Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 19 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i19.58319

PAPER

Development and Evaluation of an AR-Based Interactive System for Occupational Safety Education

Yanfei Lu, Weihang Zhang (✉), Xinjiang Mi

Hebei Key Laboratory of Children's Cognition and Digital Education, Langfang Normal University, Langfang, China

zhangweihang1971@163.com

ABSTRACT

In the field of labor-intensive production, occupational safety represents a fundamental prerequisite for safeguarding workers' health and ensuring the stable development of enterprises. With increasing industrial complexity and the proliferation of safety hazards, traditional approaches to occupational safety education—predominantly reliant on passive knowledge transmission—have proven insufficient in facilitating mastery of safety protocols, frequently resulting in non-compliant behavior during practical operations. Augmented reality (AR) technology has emerged as a promising solution, offering immersive learning environments for safety training. However, limitations persist in existing AR-based safety education systems. Specifically, the use of complex backbone networks in recognition models has resulted in slow running speed and difficulty in meeting real-time interaction requirements. The accuracy of compliance assessment remains suboptimal due to inadequate consideration of scene-specific variations. Furthermore, limited feature extraction capabilities and insufficient attention mechanisms have hindered the system's ability to distinguish between similar operations, thereby compromising recognition precision. Therefore, this study focuses on the development of a compliance assessment method for user-interactive operations based on AR gesture recognition. A lightweight network architecture, ShuffleNetv2, was introduced to replace the original backbone network, thereby reducing computational complexity and enhancing operational efficiency. Additionally, a dynamic selective attention mechanism, SKNet, was integrated into the model to enhance the extraction of critical operational features, thereby improving the accuracy of compliance determination to address limitations identified in prior research.

KEYWORDS

augmented reality (AR) technology, occupational safety education, gesture recognition, compliance assessment, ShuffleNetv2, SKNet

1 INTRODUCTION

Occupational safety has been recognized as a fundamental element in ensuring the health and well-being of workers and in maintaining the stable development of

Lu, Y., Zhang, W., Mi, X. (2025). Development and Evaluation of an AR-Based Interactive System for Occupational Safety Education. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(19), pp. 107–121. https://doi.org/10.3991/ijim.v19i19.58319

Article submitted 2025-05-11. Revision uploaded 2025-06-29. Final acceptance 2025-07-04.

enterprises in modern labor-intensive industries [1–4]. As industrial production has become increasingly complex and diversified [5, 6], potential hazards in the workplace have proliferated [7, 8]. The occurrence of various safety incidents [9–11] has not only resulted in casualties and property damage but has also exerted adverse effects on socio-economic development. Traditional methods of occupational safety education—typically comprising theoretical instruction, video demonstrations, and on-site observations [12, 13]—are largely dependent on passive learning modes. As a result, learners often fail to thoroughly understand or internalize safety protocols embedded in complex operational procedures. Consequently, non-compliant behaviors remain prevalent in practical scenarios. The rapid advancement of augmented reality (AR) technology [14] has opened new avenues for occupational safety education. By integrating virtual safety information with real-world industrial settings, AR has enabled the construction of immersive learning environments that enhance learner engagement and comprehension.

Nonetheless, existing AR-based systems for occupational safety education exhibit several limitations. In particular, the deployment of overly complex backbone networks in recognition models has impaired system responsiveness, rendering them unsuitable for real-time interactive applications. For example, recognition models based on deep learning, as discussed in prior studies [15, 16], have demonstrated high computational overhead and latency when processing complex gesture inputs. Furthermore, the accuracy of compliance assessment in these systems remains inadequate. Some studies [17, 18] have failed to sufficiently account for the variability of features across different operational contexts, thereby introducing bias into judgment outcomes. Specifically, the lack of robust feature extraction mechanisms and attention-based differentiation has hindered the system's ability to distinguish between visually similar operations, resulting in suboptimal recognition precision.

This study focuses on the development of an interactive compliance assessment method for user operations based on gesture recognition within AR environments. In the model architecture, the original backbone network was replaced with the lightweight ShuffleNetv2 framework to reduce computational complexity and enhance system efficiency. Simultaneously, a dynamic selective attention mechanism, SKNet, was integrated to enhance the extraction of salient operational features, thereby improving the accuracy of compliance assessment. This approach is intended to provide an efficient and precise solution for compliance assessment within AR-based interactive occupational safety education systems. The proposed method holds substantial theoretical and practical significance for advancing occupational safety education toward greater intelligence and operational efficiency.

2 METHODOLOGY FOR INTERACTIVE COMPLIANCE ASSESSMENT BASED ON AR GESTURE RECOGNITION

The development process of the interactive occupational safety education system based on AR was structured around system architecture design. The full development flow is illustrated in Figure 1. Initially, system performance was optimized by balancing computational capacity and power consumption, ensuring hardware adaptability. During the real-time performance phase, a system architecture was established with a focus on achieving high accuracy and generalizability. At the core of this architecture lies the interactive compliance assessment method based on AR gesture recognition, which serves as the foundational technology for evaluating

operational correctness in real time by enabling accurate identification of gesture features. Challenges such as gesture scale variability and background interference in complex labor environments were addressed, ensuring immediate and accurate feedback on user operations. This capability is pivotal in realizing the closed-loop objective of real-time performance–accuracy. During the fusion phase, the proposed method enabled seamless adaptation to the AR interface design through efficient feature extraction and data interaction mechanisms, enhancing the immersive integration of user operations with virtual environments and improving the system's interaction friendliness. Each development stage was refined through iterative optimization based on design feedback, and the overall system performance was subsequently validated through comprehensive effect evaluations. As the core technology of the system, the gesture recognition method addresses critical challenges related to efficiency and accuracy in gesture detection under AR conditions. Furthermore, it directly supports the real-time guidance of standardized labor operations.

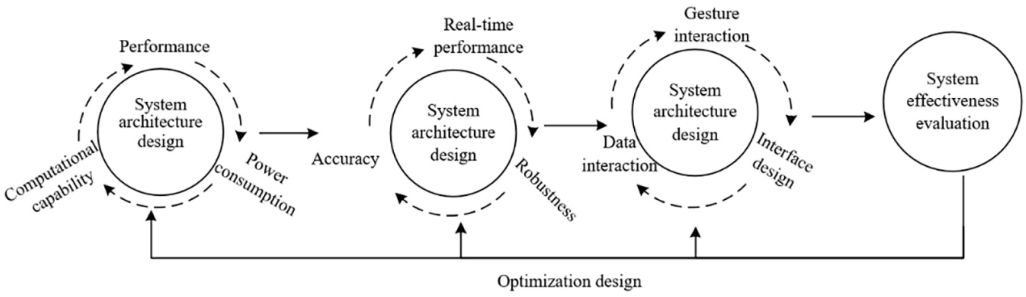

Fig. 1. Development process of the AR-based interactive occupational safety education system

The proposed method for interactive compliance assessment based on AR gesture recognition was designed primarily to meet the real-time performance requirements of interactive occupational safety education systems. To address the inefficiency of conventional gesture recognition models under AR conditions, the lightweight network ShuffleNetv2 was introduced to replace the original backbone in the gesture recognition module. By streamlining the network structure and significantly reducing both the number of parameters and computational complexity, the model size was effectively minimized. This optimization enables rapid acquisition and processing of user gesture data within simulated labor scenarios, even under the constrained hardware resources typical of AR devices. As a result, a responsive and efficient technical foundation is provided for real-time compliance assessment, enhancing both the immediacy and interactivity of the safety education process. To mitigate the potential degradation in feature extraction capability resulting from model lightweighting—which may otherwise compromise the accuracy of compliance assessment—a dynamic selective attention mechanism, SKNet, was further integrated into the ShuffleNetv2 architecture. This mechanism adaptively refines attention to salient features based on the contextual variability of operational stages within labor scenarios. It is particularly effective in environments with background interference or where gestures for compliant and non-compliant actions are visually similar. Key gesture features, such as changes in motion angle or force-related posture patterns, are accurately captured to support reliable differentiation between compliant and non-compliant operations. This capability serves as a key technical safeguard for the implementation of core system functions. The architecture of the AR gesture recognition model is illustrated in Figure 2.

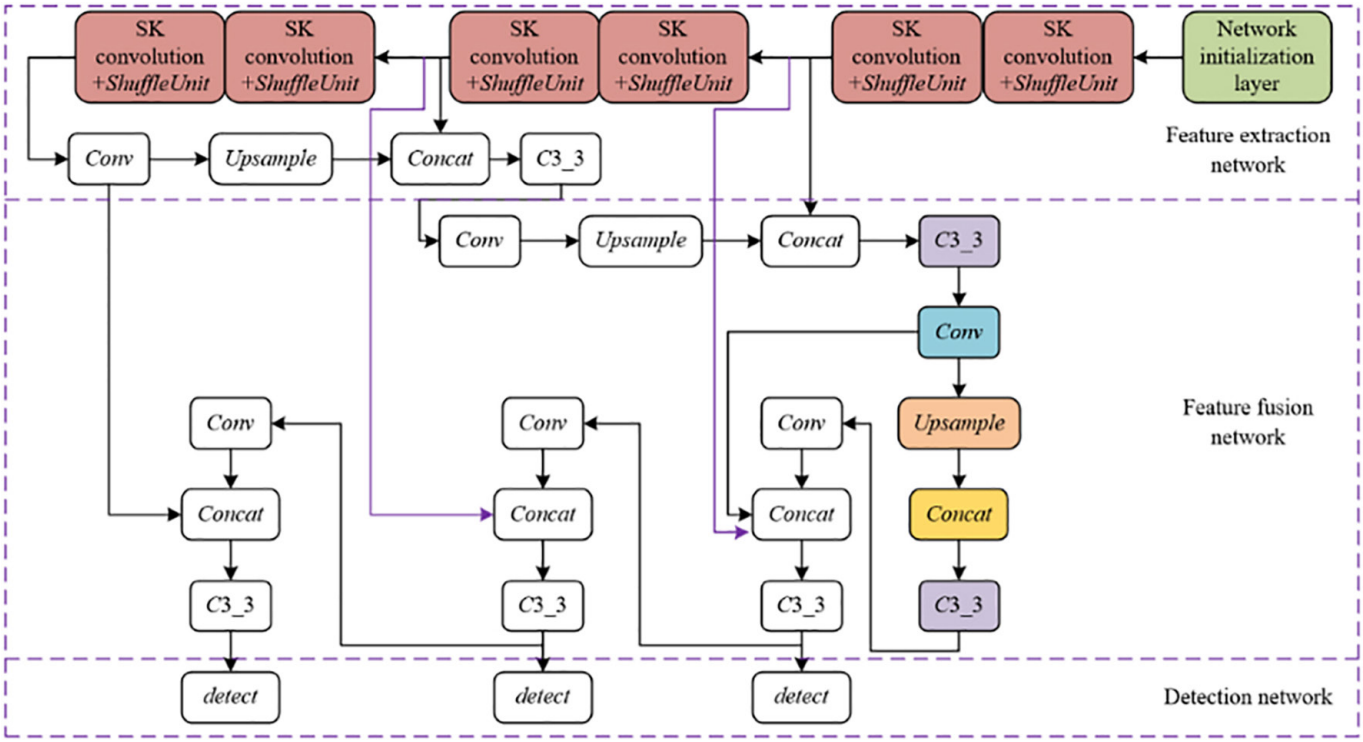

Fig. 2. Architecture of the AR gesture recognition model

2.1 Lightweight network: ShuffleNetv2

In the proposed method for interactive compliance assessment based on AR gesture recognition, the lightweight network ShuffleNetv2 was introduced to replace the original backbone architecture. This modification was primarily motivated by the hardware limitations of AR devices and the real-time performance requirements of compliance assessment. The interactive occupational safety education system relies on AR platforms—such as head-mounted displays and mobile terminals—to integrate virtual scenarios with physical operations. These devices are inherently constrained by their form factor and power consumption, resulting in limited computational and memory resources. Traditional backbone networks typically involve a large number of parameters and high computational complexity, which often lead to inference delays under AR conditions. However, compliance assessment in this context demands millisecond-level responsiveness to user gestures. For instance, during simulated equipment operations, the system must immediately determine whether a gesture conforms to safety protocols and provide feedback. Failure to do so compromises immersion and diminishes the training's instructional effectiveness. ShuffleNetv2, designed specifically for mobile and embedded applications, enables significant reduction in model size while improving inference speed, thereby offering an optimal match for the hardware constraints of AR devices and forming the technical foundation for real-time compliance assessment.

Another key rationale for employing ShuffleNetv2 lies in its advantageous balance between accuracy and speed, which aligns precisely with the system's demands for reliable gesture-based compliance evaluation. Accurate extraction of gesture features is critical for determining operational compliance—for example, distinguishing between correct and unsafe tool handling requires fine-grained recognition capability. ShuffleNetV2 addresses the performance bottleneck observed

in conventional lightweight models, where similar floating-point operation counts (FLOPs) can lead to markedly different inference speeds due to inefficient memory access patterns. By enhancing memory access efficiency, ShuffleNetv2 minimizes data read/write latency without sacrificing recognition accuracy. This combination of efficient computation and high-precision recognition ensures that critical gesture features required for compliance assessment are effectively captured, even within a lightweight framework, meeting the system's accuracy requirements.

The integration of ShuffleNetv2 in place of the original backbone network was guided by the application of its four lightweight design principles, each of which is particularly well-suited to the requirements of gesture-based compliance assessment in AR environments. First, the number of input and output channels is kept equal to minimize memory access overhead. In gesture recognition tasks, input features consist of real-time gesture images captured by AR devices, while the output corresponds to compliance-related feature vectors to be judged. By aligning channel dimensions, the time of data transmission is reduced, thereby accelerating feature propagation and enabling faster processing of continuous gesture frames. Second, the number of groups in grouped convolutions is constrained to prevent excessive memory access that would result from over-partitioning. Since gesture features span multiple dimensions—including texture and contour—moderate grouping reduces computational complexity while preserving cross-group interactions. This allows the model to retain an integrated understanding of the overall gesture structure, which is critical for distinguishing compliant operations. Third, the network structure is simplified to reduce fragmented operations. The ShuffleUnit in ShuffleNetv2 consists solely of a shortcut branch and a main branch, eliminating the inefficiencies associated with parallel multi-branch designs that typically reduce Graphics Processing Unit (GPU) execution efficiency. This enables computational resources on AR devices to be concentrated on core feature extraction, thereby improving the efficiency of continuous gesture processing. Fourth, channel concatenation is employed in place of element-wise addition to reduce memory access consumption. This design choice proves especially advantageous when handling dynamic gestures across sequential operations by accelerating the fusion of features between adjacent frames, thereby ensuring the temporal coherence of gesture recognition.

The network restructuring mechanism of ShuffleNetv2 further enhances its suitability for gesture feature extraction. Within the ShuffleUnit, input features are first divided into two parts using the ChannelSplit operation. After undergoing processing through pointwise convolution (PWConv) and depthwise convolution (DWConv), the main branch is fused with the shortcut branch via channel concatenation (Concat) and then passes through ChannelShuffle, which enables cross-group feature interaction. This architecture supports two operational modes. When the stride is set to 1, the spatial dimensions of the feature maps are preserved, making the structure well-suited for capturing fine-grained gesture details, such as variations in finger joint angles. When the stride is set to 2, downsampling is performed, allowing for the extraction of global gesture contours. The integration of depthwise separable convolution and ChannelShuffle enables parameter reduction without creating isolated feature subspaces. For instance, in the recognition of gestures associated with working at heights, it is necessary to retain both localized finger details and global posture features. The ability to fuse information across channels is thus essential for distinguishing compliant from non-compliant operations. Through this mechanism, ShuffleNetv2 is capable of operating efficiently in AR environments while delivering high-quality gesture features required for real-time compliance assessment. As such, it constitutes a key enabling technology for the core functionality of the interactive occupational safety education system.

2.2 Lightweight feature extraction network

To address the issue of critical information loss during gesture feature extraction in ShuffleNetv2, the dynamic selective attention mechanism, SKNet, was integrated into the network, thereby satisfying the stringent accuracy requirements of compliance assessment. In interactive occupational safety education, gesture inputs are frequently subjected to complex interference. For instance, when simulating machinery operation, gestures may appear smaller due to occlusion by tools or may be confused with the background due to similar coloration between uniforms and the user's skin. These challenges intensify the risk of feature degradation during the channel shuffling process in ShuffleNetv2, which in turn impairs the model's ability to differentiate between subtle actions such as compliant gripping and non-compliant contact. The dynamic selection mechanism of SKNet was introduced to mitigate these limitations. By adaptively adjusting convolutional kernel sizes to accommodate gesture features at varying spatial scales, this mechanism enhances the model's sensitivity to critical motion patterns—such as changes in gesture posture related to angle or applied force—while concurrently suppressing irrelevant background noise. In this way, essential features required for compliance assessment are preserved within the lightweight model framework.

More specifically, SK convolutional units were embedded within the ShuffleUnit of ShuffleNetv2 to enrich feature representation through a decomposition–fusion process. In the decomposition phase, parallel convolutions using multiple kernel sizes were applied to the feature maps in the main branch of the ShuffleUnit. For example, in the recognition of a valve-twisting gesture, small kernels capture fine-grained rotations at the finger joints, while large kernels extract the broader swing of the arm. This multiscale convolutional structure retains the efficiency benefits of ShuffleNetv2 while covering hierarchical features essential to compliance assessment. In the fusion phase, output feature maps from the parallel branches were aggregated. Global average pooling was used to compress spatial information, followed by fully connected layers that generate feature weights, providing a preliminary selection of relevant outputs from the multiscale convolutional kernels, thereby establishing the basis for subsequent dynamic selection.

The key to the proposed method lies in the dynamic adaptability conferred upon the network through a selection mechanism, ultimately enabling the construction of a high-precision, lightweight feature extraction network. During the selection phase, the fused feature weights generated in the fusion phase were applied to the multiple convolutional outputs obtained during the decomposition stage. This weighted fusion allows the network to dynamically assign importance based on real-time gesture characteristics. For instance, when recognizing gestures such as donning protective equipment, if the hand is partially occluded, the network automatically increases the weights corresponding to small convolutional kernels to focus on the exposed fingertip features. Conversely, when gestures are presented against an open background, greater emphasis is placed on the global postural features extracted by larger kernels. This adaptive adjustment addresses the feature loss problem caused by channel shuffling in ShuffleNetv2 and overcomes the rigidity of traditional multi-branch structures such as those used in Inception networks, where convolutional parameters are fixed. By integrating SK units with the depthwise convolutions and channel shuffle operations of ShuffleNetv2, a feature extraction network was constructed that not only satisfies the lightweight constraints of AR devices but also retains the capacity to accurately capture critical gesture features for compliance assessment. This architecture provides a foundational technical guarantee for achieving real-time responsiveness and precise judgment within the interactive occupational safety education system.

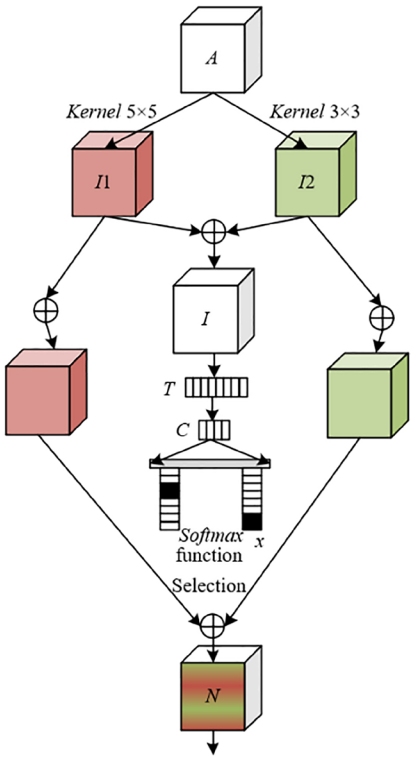

Fig. 3. SK convolutional operation procedure

The SK convolutional operation consists of three primary steps: splitting, fusion, and selection. A schematic representation is provided in Figure 3. In the splitting phase, dual-channel convolution was applied to the input feature maps from the main branch of the ShuffleUnit in ShuffleNetv2. These feature maps were derived from real-time gesture data captured by AR devices and contain critical information such as gesture contours and joint angles, which are essential for compliance assessment. First, a standard 3×3 convolution was employed to extract local detail features—for example, in the simulation of wiring operations, fine-grained features such as the contact angle between fingers and wires are emphasized. Second, a dilated convolution with a dilation rate of 2 was used in place of a 5×5 kernel to expand the receptive field, thereby capturing global gesture patterns such as the swing amplitude of the arms and the spatial relationship between limbs in tasks such as heavy object lifting. This design ensures that both local fine-grained actions and global postural features required for compliance assessment are extracted, all while preserving the lightweight nature of the network. As a result, multi-scale feature representations are provided to distinguish between compliant and non-compliant operations.

The fusion step is intended to integrate dual-channel features and extract global information, thereby laying the foundation for adaptive receptive field adjustment. Initially, the outputs of the 3×3 convolution ($I1$) and the dilated convolution ($I2$) were element-wise summed to perform a preliminary fusion of local and global gesture features. For instance, in recognizing the gesture of donning a safety helmet, both the detailed finger motion involved in fastening the chin strap and the overall spatial relationship between the head and hand are simultaneously retained.

$$I = I1 + I2 \tag{1}$$

Next, global average pooling was applied to compress the spatial dimensions, extracting statistical representations of gesture features while suppressing

redundancy caused by complex AR scene backgrounds such as equipment or tools. Let the global average pooling operation function be denoted by D_{ho}, where the output of the z-th channel is represented by t_z, and the coordinates of the z-th channel are given by $I_z(u,k)$. Let G and Q denote the height and width of the feature map, respectively, and let u and k denote their corresponding coordinates. The expression is as follows:

$$t_z = D_{ho}(I_z) = \frac{1}{G \times Q} \sum_{u=1}^{G} \sum_{k=1}^{Q} I_z(u,k) \tag{2}$$

Finally, a fully connected layer was applied to reduce the dimensionality of the channel-wise features, resulting in a vector z. This process simultaneously reduces computational complexity and emphasizes core features closely associated with operational compliance—such as gesture cues linked to tool gripping force—ensuring that the subsequent selection mechanism can operate in a timely manner, fully compatible with the real-time processing constraints of AR devices. Let the fully connected operation function be denoted as D_{dz}, the nonlinear activation function as σ, and the batch normalization layer as α. A reduction ratio e was used to control the fully connected layer, expressed as f, where the minimum value is M. Let $Q \in R^{f \times z}$, and $c \in R^{f \times m}$. The corresponding mathematical formulation is:

$$c = D_{dz}(t) = \sigma(\alpha(Q_t)) \tag{3}$$

$$f = MAX\left(\frac{z}{e}, M\right) \tag{4}$$

In the selection step, a dynamic attention mechanism was utilized to perform precise selection and fusion of features, directly serving the objective of accurate compliance assessment. Channel-level attention weights were generated based on the vector c, allowing for dynamic adjustment depending on real-time gesture features. For instance, during the simulation of rotary machine operation, if the user's hand is partially occluded by equipment and only a few fingertips are visible, the attention mechanism assigns greater weight to the local detail features extracted by the 3×3 convolution, thus focusing on compliant contact points. Conversely, in scenarios with an unobstructed background, attention is shifted toward the global operational trajectory captured by the dilated convolution. Let $X, Y \in R^{z \times f}$, and let the attention vectors corresponding to $I1$ and $I2$ be denoted by x_z and y_z, respectively. The z-th row is denoted as X_z, and the z-th element of x is represented as x_z. The process is then defined as:

$$\begin{cases} x_z = \dfrac{r^{X_z c}}{r^{X_z c} + r^{Y_z c}} \\ y_z = \dfrac{r^{Y_z c}}{r^{X_z c} + r^{Y_z c}} \end{cases} \tag{5}$$

The weights were subsequently applied to perform weighted fusion of the dual-channel features, resulting in the final feature vector N_z. This dynamic selection mechanism not only compensates for feature loss caused by the channel shuffling process in ShuffleNetv2 but also selectively enhances gesture features critical to compliance assessment. For instance, spatial features used to distinguish between compliant inbound operations and non-compliant boundary-crossing gestures are effectively emphasized. As a result, the fused lightweight network maintains high computational efficiency under AR constraints while delivering precise

compliance judgments for the interactive occupational safety education system. Let $N = [N_1, N_2 \ldots N_z]$, where $N_z \in R^{G \times Q}$. The computation is expressed as:

$$N_z = x_z I_z + y_z I_z; x_z + y_z = 1 \tag{6}$$

The proposed method for interactive compliance assessment based on AR gesture recognition adopts a modular network architecture in which the core building block is a newly designed unit that integrates SK convolution modules with the ShuffleUnit structure. A multi-stage stacking strategy was employed to construct a lightweight and high-performance feature extraction network. This new module adopts a residual structure embedded with lightweight attention mechanisms, overcoming the limitations of ShuffleNetv2 that relies solely on 3×3 depthwise convolution. For the case of stride = 1 (New Module-1), the input features were first split using ChannelSplit into a shortcut branch and a main branch. The shortcut branch preserves the original features to minimize information loss, while the main branch is dimensionally reduced via 1×1 convolution and subsequently passed through a residual block containing the SK module. This pathway is designed to enhance fine-grained gesture features relevant to compliance, such as tool manipulation angles. Afterward, their original dimensions were restored via another 1×1 convolution. Finally, both branches were concatenated to maintain consistent feature map dimensions and channel count. ChannelShuffle was then applied to achieve cross-group information fusion, ensuring effective interaction of local fine-grained actions and global operational postures.

To accommodate dynamic gesture variations, the stride = 2 version of the newly constructed module—New Module-2—was structurally optimized. In this configuration, the shortcut branch incorporates a 3×3 depthwise convolution with stride = 2 followed by a 1×1 convolution. This design enables feature map downsampling, thereby capturing global gesture patterns over a wider spatial range. Meanwhile, the main branch preserves the processing flow of New Module-1, consisting of dimensionality reduction, residual block, and dimensionality restoration. This ensures that critical fine-grained features are retained during downsampling. The differential design allows the network to respond to both close-range, detailed gestures and large-scale limb movements across diverse operational contexts.

The overall architecture of the network—based on the newly designed modules—was constructed through multi-stage stacking to enable progressive feature extraction. Each stage selectively stacks either New Module-1 or New Module-2 according to the stride parameter. When $S = 1$, New Module-1 is stacked to preserve feature map dimensions and to focus on in-depth learning of gesture detail features. When $S = 2$, New Module-2 is stacked to perform downsampling, thereby progressively expanding the receptive field to capture more global features associated with broader operational contexts. A channel scaling factor $g = 1$ was adopted to control the number of output channels, ensuring sufficient feature richness while avoiding unnecessary increases in model complexity, thus maintaining compatibility with the limited computational resources of AR devices.

The core advantages of this architecture are reflected in three key aspects: a) By integrating residual structures with attention mechanisms within the newly designed modules, the lightweight network significantly enhances feature extraction capabilities, effectively addressing the issue of critical feature loss caused by channel shuffling in ShuffleNetv2. As a result, fine-grained distinctions between compliant operations and visually similar non-compliant actions can be accurately achieved. b) The differentiated design of the two module variants allows for simultaneous focus on fine gesture details and global limb movements, thereby adapting to diverse training scenarios in interactive occupational safety education—from precision tool manipulation to large-scale

body motions. c) By limiting the number of channels and simplifying the network branching structure, the architecture preserves its lightweight characteristics. This ensures high-speed inference in AR devices while satisfying the real-time compliance assessment requirements. Together, these attributes provide core "high-precision + high-efficiency" technical support for the interactive occupational safety education system.

3 EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, the performance of ShuffleNetv2 demonstrates a clear advantage in terms of lightweight characteristics, with a parameter count of 3.1 million, a model size of 8.3 MB, and an inference time of 7.12 ms. The mean Average Precision (mAP) reached 96.2%, which is substantially higher than that of GhostNet (94.6%) and only marginally lower than that of EfficientNet (97.5%). In terms of computational cost, ShuffleNetv2 reported 4.8 billion FLOPs, comparable to GhostNet's 4.2 billion, while maintaining a relatively efficient inference speed. Given the computational constraints of AR devices, ShuffleNetv2 was adopted as the backbone network due to its effective balance between accuracy and model efficiency. When integrated with the dynamic selective attention mechanism, SKNet, the ability to extract critical features related to operational compliance was further enhanced, enabling improvements in mAP without incurring significant additional computational cost or latency. Experimental data confirm that ShuffleNetv2 outperforms EfficientNet in lightweight metrics and exceeds GhostNet in recognition accuracy, making it an ideal structural foundation for embedding the SKNet attention mechanism. The final proposed architecture—combining ShuffleNetv2-based lightweight modeling with SKNet-based feature enhancement—satisfies the real-time performance requirements of AR systems while improving gesture recognition accuracy, thereby substantially increasing the precision of interactive gesture-based compliance assessment without compromising system efficiency.

Table 1. Performance comparison of different feature extraction networks

Feature Extraction Network	mAP (%)	Parameters (M)	FLOPs ($\times 10^9$)	Model Size (MB)	Inference Time (ms)
EfficientNet	97.5	12.5	16.8	25.6	11.23
GhostNet	94.6	3.4	4.2	8.8	6.54
ShuffleNetv2	96.2	3.1	4.8	8.3	7.12

Table 2. Comparative performance of different algorithms and the proposed method

Model	mAP (%)	Parameters (M)	Model Size (MB)	Inference Time (ms)
Faster-RCNN	946	61.2	148	43.21
SSD	92.3	24.5	91.2	21.56
YOLOv5s	93.7	61.2	223	15.89
OpenPose	95.6	7.12	14.6	9.12
MediaPipe	97.4	12.6	25.8	11.23
Proposed method	97.2	3.2	11.2	7.46

As shown in Table 2, the proposed method demonstrates outstanding performance across multiple critical metrics. The mAP reached 97.2%, slightly below MediaPipe (97.4%) but with significantly superior model compactness and inference speed, achieving a balanced trade-off between high accuracy, lightweight design,

and fast inference. When compared to classical object detection algorithms, the proposed method exhibited an mAP improvement of 2.6%, 4.9%, and 3.5%, respectively. Furthermore, the model size was reduced to just 7.57%, 12.28%, and 5.04% of those same models, while inference time was shortened to 17.26%, 34.6%, and 47.0%, respectively. These results reflect consistent superiority in terms of accuracy, efficiency, and speed. By integrating ShuffleNetv2, model complexity was substantially reduced, enabling deployment under the resource constraints of AR devices. The incorporation of the dynamic selective attention mechanism (SKNet) further enhanced the model's ability to extract critical gesture features in complex labor scenarios. As a result, mAP performance approached the high accuracy level of MediaPipe while maintaining an inference time of only 7.46 ms—suitable for real-time interaction. The experimental findings indicate that the proposed method effectively overcomes the limitations of traditional algorithms in AR-based gesture recognition scenarios, which are often burdened by model size and slow inference. At the same time, it addresses the lightweight inadequacies of algorithms such as MediaPipe. The resulting combination of high accuracy, lightweight design, and rapid inference provides core technical support for the interactive occupational safety education system.

Table 3. Performance comparison of each improved model

Model	ShuffleNetv2	SKNet	mAP (%)	FLOPs ($\times 10^9$)	Model Size (MB)	Inference Time (ms)
YOLOv5s			97.5	16.8	25.6	11.23
Improved model 1	√		96.2	4.8	8.3	7.14
Proposed method	√	√	97.5	5.2	11.5	7.45

As shown in Table 3, the proposed method achieved an mAP of 97.5%, equivalent to that of You Only Look Once version 5—small variant (YOLOv5s), while offering significant improvements in model compactness, FLOPs, and inference time. This demonstrates the attainment of both lightweight deployment and acceleration without sacrificing recognition accuracy. In comparison with Improved Model 1, which integrates only ShuffleNetv2, the proposed method achieved a 1.3% improvement in mAP. This enhancement is attributed to the incorporation of SKNet, which effectively strengthens the extraction of key operational features—particularly in capturing multi-scale gesture details and distinguishing compliant gestures from background interference in complex labor scenarios, thereby compensating for the accuracy trade-off typically associated with lightweight architectures such as ShuffleNetv2. Specifically, the lightweight design of ShuffleNetv2 ensures compatibility with the limited computational resources of AR devices, while SKNet's dynamic attention mechanism enables precise capturing of differences between compliant and non-compliant gestures. As a result, the proposed method retains the advantages of a compact architecture while restoring the level of precision. The experimental results confirm that the proposed method successfully resolves the bulky structure and slower inference speed of YOLOv5s while overcoming the accuracy limitations commonly associated with lightweight networks.

The training results illustrated in Figure 4 indicate that the loss values of both training and validation sets exhibit a rapid decline followed by gradual convergence. The final training set loss dropped below 0.5, while the validation set loss remained only marginally higher, indicating that the model converged well during the training process and effectively avoided overfitting problems, possessing stable learning ability. Moreover, the accuracy curves demonstrate a clear upward trend. The training set accuracy approached near-perfect performance, while the validation set accuracy stabilized

above 80%, indicating that the model not only achieved high fitting capability on the training data but also maintained robust generalization on previously unseen data.

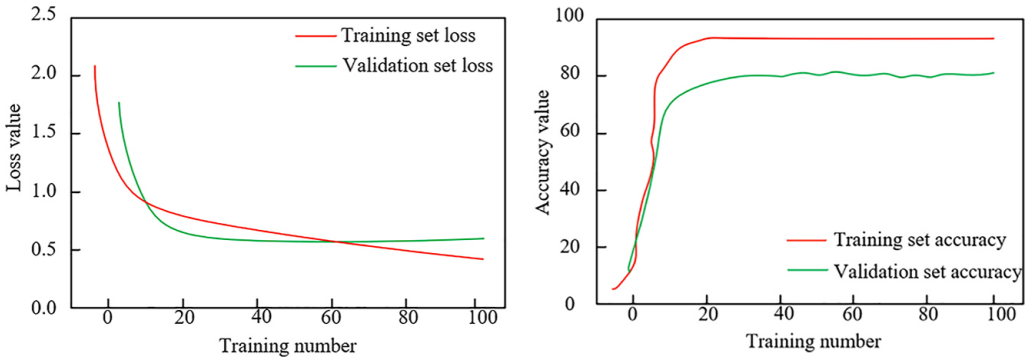

Fig. 4. Training curves of the proposed model

As illustrated in Figure 5, the classification accuracy across all gesture categories remained consistently high. For instance, Gestures 4 and 5 achieved prediction accuracies of 0.96 and 0.98, respectively, indicating that the proposed model exhibited strong recognition capability for standard-compliant gestures by accurately capturing their defining features. Off-diagonal elements of the confusion matrix were generally low, with mutual misclassification rates between Gestures 1 and limited to 0.04 and 0.08, respectively. These results demonstrate the model's high discriminative capacity, even for visually similar gestures, thereby reducing the likelihood of compliance misjudgments due to gesture resemblance. This performance can be attributed to the integrated design strategy, in which the lightweight ShuffleNetv2 backbone was employed to reduce computational complexity, while the SKNet dynamic attention mechanism was incorporated to enhance the extraction of salient operational features in labor-intensive scenarios. As a result, superior classification accuracy was attained on the validation set. In summary, the experimental findings confirm that the synergistic use of a lightweight network and dynamic attention mechanism led to outstanding gesture recognition accuracy. The proposed method was shown to accurately identify user operation gestures in complex labor settings, thereby verifying its practical applicability and effectiveness within AR-based interactive occupational safety education systems. This performance foundation strongly supports further deployment and development of such systems.

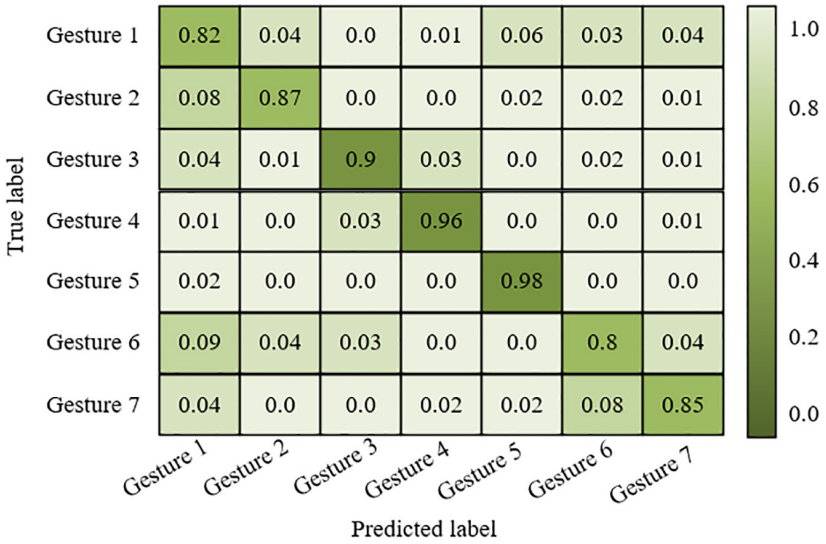

Fig. 5. Confusion matrix of the algorithm on the validation set

4 CONCLUSION

A methodology for assessing user interactive operational compliance in AR environments based on gesture recognition was developed. A lightweight backbone network, ShuffleNetv2, was employed to replace the original feature extractor, and innovative modules were constructed by integrating the SKNet dynamic selective attention mechanism. This design enabled the construction of a feature extraction network that balances computational efficiency and recognition accuracy. The experimental results demonstrated that the architecture, through the differentiated design of innovative modules with strides of 1 and 2, successfully captured both fine-grained gesture features—such as tool handling angles—and global operational features—such as body safety distances. The proposed approach significantly enhanced the accuracy of operational compliance assessment while maintaining a lightweight model structure. Precise differentiation was achieved between visually similar operations such as "standard tool gripping" and "finger overreach." As a result, the proposed method provided core technical support for interactive occupational safety education systems and addressed the challenge of balancing efficiency and accuracy in AR environments, enabling real-time guidance on safety compliance, thereby contributing to enhanced immersion and practical effectiveness in occupational safety education.

5 ACKNOWLEDGEMENT

This paper was funded by the Higher Education Teaching Reform Research and Practice Project of Hebei Province: Construction and practice of the curriculum system of "Labor Education" in local normal universities under the background of application transformation development—based on the perspective of "vocational maturity" (Grant No.: 2023GJJG371); the Higher Education Teaching Reform Research and Practice Project of Hebei Province: Construction and Practice of Innovation and Entrepreneurship Education Curriculum System Based on "Whole Field Double Thread" (Grant No.: 2023GJJG364); and the Research and Practice Project on Teaching Reform of Innovation and Entrepreneurship Education of Hebei Provincial Department of Education: Construction of Innovation and Entrepreneurship Education Curriculum System Based on the Linkage of Three Courses: A Case Study of Primary Education (Grant No.: 2023CXCy175).

6 REFERENCES

[1] K. Zhou, Q. Wang, and J. Tang, "Tripartite evolutionary game and simulation analysis of coal mining safe production supervision under the Chinese central government's reward and punishment mechanism," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 5298890, 2021. https://doi.org/10.1155/2021/5298890

[2] M. Chansunthorn and P. Pochanart, "Risk assessment in developing occupational standards for environmental work in Thailand," *International Journal of Sustainable Development and Planning*, vol. 19, no. 1, pp. 187–196, 2024. https://doi.org/10.18280/ijsdp.190117

[3] Ş. Toptancı, "Ergonomic performance evaluation in Türkiye's metal industry: Occupational health and safety indicators through VIKOR methodology," *Journal of Operational and Strategic Analytics*, vol. 1, no. 3, pp. 140–146, 2023. https://doi.org/10.56578/josa010305

[4] L. Aouragh, N. Ouazraoui, L. Boubaker, N. Bourmada, and A. Sekhri, "Integrating human performance factors to improve occupational risk assessment," *International Journal of Safety and Security Engineering*, vol. 15, no. 3, pp. 609–619, 2025. https://doi.org/10.18280/ijsse.150319

[5] M. D. Korneeva, "Modeling of production-economic relations in innovation-industrial complexes," *Automation and Remote Control*, vol. 67, pp. 161–165, 2006. https://doi.org/10.1134/S0005117906010103

[6] C. S. Longo and C. Fantuzzi, "Simulation and optimization of industrial production lines," *at-Automatisierungstechnik*, vol. 66, no. 4, pp. 320–330, 2018. https://doi.org/10.1515/auto-2017-0126

[7] E. S. Erondu and P. E. Anyanwu, "Potential hazards and risks associated with the aquaculture industry," *African Journal of Biotechnology*, vol. 4, no. 13, pp. 1622–1627, 2005. https://doi.org/10.4314/ajfand.v4i13.71775

[8] K. Bhattacharjee, S. Chaudhary, A. Vishnoi, D. A. Patel, and N. Bugalia, "Characterization of health and safety hazards of deconstruction activities," *American Journal of Industrial Medicine*, vol. 68, no. S1, pp. S71–S87, 2025. https://doi.org/10.1002/ajim.23652

[9] N. N. Hien, N. T. Kiet, and A. T. Nguyen, "The influence of safety culture on safety attitude, personnel error behavior, and safety citizenship behavior: Research in the Vietnam oil and gas industry," *International Journal of Safety and Security Engineering*, vol. 14, no. 2, pp. 399–409, 2024. https://doi.org/10.18280/ijsse.140208

[10] Y. A. Kim, B. Y. Ryoo, Y. S. Kim, and W. C. Huh, "Major accident factors for effective safety management of highway construction projects," *Journal of Construction Engineering and Management*, vol. 139, no. 6, pp. 628–640, 2013. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000640

[11] M. Damjanović, A. Petrović, V. Ilić, M. Radetić, and P. Stanojević, "Risk assessment in the transportation of dangerous goods: Application of ALOHA and GIS tools in Montenegro," *Journal of Operational and Strategic Analytics*, vol. 2, no. 4, pp. 254–265, 2024. https://doi.org/10.56578/josa020404

[12] C. Ma, M. Jing, S. Hou, J. Jiang, and B. Zhang, "Current status of safety engineering education in China," *Process Safety Progress*, vol. 41, no. 2, pp. 218–225, 2022. https://doi.org/10.1002/prs.12306

[13] S. E. Lee, V. S. Dathinten, and H. Do, "Patient safety education in pre-registration nursing programmes in South Korea," *International Nursing Review*, vol. 67, no. 4, pp. 512–518, 2020. https://doi.org/10.1111/inr.12630

[14] K. Sdravopoulou, J. J. G. Castillo, and J. M. M. González, "Naturalistic approaches applied to AR technology: An evaluation," *Education and Information Technologies*, vol. 26, pp. 683–697, 2021. https://doi.org/10.1007/s10639-020-10283-4

[15] F. Wang, X. Ao, M. Wu, S. Kawata, and J. She, "Explainable deep learning for sEMG-based similar gesture recognition: A Shapley-value-based solution," *Information Sciences*, vol. 672, p. 120667, 2024. https://doi.org/10.1016/j.ins.2024.120667

[16] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, pp. 3941–3951, 2017. https://doi.org/10.1007/s00521-016-2294-8

[17] F. Jalundhwala and V. Londhe, "A systematic review on implementing operational excellence as a strategy to ensure regulatory compliance: A roadmap for Indian pharmaceutical industry," *International Journal of Lean Six Sigma*, vol. 14, no. 4, pp. 730–758, 2023. https://doi.org/10.1108/IJLSS-04-2022-0078

[18] F. G. Habtemichael and L. de Picado Santos, "Safety and operational benefits of variable speed limits under different traffic conditions and driver compliance levels," *Transportation Research Record*, vol. 2386, no. 1, pp. 7–15, 2013. https://doi.org/10.3141/2386-02

7 AUTHORS

Yanfei Lu graduated from Renmin University of China in 2018, works at Langfang Normal University, and is engaged in art education (E-mail: RDYS1937LYF@163.com).

Weihang Zhang graduated from the University of Waikato, New Zealand, in 2003, works at Langfang Normal University, and is engaged in teacher education (E-mail: zhangweihang1971@163.com).

Xinjiang Mi graduated from Hebei University in 1985, works at Langfang Normal University, and is engaged in Computer education (E-mail: mixinjiang1963@163.com).