrared grayscale, and depth images. RGB and infrared images serve as the dual-modal inputs of the human pose estimation network, while depth images are used for three-dimensional back-projection of two-dimensional human keypoints and joint-motion parameter calculation. To construct the RGB-IR human pose dataset, synchronized RGB, infrared, and depth images were collected using the Intel RealSense D435 under different illumination conditions and human postures. The resulting dataset contains 12,440 paired RGB-IR samples, including 7,808 training samples, 2,281 validation samples, and 2,351 test samples. Human keypoints were annotated according to the COCO 17-keypoint format. The dataset includes samples acquired under normal illumination, low-light, and backlight conditions. Table I summarizes the dataset composition.

Table I Composition of the rgb-ir human pose dataset

Item	Description
Total paired RGB-IR samples	12440
Training	7808
Validation	2281
Test	2351
Keypoint format	COCO 17-keypoint
Image resolution	640 × 640
Modalities	RGB, IR, and depth
Acquisition device	Intel RealSense D435

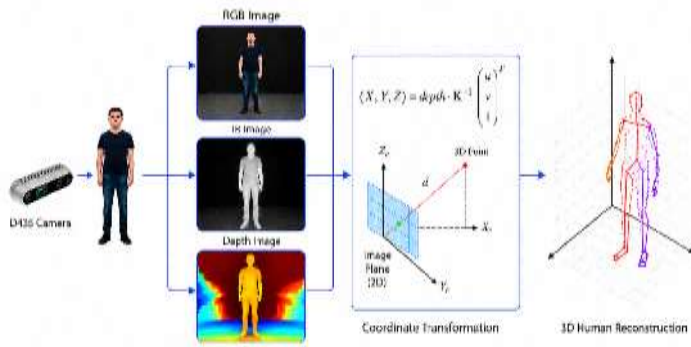


Figure 1. Multimodal data acquisition system

As illustrated in Figure 1, the Intel RealSense D435 simultaneously captures RGB, infrared, and depth images. The RGB and infrared images provide human appearance and contour information, whereas the depth image provides distance information corresponding to individual keypoints. First, two-dimensional human keypoints (u, v) are detected in the image plane. The corresponding depth value z , together with the intrinsic parameters of the camera, is then used to back-project the two-dimensional pixel coordinates into the camera coordinate system, obtaining three-dimensional keypoint coordinates (x, y, z) . Finally, the joints are connected to construct a three-dimensional human skeleton and reconstruct the human pose in space.

B. Image Preprocessing

The RGB and infrared sensors of the D435 camera have different imaging fields of view and optical coordinate systems. Consequently, the position, scale, and edge contours of the same human target may differ between RGB and infrared images. Directly feeding unprocessed bimodal images into the network may cause spatial inconsistency between corresponding human regions in the feature space, thereby affecting cross-modal feature fusion and keypoint localization. Therefore, spatial alignment of RGB and infrared images is performed before human pose estimation.

In this study, the RGB image coordinate system is used as the reference. The intrinsic and extrinsic parameters obtained through camera calibration are used to transform pixels in the infrared image into the RGB camera coordinate system. Coordinate mapping and image resampling are subsequently performed to obtain an infrared image with the same spatial position and image dimensions as the RGB image. During alignment, three-dimensional coordinate transformation is performed according to the rotation and translation relationships between the sensors. The transformed spatial points are then projected onto the RGB imaging plane to establish pixel-level correspondence between RGB and infrared images. Interpolation is used to fill holes and non-integer pixel positions generated during the mapping process to preserve image structural integrity. The comparison between the RGB image and the infrared images before and after alignment is shown in Figure 2.



Figure 2. Comparison of RGB and infrared images

As shown in Figure 2, before alignment, noticeable positional deviations exist around the head, shoulders, and upper limbs between the RGB and infrared images. After spatial alignment, the human regions and major joint locations in the two modalities largely overlap, and the human contours exhibit good consistency. The alignment

procedure reduces spatial errors caused by differences between the imaging sensors and provides accurately paired data for subsequent RGB and infrared dual-branch feature extraction, cross-modal information fusion, and human keypoint detection.

C. Multimodal Human Pose Estimation

Human pose estimation is a key component in acquiring human motion information and implementing robot interaction control. The conventional YOLOv8-Pose model mainly uses RGB images as input and exploits human texture, color, edge, and other visual information for human detection and keypoint localization. However, its recognition performance can be affected by low illumination, backlighting, and complex backgrounds. In comparison, infrared images are less affected by visible-light variations and can preserve relatively clear human contours under low-illumination conditions. Therefore, this study adopts YOLOv8-Pose as the baseline model and introduces RGB and infrared images to construct a dual-branch feature extraction network with a Cross-Modal Adaptive Gated Fusion (CM-AGF) mechanism, thereby improving the stability of human keypoint detection under complex illumination conditions.

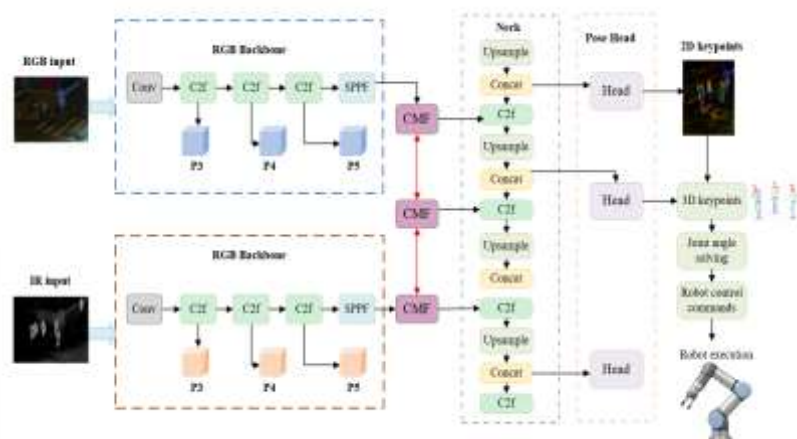


Figure 3. Architecture of the bimodal YOLOv8-Pose network with CM-AGF

As shown in Figure 3, the improved model mainly consists of an RGB feature extraction branch, an infrared feature extraction branch, multi-scale CM-AGF modules, a multi-scale feature aggregation network, and a human pose prediction head. Spatially aligned RGB and infrared images are separately fed into two feature extraction branches with identical network structures but independent parameters. The RGB branch mainly extracts human texture, color, and local-detail features, while the infrared branch focuses on human contours, edges, and grayscale-distribution features. By maintaining independent feature extraction processes for the two modalities, information interference caused by directly concatenating heterogeneous modal data at the network input stage can be reduced, while complementary visual information is retained for subsequent feature-level fusion.

Specifically, the RGB and infrared backbones extract multi-scale features at the P3, P4, and P5 levels. At each scale, the corresponding RGB and infrared features are fed into an independent CM-AGF module. Rather than directly concatenating the two modality features or assigning them fixed fusion weights, CM-AGF dynamically estimates their relative contributions at the channel level. First, global average pooling is applied to the RGB and infrared features to obtain compact channel descriptors. The two descriptors are then concatenated and processed by a lightweight gating network consisting of two 1×1 convolutional layers with a SiLU activation function. The gating network generates channel-wise response scores for the RGB and infrared modalities. A competitive Softmax operation is subsequently performed between the two modality scores for each channel, producing normalized RGB and infrared weights whose sum equals one. Finally, the learned weights are used to reweight the corresponding RGB and infrared features, and the two reweighted features are integrated through weighted element-wise addition. In this manner, CM-AGF enables the network to adaptively emphasize the more informative modality for different feature channels and scales while preserving complementary cross-modal information.

Robot Interaction Teaching Platform and Experiments

A. Hardware Composition and Platform Architecture

To establish an effective connection between multimodal human pose estimation and robot motion control, an interactive teaching platform is constructed using an Intel RealSense D435 camera, a deep-learning-based human pose estimation model, and a robot control module. The platform mainly consists of a multimodal visual acquisition module, a human pose estimation module, a three-dimensional motion parameter calculation module, a communication and control module, and a robot execution module. The overall architecture is shown in Figure 4.

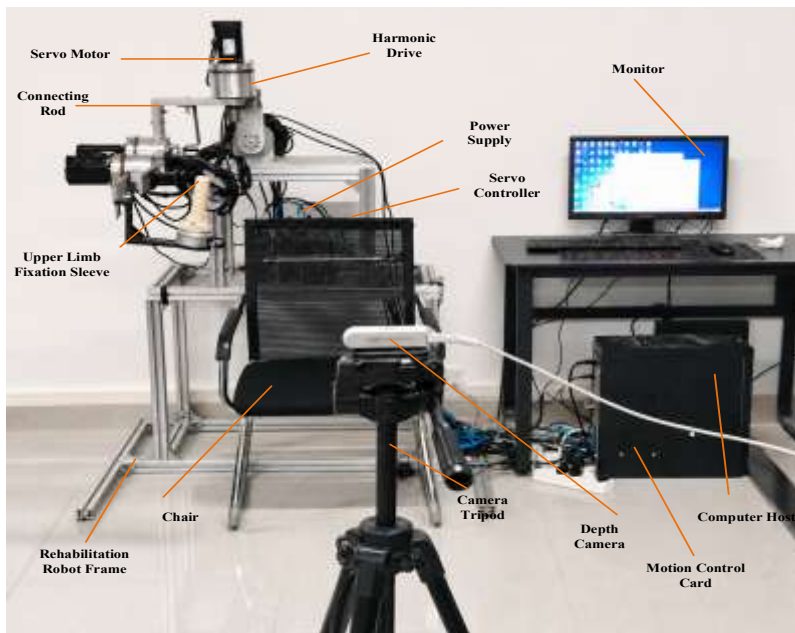


Figure 4. Overall architecture of the multimodal human pose estimation and robot interaction teaching platform

As shown in Figure 4, operation of the platform is divided into two stages: offline training and online interaction. During the offline training stage, multimodal dataset construction, keypoint annotation, network training, and model parameter optimization are performed. During the online interaction stage, aligned RGB-IR images acquired in real time using the D435 camera are fed into the trained model to obtain two-dimensional human keypoints. The depth information is subsequently used to calculate three-dimensional joint coordinates and motion angles. Finally, control commands are transmitted to the robot system through the communication module. This process establishes a closed-loop teaching workflow of “multimodal perception-pose estimation-motion calculation-control decision-robot execution.” Through the construction of the overall platform architecture, visual sensing, deep learning, spatial coordinate transformation, and robot control are connected within a unified system, enabling students to understand the relationships among different knowledge modules through an integrated robot interaction experiment.

B. RGB-IR Human Pose Estimation Experiment

To verify the feasibility of the RGB-IR bimodal human pose estimation model, the constructed dataset is used for model training, and pose estimation experiments are conducted under different illumination conditions and human motions. Model training and validation are performed on a cloud server. The hardware configuration consists of an NVIDIA GeForce RTX 5090 GPU with 32 GB of video memory, a 24-core CPU, and 32 GB of system memory. The software environment consists of Windows 10, Python 3.10, and PyTorch 2.3.0. The input image size is 640×640 , the batch size is 8, the initial learning rate is 1×10^{-3} , and the model is trained for 100 epochs. After training, the model weights are loaded into the human pose estimation program to process paired RGB and infrared images acquired using the D435 camera. The human pose estimation results are shown in Figure 5.

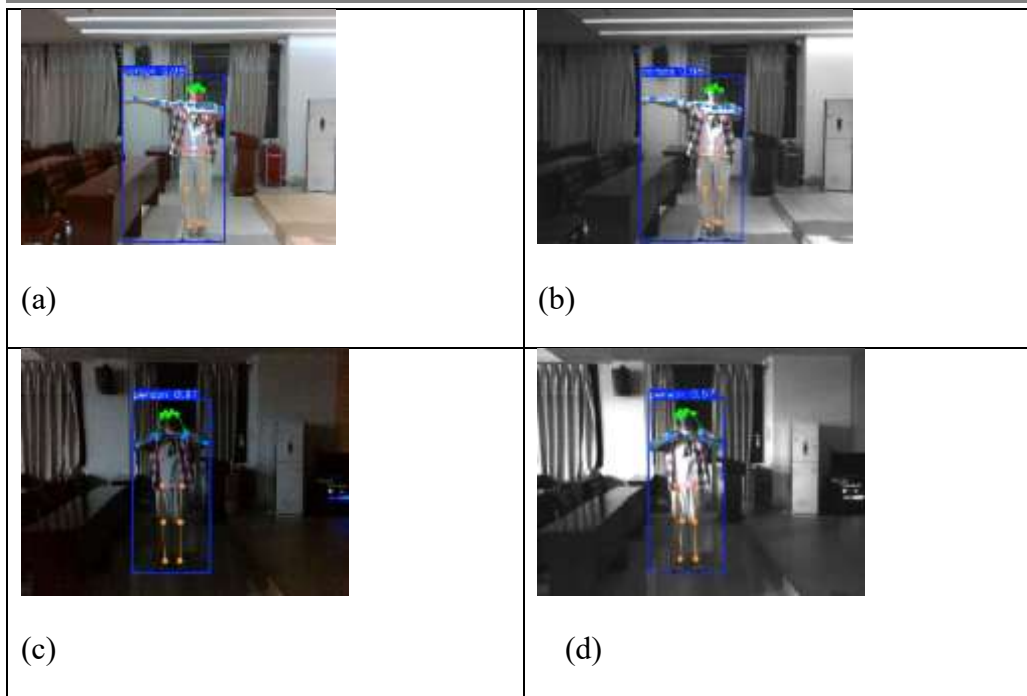


Figure 5. Human pose estimation results

As illustrated in Figure 5, students can directly observe the differences between RGB and infrared images under different environmental conditions and examine the processes of dual-branch feature extraction, cross-modal fusion, and multi-scale pose prediction. Through model training, parameter configuration, inference-result visualization, and keypoint error analysis, the experiment provides practical opportunities to understand the complete workflow from multimodal data input to human pose output and to practice model debugging, result analysis, and engineering problem solving.

Table II Comparison of rgb-only and rgb-ir pose estimation performance

Parameter	RGB-only	RGB-IR
Precision	99.59%	99.90%
Recall	99.69%	99.96%
mAP@0.5 (%)	99.40%	99.44%
mAP@0.5:0.95(%)	77.87%	92.35%
Parameters(M)	3.290	4.418
Inference(ms)	0.465	0.839

As shown in Table II, the RGB-IR model achieved comparable mAP@0.5 to the RGB-only baseline, while improving mAP@0.5:0.95 from 77.87% to 92.35%, an increase of 14.48 percentage points. Precision and recall also increased from 99.59% and 99.69% to 99.90% and 99.96%, respectively. Although the parameter count increased from 3.29 M to 4.42 M, the marked improvement in mAP@0.5:0.95 indicates that multimodal feature fusion enhances keypoint localization accuracy.

C. Human Pose-Driven Robot Interaction Experiment

To verify the feasibility of applying human pose estimation results to robot interaction control, a human pose-driven robot motion experiment is conducted. The Intel RealSense D435 camera simultaneously captures RGB, infrared, and depth images of the human body. After spatial alignment, RGB and infrared images are fed into

the trained bimodal YOLOv8-Pose model to obtain two-dimensional keypoints including the shoulder, elbow, and wrist. Depth information is then used to reconstruct three-dimensional keypoints and calculate joint angles. The resulting angle data undergo smoothing, range limiting, and mapping. Robot joint control commands are subsequently generated, establishing an interactive process of “human motion acquisition-pose estimation-three-dimensional reconstruction-robot execution.” This process is consistent with the experimental concept of closed-loop robot control driven by visual pose and depth information. The control principle is shown in Figure 6.

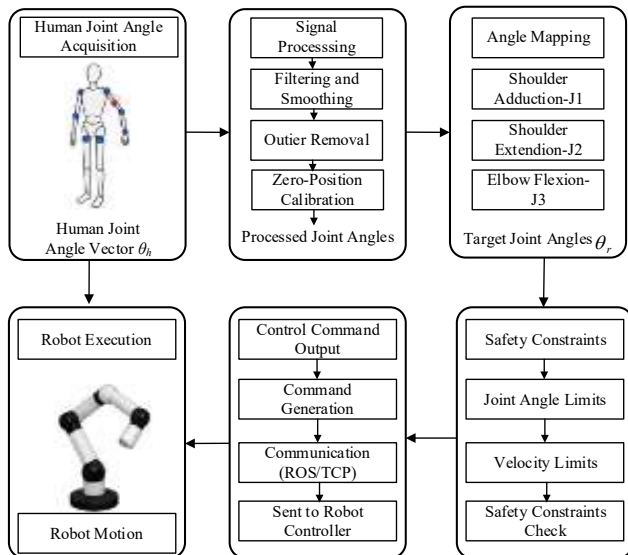


Figure 6. Mapping principle from human pose to robot joint control

As shown in Figure 6, the robot control mapping process mainly includes human joint-angle calculation, temporal filtering, zero-position calibration, angle mapping, safety limiting, and control-command transmission. By establishing a mapping relationship between human joint angles and robot target joint angles, human motion information can be converted into executable robot joint control commands.

Typical movements, including shoulder adduction, shoulder extension, and elbow flexion, are selected for the experiments. When the operator performs a specified movement, the robot executes the corresponding motion according to the joint angles calculated in real time. The feasibility of the proposed interaction method is evaluated by observing the consistency between human movements and the corresponding robot motions. The experimental results are shown in Figure 7.

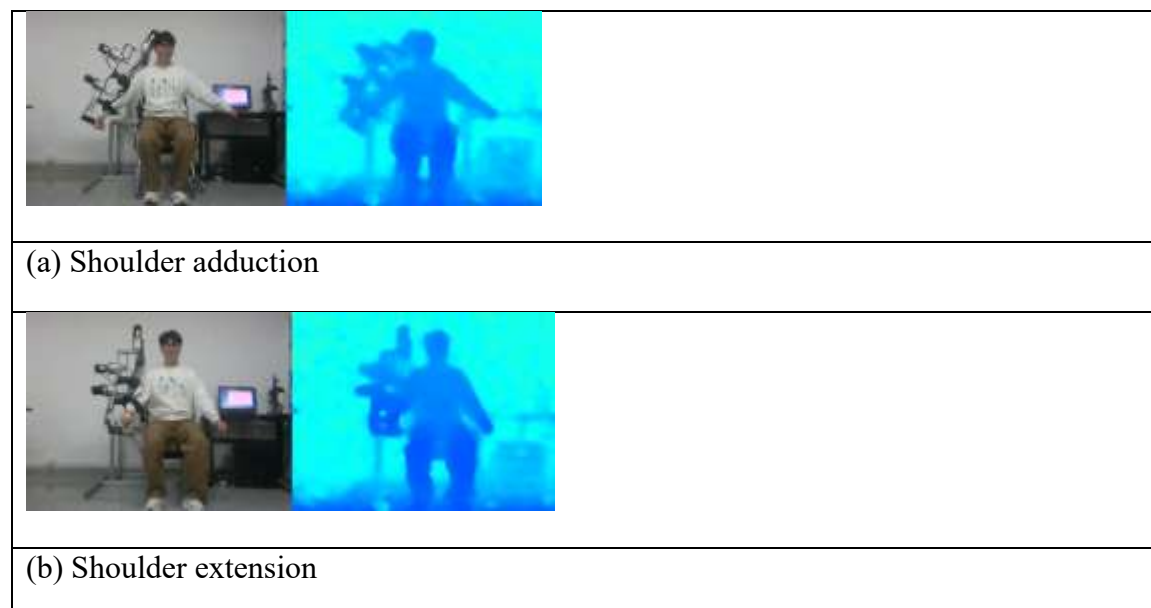




Figure 7. Human pose-driven robot interaction experimental results

The experimental results show that, under the specified movements and experimental conditions, the robot can perform corresponding following motions according to the human pose estimation results, verifying the feasibility of applying multimodal visual pose estimation to robot interaction teaching. Through the human pose-driven robot interaction experiment, students can directly observe how multimodal visual perception results are transformed into robot motion commands and understand the relationships among human keypoint detection, three-dimensional coordinate reconstruction, joint-angle calculation, and robot control. The experiment combines deep-learning algorithms with robot kinematics, communication control, and system debugging, providing students with opportunities to practice programming, system integration, and complex engineering problem solving.

CONCLUSION

This paper integrates an Intel RealSense D435 multimodal visual acquisition device, an RGB-IR bimodal YOLOv8-Pose network, and a robot control platform to construct an interactive robotics teaching system based on multimodal human pose estimation. The system organizes practical teaching activities around RGB-IR image acquisition and alignment, cross-modal feature fusion, human keypoint detection, three-dimensional keypoint reconstruction, joint-angle calculation, and robot control, thereby establishing a complete experimental workflow from multimodal perception to robot execution.

The proposed system connects the main technical stages of robotics, including perception, information processing, decision-making, and control, within a unified practical teaching platform. It provides students with opportunities to integrate knowledge of multimodal vision, deep learning, robot kinematics, and motion control, and offers a practical environment for training in algorithm implementation, system integration, engineering debugging, and interdisciplinary problem solving.

The current teaching system still has limitations in dataset scale, model performance, platform functionality, and educational evaluation. Future work will expand the multi-scenario human pose dataset, improve the real-time performance and robustness of the bimodal network, further enhance the experimental teaching platform, and establish a systematic teaching evaluation framework to assess learning outcomes and the effectiveness of the proposed teaching approach.

ACKNOWLEDGMENT

This work was supported by the National First-Class Undergraduate Course Program of the Ministry of Education of the People's Republic of China (Robotics I) and the Quality Engineering Project of the Anhui Provincial Department of Education (Provincial First-Class Course, Robotics I, Grant No. 2023xxkc130).

REFERENCES

1. Lu, L., Dong, Q., Zhang, T., et al. (2023) Review of robot kinematics and motion planning algorithms, Research on Printing and Digital Media Technology, (5), 1–16. (in Chinese)
2. Xiong, Y., Tang, L., Ding, H., et al. (2008) Fundamentals of Robotics Technology, Huazhong University of Science and Technology Press, Wuhan, China. (in Chinese)
3. Yang, K., Yu, T., Wei, W., et al. (2025) Reform and practice of the blended teaching model for Robot

- Perception Technology under the background of emerging engineering education, *Intelligent Manufacturing*, (3), 124–128. (in Chinese)
4. Xu, G. and Wu, X. (2023) Application of a human pose estimation method based on the LSA-HRNet network in Tai Chi movements, *Journal of South-Central Minzu University (Natural Science Edition)*, 42(6), 839–845. (in Chinese)
 5. Upadhyay, A., Dhupar, B., Sharma, M., et al. (2024) LWIRPose: A novel LWIR thermal image dataset and benchmark, *arXiv preprint*, arXiv:2404.10212.
 6. Tang, H., Zhang, X., Zhuo, S., et al. (2015) High resolution photography with an RGB-infrared camera, in *2015 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10.
 7. Du, J., Zhou, H., Qian, K., et al. (2020) RGB-IR cross input and sub-pixel upsampling network for infrared image super-resolution, *Sensors*, 20(1), 281.
 8. Chen, H., Feng, R., Wu, S., et al. (2023) 2D human pose estimation: A survey, *Multimedia Systems*, 29(5), 3115–3138.
 9. Dong, C., Tang, Y. and Zhang, L. (2024) HDA-Pose: A real-time 2D human pose estimation method based on modified YOLOv8, *Signal, Image and Video Processing*, 18(8), 5823–5839.
 10. Zhang, Y., Yan, W., Shan, T., et al. (2026) Research on passenger posture prediction in subway station based on lightweight improved YOLOv8-Pose, *Electronics*, 15(14), 3031.