RepCA-GCN: A Lightweight and Efficient Method for Real-Time Construction Worker Action Recognition via Structural Re-parameterization

Jinhui Liang[1], Hui Deng[1], Zhou Zhang[2], Wenhao Li[1], Yichuan Deng[1, *]

[1] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China.
[2] School of Management Science and Engineering, Guizhou University of Finance and Economics, Guizhou, China.
jhliang29@gmail.com, hdeng@scut.edu.cn, zhangzhou@mail.gufe.edu.cn, 19098092885@163.com, ctycdeng@scut.edu.cn

Abstract –

Accurate and real-time recognition of construction workers' behaviors is critical to intelligent safety supervision and productivity assessment. However, existing skeleton-based methods struggle to balance recognition accuracy with inference latency, while lacking robustness against noisy keypoints caused by frequent occlusions. This study proposes RepCA-GCN, a high-efficiency architecture specifically tailored for construction site action recognition. The method integrates Structural Re-parameterization to decouple training capacity from inference latency, Coordinate Attention to capture long-range spatial dependencies, and a Confidence Gating mechanism to suppress noise from unreliable pose estimations extracted via YOLOv11-Pose. Validated on our constructed SiteWorker Action Recognition (SWAR) dataset of 7,424 video clips, we systematically explored skeletal configurations, finding that a streamlined 14-keypoint scheme helps to mitigate facial keypoint noise. Experimental results demonstrate that RepCA-GCN achieves 95.35% accuracy, matching state-of-the-art complex models (e.g., CTR-GCN) while significantly reducing FLOPs to 0.413G and achieving a low latency of 1.36 ms. Furthermore, a deployed monitoring system utilizing a dual sliding-window strategy achieves over 60 FPS in multi-person scenarios on a workstation equipped with an RTX 4090 GPU. This framework offers a superior speed-accuracy trade-off, making it well-suited for edge deployment in dynamic construction environments.

Keywords –

Construction Safety; Action Recognition; Pose Estimation; Structural Re-parameterization; Graph Convolutional Network; Real-time Monitoring

1 Introduction

The construction industry remains a high-risk sector, where unsafe behaviors of workers are a primary cause of accidents [1,2]. While traditional manual inspections suffer from insufficient coverage and high costs [3], the advent of computer vision has enabled automated monitoring using standard RGB cameras, which are cost-effective and non-intrusive compared to wearable sensors [4,5] or depth cameras [6].

In recent years, skeleton-based action recognition has emerged as a promising direction. By extracting human keypoints using pose estimation models (e.g., OpenPose [7], YOLO-Pose [8]) and processing them with deep learning networks, researchers can effectively classify worker actions [9,10]. Compared with Long Short-Term Memory (LSTM) based methods [10], the Spatio-Temporal Graph Convolutional Network (ST-GCN) [11] serves as a more advanced foundational framework with superior performance. However, applying ST-GCN directly to construction sites faces two critical bottlenecks:

Noise Sensitivity: Construction sites are cluttered, leading to frequent occlusions. Standard GCNs treat all keypoints equally, allowing low-confidence (occluded) keypoints to introduce noise and degrade accuracy.

Computational Efficiency vs. Accuracy: Advanced GCN variants (e.g., 2s-AGCN [12], CTR-GCN [13]) improve accuracy but introduce heavy computational burdens (high FLOPs and latency), making them unsuitable for real-time edge deployment on site. Conversely, lightweight models often sacrifice recognition performance.

To bridge this gap, this paper proposes RepCA-GCN, a lightweight yet powerful network tailored for real-time construction monitoring. We leverage Structural Re-parameterization technology [14], which allows the model to utilize a complex multi-branch architecture for feature extraction during training but collapses into a single, efficient convolution layer during inference. This effectively decouples training capacity from inference speed. Furthermore, we integrate Coordinate Attention (CA) to enhance spatial feature representation and a Confidence Gating mechanism to filter noise based on keypoint reliability.

Our main contributions are:

We propose RepCA-GCN, incorporating Structural Re-parameterization modules, Coordinate Attention, and Confidence Gating. This architecture achieves SOTA accuracy (95.35%) with significantly lower latency (1.36ms) compared to existing methods.

We construct the SiteWorker Action Recognition (SWAR) dataset. To enhance robustness, we employ a rigorous occlusion-aware data augmentation strategy (including scaling, flipping, and random occlusion) on the training set, expanding the data scale to four times its original size to ensure generalization.

We validate the system's effectiveness through comprehensive experiments, demonstrating its capability for both safety monitoring (e.g., fall detection) and efficiency analysis in real-world scenarios.

2 Related Work

2.1 Vision-based Construction Safety Monitoring

Computer vision has been widely adopted for site safety. Early works focused on object detection, such as detecting Personal Protective Equipment (PPE) using YOLO or Faster R-CNN [15–17]. While effective for static compliance checking, these methods fail to capture temporal dynamics required for identifying complex actions (e.g., working vs. idling). Video-based approaches like 3D-CNNs [18] and Two-Stream networks [19] can capture motion but suffer from high computational costs, limiting real-time application [20].

2.2 Skeleton-based Action Recognition

Skeleton-based methods offer a balance between performance and efficiency. Yan et al. [11] proposed ST-GCN to model skeletal data as graphs. In construction, Wu et al. [9] and Zhou et al. [10] applied pose-based methods for behavior analysis. Zhou et al. [10] utilized a lightweight LSTM model; however, RNN-based structures lack the ability to explicitly model spatial structural dependencies compared to GCNs.

Recent advancements in GCNs focus on adaptive graphs (e.g., 2s-AGCN, CTR-GCN) to improve accuracy, but their complex topology computations increase inference latency. To address the efficiency issue, Structural Re-parameterization (e.g., RepVGG [14]) has gained attention in CNNs for speeding up inference without accuracy loss. In this study, we adapt this technique to the temporal modeling of GCNs, combined with YOLOv11-Pose, to achieve a high-performance, real-time solution for construction sites.

3 Methodology

To address the challenges of occlusion and high computational latency in construction site monitoring, we propose a high-efficiency framework for worker action recognition. As illustrated in Figure 1, the system operates in a two-stage pipeline. First, the YOLOv11-Pose model is employed to extract spatiotemporal human skeletal keypoint sequences from raw video streams. These sequences, which contain both coordinate data and confidence scores, serve as the input for the subsequent stage. Second, the proposed RepCA-GCN model processes these skeletal sequences to classify worker behaviors. This decoupled design allows for the independent optimization of pose estimation and action recognition, ensuring both precision and system flexibility.

3.1 RepCA-GCN Model Architecture

The core of our methodology is the RepCA-GCN (Reparameterized Coordinate-Attention Graph Convolutional Network). As shown in Figure 2, the network processes a skeletal sequence with dimensions (B, T, N, C). The backbone comprises four stacked RepBlock stages. To primarily resolve the conflict between model capacity and real-time inference speed, we employ Structural Re-parameterization in the temporal modeling layers. During the training phase, the temporal block adopts a multi-branch topology consisting of parallel 5×1, 3×1, 1×1 convolutions, and an identity mapping. This design enables the model to learn multi-scale temporal features. However, for deployment, these branches are mathematically fused into a single 9×1 convolutional layer via linear transformations of kernels and biases. This decoupling ensures the model retains the representational power of a complex topology while achieving the low memory access cost and latency of a single-path network during inference.

Complementing this efficient architecture, we integrate Coordinate Attention to refine features. Unlike standard channel attention, Coordinate Attention aggregates the $C \times T \times N$ skeletal graph features along distinct temporal (T) and spatial (N) directions, capturing long-range dependencies with minimal overhead. Additionally, a Confidence Gating mechanism is applied. By using the input confidence score c_n to modulate features via $X_{\text{out}} = X_{\text{gcn}} \odot \sigma(W \cdot c_n + b)$, the model effectively suppresses noise from occluded keypoints.

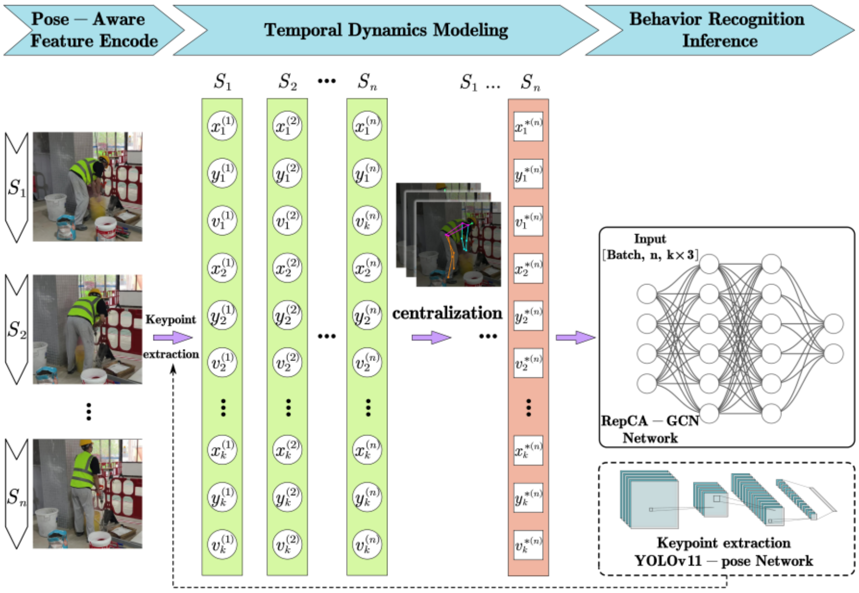

Figure 1. The Proposed Methodological Framework for Skeleton-Based Action Recognition

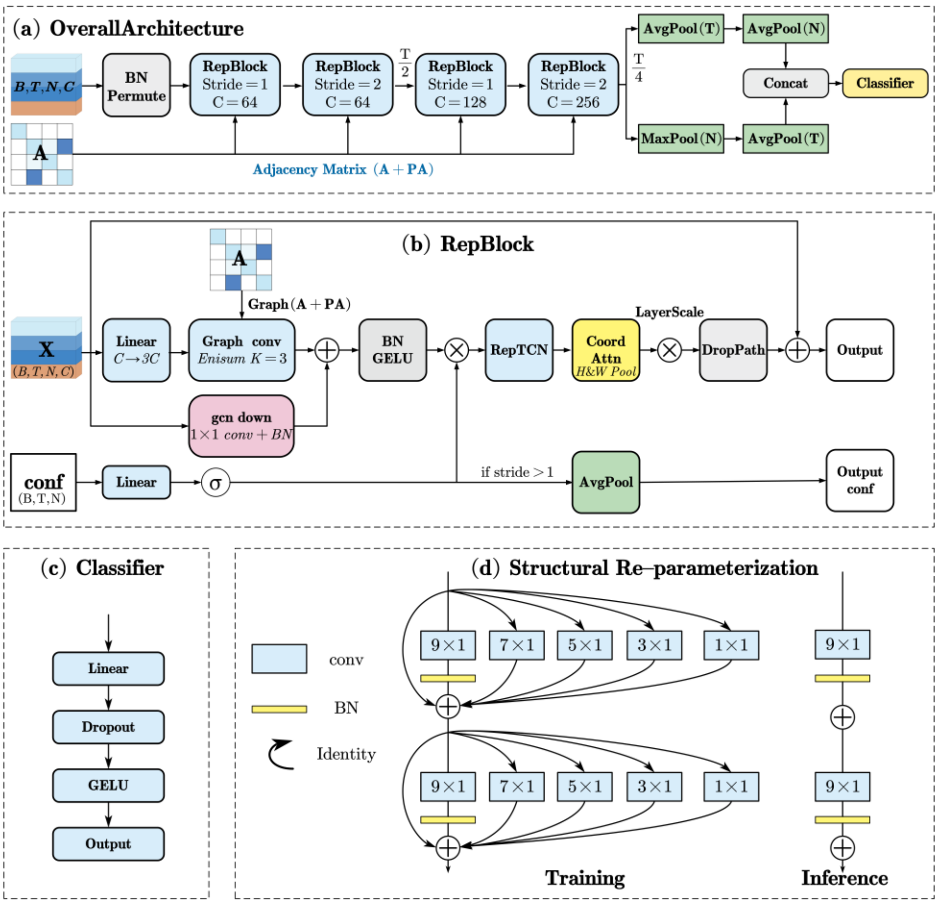

Figure 2. RepCA-GCN Model Architecture

3.2 Data Preparation and Workflow

The comprehensive workflow, ranging from data preparation to model training and practical recognition, is depicted in Figure 3. The process begins with the construction of the SiteWorker Action Recognition (SWAR) dataset. We collected video data from diverse sources, including real-world construction site surveillance, drone footage, and simulated environments, to ensure scenario diversity.

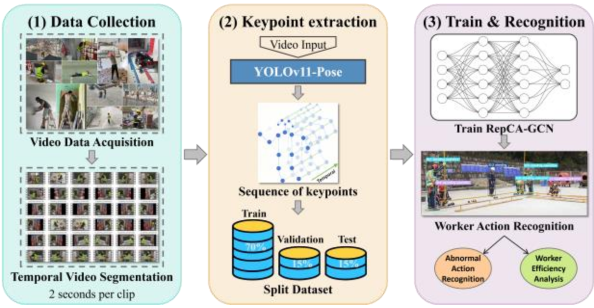

Figure 3. Workflow of data preparation, model training and recognition for the RepCA-GCN

As shown in Figure 4, the dataset is collected from actual construction sites, covering various environmental conditions and camera angles. It includes multiple categories of behaviors recorded under different lighting levels to reflect the complexity of real-world scenes.

Figure 4. SWAR dataset

The composition of the dataset is analyzed in Figure 5. The dataset exhibits a relatively balanced distribution across the six behavioral categories. 'Squatting operation' constitutes the largest portion (24.2%), while 'fall' represents the smallest category (13.1%), thereby mitigating the risk of severe class imbalance. The 'squatting operation' and 'standing operation' categories specifically denote the active construction states of workers in their respective postures. Most videos in the SWAR dataset are in 1080p resolution (1920 × 1080 or 1080 × 1920), with a small portion recorded at 720p. The specific number of video clips for each category is summarized in Table 1.

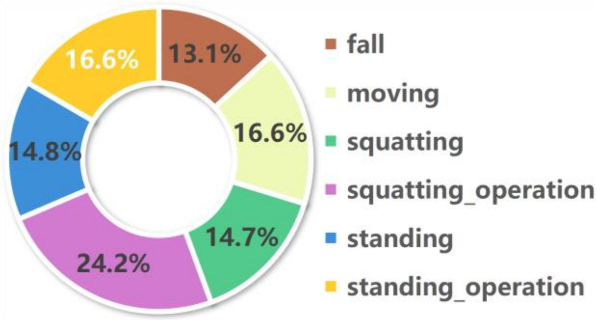

Figure 5. Composition of the SWAR Dataset

Table 1. Number of Video Clips in the SWAR Dataset

Behavior	Count
Fall	976
Moving	1230
Squatting	1091
Squatting operation	1794
Standing	1100
Standing operation	1233
Total	7424

For skeletal representation, we adopt the standard 17-keypoint configuration defined by the COCO format (Figure 6), with all coordinates undergoing normalization and centralization to ensure spatial consistency. Additionally, alternative 13-keypoint and 14-keypoint schemes (Figure 7) were explored and evaluated in the subsequent experiments.

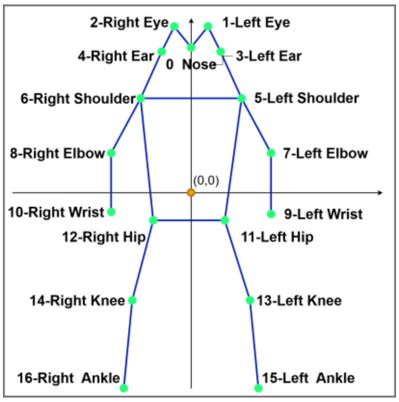

Figure 6. COCO 17 Keypoints Layout

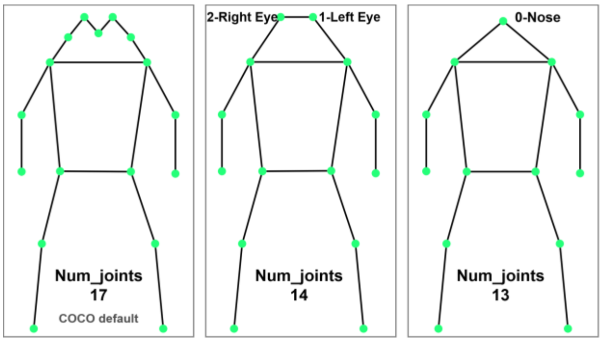

Figure 7. Three Keypoint Selection Schemes

4 Experiments and Results

4.1 Experimental Setup

The dataset partitioning and augmentation strategy are detailed in Table 2. The original SWAR dataset was split into training, validation, and testing sets with a ratio of 70:15:15. To strictly prevent data leakage and ensure realistic evaluation, we implemented a hybrid splitting strategy. Specifically, 30% of the validation and testing samples were randomly selected to ensure global data representation, while the remaining data were allocated sequentially based on video indices. Because raw videos capture continuous actions of identical subjects, this sequential allocation guarantees that frames from the same subject do not overlap across the training and testing sets.

Table 2. Dataset Partition

Subset	Original Samples	Augmented Samples
Train	5195	20780
Validation	1111	-
Test	1118	-
Total	7424	-

To enhance robustness against scale variations and frequent occlusions on construction sites, we exclusively applied three distinct data augmentation pipelines to the training set: (1) adaptive spatial scaling combined with a 50% probability of horizontal flipping; (2) scaling and flipping coupled with random occlusion; and (3) pure random occlusion. The occlusion operation intentionally masks 1 to 2 skeletal joints to simulate the noisy keypoints commonly generated by pose estimators in cluttered environments. This process expanded the training set to four times its original size.

All experiments were conducted on an Ubuntu workstation equipped with an NVIDIA RTX 4090 GPU using PyTorch 2.5.1. The models were trained for 50 epochs with a batch size of 128, optimized via AdamW with a cosine annealing learning rate schedule (decaying from 1e-3 to 1e-5) and gradient clipping for training stabilization.

4.2 Performance Analysis

The training dynamics of the proposed model are illustrated in Figure 8, demonstrating stable convergence without severe overfitting, owing to the rigorous anti-leakage data splitting and occlusion-aware augmentation. As presented in Table 3, the SWAR dataset under the new strict partition proves to be a challenging benchmark. RepCA-GCN achieves a highly competitive accuracy of 92.58%, outperforming the baseline ST-GCN (91.32%) and performing on par with the computationally heavy 2s-AGCN (92.47%).

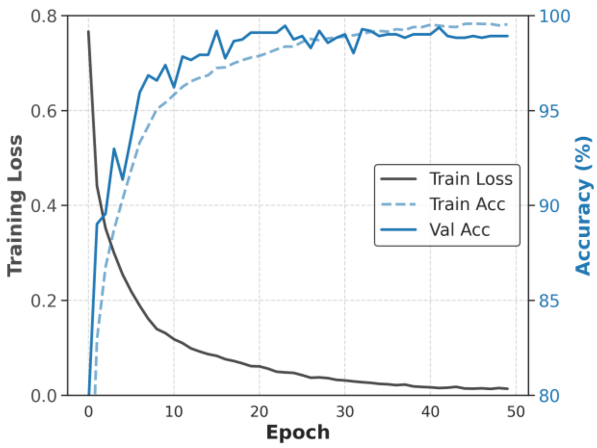

Figure 8. Training and validation results of RepCA-GCN

Table 3. Comparative performance of different models on the SWAR dataset (17 Keypoints)

Model	Acc(%)	Params(M)	FLOPs(G)	latency(ms)
ST-GCN	91.32	2.07	0.795	1.76
2s-AGCN	92.47	3.45	1.305	3.41
CTR-GCN	**93.11**	1.42	0.595	5.53
RepCA-GCN	92.58	**1.19**	**0.413**	**1.36**

Although slightly lower than CTR-GCN (93.11%), RepCA-GCN demonstrates a vastly superior speed-accuracy trade-off. By leveraging structural re-parameterization, it requires only 0.413 GFLOPs and achieves an ultra-low inference latency of 1.36ms, making it a well-suited choice for resource-constrained edge deployment.

Table 4. Accuracy (%) comparison across different models and keypoint configurations

Model	13 Points	14 Points	17 Points
ST-GCN	92.13	90.16	91.32
2s-AGCN	90.52	95.08	92.47
CTR-GCN	**92.31**	94.54	**93.11**
RepCA-GCN	92.22	**95.35**	92.58

Furthermore, we investigated the influence of skeletal representations on model performance (Table 4). Interestingly, RepCA-GCN achieves its peak accuracy of 95.35% when utilizing the 14-keypoint scheme, significantly surpassing the standard 17-keypoint COCO format (92.58%). This confirms that standard pose layouts may not be optimal for construction action recognition. The 17-keypoint format contains dense, highly localized facial keypoints that contribute little to macroscopic body actions but introduce significant noise when partially occluded by hardhats. The streamlined 14-keypoint scheme effectively filters out this redundant facial noise while preserving essential structural integrity, thereby maximizing the efficacy of our Confidence Gating mechanism and yielding the best recognition performance.

4.3 Ablation Study and Efficiency Analysis

As shown in Table 5, activating the structural re-parameterization deployment mode (ON) losslessly reduces parameters and FLOPs by ~40%, lowering latency to 1.36 ms without accuracy degradation.

Table 5. Efficiency comparison of RepCA-GCN before and after structural re-parameterization

Rep deploy	Params(M)	FLOPs(G)	latency(ms)
OFF	2.01	0.697	1.80
ON	**1.19**	**0.413**	**1.36**

This efficiency gain is mathematically guaranteed by the additivity of linear transformations. During training, the multi-branch temporal block outputs $Y = \sum_{i\in S} \mathrm{BN}_i\,(\mathrm{Conv}_i(X))$, where $S = \{9\times 1, 5\times 1, 3\times 1, 1\times 1, \text{identity}\}$. For inference, each Conv-BN sequence is linearly fused into a single weight W_i' and bias b_i'. By zero-padding smaller kernels and the identity matrix (formulated as a 1×1 Dirac delta) to match the maximal 9×1 kernel (denoted by a padding operator $\mathcal{P}$), all branches equivalently collapse into a single generic convolution:

$$Y = \left(\sum_{i\in S} \mathcal{P}(W_i')\right) X + \sum_{i\in S} b_i' = W_{\text{fused}} X + b_{\text{fused}} \quad (1)$$

This structural decoupling ensures high training capacity while yielding an ultra-fast, single-path network for edge deployment.

5 Application and Evaluation

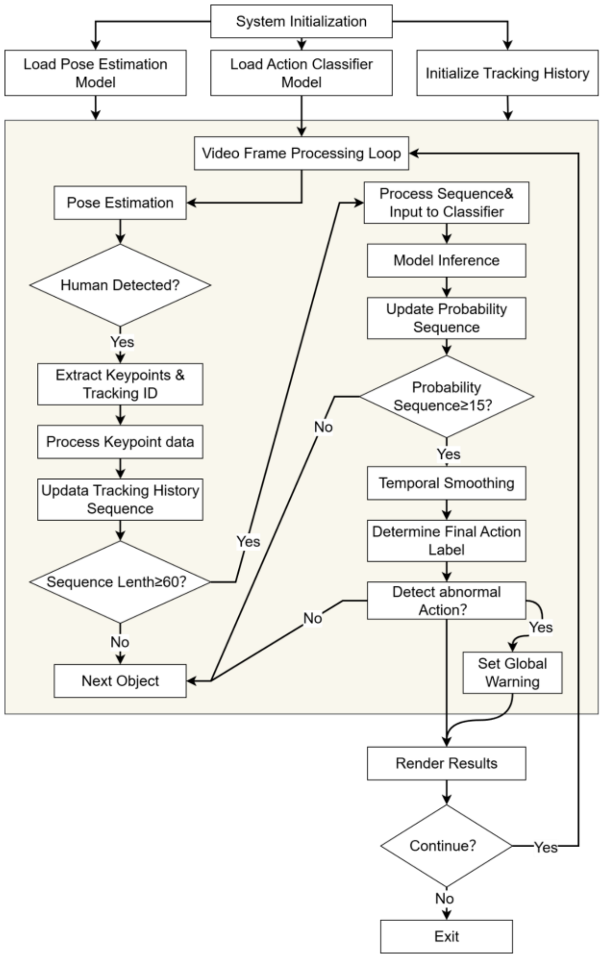

Figure 9. Flowchart of the proposed worker action recognition system for construction sites

To verify its practical utility, we developed a real-time monitoring system (Figure 9) that integrates YOLOv11-Pose for multi-object tracking with RepCA-GCN for action classification. The pipeline assigns unique IDs to detected individuals to extract and preprocess continuous skeletal keypoint streams.

5.1 Real-time Monitoring System Design and Visualization

For robust monitoring, a dual sliding-window mechanism—a 60-frame Temporal Buffer for context and a 15-frame Probabilistic Smoothing Window ($\alpha = 0.96$) —is implemented to filter transient noise and classification jitter. As visualized in Figure 10, the system accurately identifies diverse actions (e.g., "falls") and provides real-time visual feedback for immediate safety alerts.

5.2 System Performance Evaluation

Table 6 details the end-to-end throughput in multi-person scenarios, achieving 60.5 FPS on an RTX 4090 and 22.6 FPS on an RTX 4060. These baseline PyTorch metrics already satisfy real-time constraints, with further acceleration possible via TensorRT, proving RepCA-GCN an efficient and scalable solution for safety supervision.

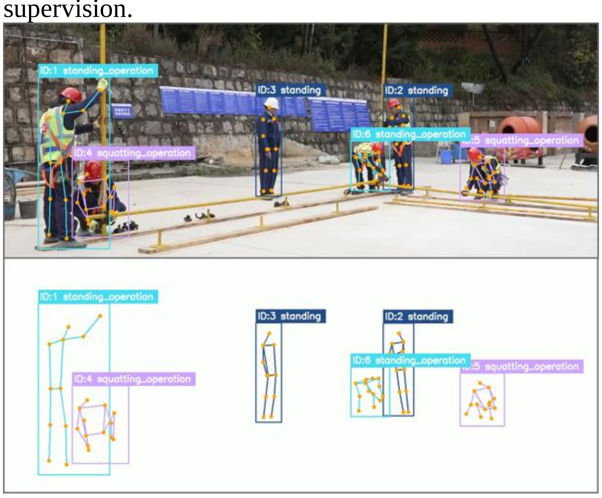

Figure 10. Visualization of Recognition Results

Table 6. End-to-End System Throughput (FPS) on Different Hardware Platforms

Model	Server (RTX 4090)			PC (RTX 4060)		
	Single-person	Multi-person	Avg	Single-person	Multi-person	Avg
ST-GCN	65.1	41.5	53.3	**24.9**	18.9	21.9
2s-AGCN	52.1	27.7	39.9	23.9	15.2	19.6
CTR-GCN	47.6	24.0	35.8	18.9	10.9	14.9
RepCA-GCN	**70.6**	**50.3**	**60.5**	24.1	**21.0**	**22.6**

6 Conclusion and Limitations

This study proposes RepCA-GCN, a lightweight framework for construction worker action recognition. By integrating Structural Re-parameterization and reliability-aware feature modulation, the model addresses computational latency and noisy keypoints caused by occlusions. We constructed the SWAR dataset with extensive augmentation to facilitate robust training. Results show RepCA-GCN achieves state-of-the-art accuracy (95.35%) with reduced computational costs (0.413 GFLOPs) and low latency (1.36 ms). The monitoring system demonstrates robust real-time performance for intelligent safety supervision. However, skeleton-based methods completely discard background information, making them unable to recognize environment-interactive actions.

Acknowledgments

The authors would like to acknowledge the support by National Science Foundation of China (52308314), the support by Guangdong Basic and Applied Basic Research Foundation (2023A1515030169) and the support by Guangzhou Science and Technology Project (2025A04J5216).

References

[1] B. Choi, S. Ahn, S. Lee, Construction Workers' Group Norms and Personal Standards Regarding Safety Behavior: Social Identity Theory Perspective, J. Manage. Eng. 33 (2017) 04017001. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000511.

[2] M.A. Nabi, I.H. El-adaway, C. Dagli, A System Dynamics Model for Construction Safety Behavior, Procedia Computer Science 168 (2020) 249–256. https://doi.org/10.1016/j.procs.2020.02.254.

[3] S. Chi, C.H. Caldas, Automated Object Identification Using Optical Video Cameras on Construction Sites, Comput.-Aided Civil Infrastruct. Eng. 26 (2011) 368–380. https://doi.org/10.1111/j.1467-8667.2010.00690.x.

[4] A. Kelm, L. Laussat, A. Meins-Becker, D. Platz, M.J. Khazaee, A.M. Costin, M. Helmus, J. Teizer,

Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites, Autom. Constr. 36 (2013) 38–52. https://doi.org/10.1016/j.autcon.2013.08.009.

[5] I. Awolusi, E. Marks, M. Hallowell, Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices, Autom. Constr. 85 (2018) 96–106. https://doi.org/10.1016/j.autcon.2017.10.010.

[6] S. Subedi, N. Pradhananga, Sensor-based computational approach to preventing back injuries in construction workers, Autom. Constr. 131 (2021) 103920. https://doi.org/10.1016/j.autcon.2021.103920.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 172–186. https://doi.org/10.1109/TPAMI.2019.2929257.

[8] D. Maji, S. Nagori, M. Mathew, D. Poddar, YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2022, IEEE, New York, 2022: pp. 2636–2645. https://doi.org/10.1109/CVPRW56347.2022.00297.

[9] H. Wu, Y. Han, M. Zhang, B.D. Abebe, M.B. Legesse, R. Jin, Identifying Unsafe Behavior of Construction Workers: A Dynamic Approach Combining Skeleton Information and Spatiotemporal Features, J. Constr. Eng. Manage. 149 (2023) 04023115. https://doi.org/10.1061/JCEMD4.COENG-13616.

[10] X. Zhou, S. Li, J. Liu, Z. Wu, Y.F. Chen, Construction Activity Analysis of Workers Based on Human Posture Estimation Information, Engineering 33 (2024) 225–236. https://doi.org/10.1016/j.eng.2023.10.004.

[11] S. Yan, Y. Xiong, D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence / Thirtieth Innovative Applications of Artificial Intelligence Conference / Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, Assoc Advancement Artificial Intelligence, Palo Alto, 2018: pp. 7444–7452.

[12] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), IEEE, New York, 2019: pp. 12018–12027. https://doi.org/10.1109/CVPR.2019.01230.

[13] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021), IEEE, New York, 2021: pp. 13339–13348. https://doi.org/10.1109/ICCV48922.2021.01311.

[14] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, RepVGG: Making VGG-style ConvNets Great Again, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, IEEE Computer Soc, Los Alamitos, 2021: pp. 13728–13737. https://doi.org/10.1109/CVPR46437.2021.01352.

[15] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New York, 2016: pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.

[16] N.D. Nath, A.H. Behzadan, S.G. Paal, Deep learning for site safety: Real-time detection of personal protective equipment, Autom. Constr. 112 (2020) 103085. https://doi.org/10.1016/j.autcon.2020.103085.

[17] L. Lopez, J. Suarez-Ramirez, M. Aleman-Flores, N. Monzon, Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation, Autom. Constr. 176 (2025) 106231. https://doi.org/10.1016/j.autcon.2025.106231.

[18] J. Carreira, A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, in: 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), IEEE, New York, 2017: pp. 4724–4733. https://doi.org/10.1109/CVPR.2017.502.

[19] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New York, 2016: pp. 1933–1941. https://doi.org/10.1109/CVPR.2016.213.

[20] C. Feichtenhofer, H. Fan, J. Malik, K. He, SlowFast Networks for Video Recognition, in: 2019 IEEE/CCF International Conference on Computer Vision (ICCV 2019), IEEE, New York, 2019: pp. 6201–6210. https://doi.org/10.1109/ICCV.2019.00630.