A Multimodal Late-Fusion Framework for Reliable Crane Hand-Signal Recognition in Low-Light Environments

Zhihao Wei[1,2], Moein Maleki[2,3], Ghulam Muhammad Ali[3], and Xinming Li[1,*]

[1]School of Mechanical Engineering, University of Alberta, Canada
[2]School of Electrical and Computer Engineering, University of Alberta, Canada
[3]SensiImage Technologies Ltd. Canada
wzhihao@ualberta.ca, mmaleki@ualberta.ca, gmail@sensiimage.com, xinming.li@ualberta.ca

Abstract –

This study presents a multimodal crane hand-signal recognition framework that integrates helmet-mounted binocular vision with wearable sensing. Dual-view visual inputs are processed using Vision Transformer models and augmented by a few-shot visual prior based on Contrastive Language–Image Pretraining (CLIP) to improve generalization under unseen environments. In parallel, a hybrid CNN–LSTM architecture is employed for sensor-based recognition to capture temporal hand motion patterns. During inference, left and right visual predictions are first fused at the visual level, followed by cross-modal fusion with sensor outputs using an optimal transport formulation with a task-informed cost matrix. Experimental results show that the proposed fusion strategy maintains 92.93% accuracy under unseen background conditions and 92.73% accuracy under unseen background with low-light conditions, demonstrating strong robustness and reliability for crane hand-signal recognition in realistic construction scenarios.

Keywords –

Crane operations; Few-shot learning; Multimodal fusion; Vision Transformer; Wearable sensing; Safety in construction.

1 Introduction

Crane hand signals serve as a critical communication medium between signalers and crane operators, enabling safe and efficient coordination in construction environments. Accurate recognition of these standardized gestures is essential for preventing misoperation and ensuring on-site safety. However, manual interpretation of hand signals remains error-prone due to variations in gesture execution and operator experience, which has motivated increasing interest in automated crane hand-signal recognition systems.

Existing gesture recognition approaches can be broadly categorized into vision-based and wearable sensor–based methods. Vision-based systems leverage camera inputs and deep learning models to recognize gestures from visual appearance, offering intuitive and non-intrusive sensing. Wearable sensor–based approaches directly capture hand motion and finger articulation through inertial and flex sensing, providing reliable measurements of gesture dynamics. Each modality offers complementary information, yet their performance may degrade when deployed independently under distribution shifts between training and real-world operation. In particular, sensor-based methods, while generally robust, can be sensitive to motion disturbances, rapid gesture transitions, and interference from whole-body movements, which introduce ambiguity in temporal signals.

To address these challenges, this work proposes a multimodal crane hand-signal recognition framework that integrates helmet-mounted binocular vision with wearable sensing. Dual-view visual predictions are enhanced by a CLIP-based few-shot visual prior to improve generalization under unseen conditions, and are subsequently fused with sensor-based predictions using an optimal transport formulation with a task-informed cost matrix. By explicitly incorporating gesture semantics into the fusion process, the proposed system enables robust and reliable recognition performance under unseen background and low-light scenarios.

2 Related works

2.1 Vision-based HGR

Vision-based methods represent a major research direction for hand gesture recognition. Common optical sensing modalities include RGB cameras [1], [2], [3], [4], depth and RGB-D cameras [5], [6], [7], infrared sensors, and stereo vision systems [1], [8]. By combining visual observations with machine learning or deep neural networks, these approaches have demonstrated strong

performance in recognizing hand gestures under controlled conditions.

With the advancement of deep learning, convolutional neural networks (CNNs) have become the dominant backbone for image-based gesture recognition [6]. Recurrent architectures such as Long Short-Term Memory (LSTM) networks have further enabled temporal modeling in video-based tasks [1], while 3D-CNNs integrate spatial and temporal learning within a unified framework, achieving high accuracy for dynamic gesture recognition [5], [9]. More recently, Vision Transformers (ViT) [4], [9] have attracted increasing attention due to their ability to model global context via self-attention, leading to improved robustness against background variation and viewpoint changes.

In construction-related applications, vision-based systems have been explored for crane hand-signal recognition. Mansoor et al. employed YOLO-based multi-view video analysis [1] to recognize standardized crane gestures, while Wang et al. [5] achieved over 93% accuracy using a 3D-CNN trained on RGB-D hand-signal datasets. However, most existing vision-based datasets adopt second-person viewpoints with fixed external cameras, which limits their applicability in outdoor and dynamic construction environments. First-person vision datasets, such as EgoGesture [3], provide an alternative by enabling wearable or body-mounted data collection, offering better suitability for mobile and real-world deployment.

2.2 Sensor-based HGR

Wearable sensor-based gesture recognition has received extensive attention, particularly for scenarios where visual sensing is unreliable. A wide range of sensor types have been employed, including inertial measurement units (IMUs) [9], [10], electromyography (EMG) [11], [12], [13], [14], flex sensors [10], optical sensors [15], radar [16], [17], [18], acoustic sensors [12], and photoplethysmography (PPG) [19]. Among these, IMUs are well suited for capturing dynamic gestures through acceleration and angular velocity measurements and are widely integrated into wearable devices. Mansoor et al. [10] combined IMU and flex sensors within a glove to classify crane hand signals, achieving 94.22% accuracy using a CNN–LSTM architecture. Similarly, Wang et al. [20] designed a ring-shaped accelerometer for fine-grained hand and finger motion tracking, demonstrating the adaptability of compact IMU-based sensing.

EMG-based approaches measure muscle activation signals and excel at capturing subtle finger motions and postural changes. Neacsu et al. [14] reported 99.78% accuracy across seven gesture classes using time-domain EMG features and neural networks. Optical tendon-sensing techniques and flex sensors have also been widely applied to detect finger bending and deformation states.

Despite their effectiveness, single-sensor systems face inherent limitations. IMUs are susceptible to environmental drift, EMG signals are sensitive to electrode placement and skin conditions, optical and radar sensors are affected by lighting and interference, and flex sensors struggle to capture dynamic transitions.

To address these challenges, multimodal fusion has been extensively investigated. Prior studies have shown that integrating complementary sensors, such as EMG and ultrasound [12] or IMU and flex sensors [10], significantly improves recognition accuracy and robustness. Large-scale multimodal datasets and attention-based fusion mechanisms [12], [21], [22] further demonstrate the advantages of deep learning–based multimodal gesture recognition in complex and variable environments, including industrial and construction scenarios.

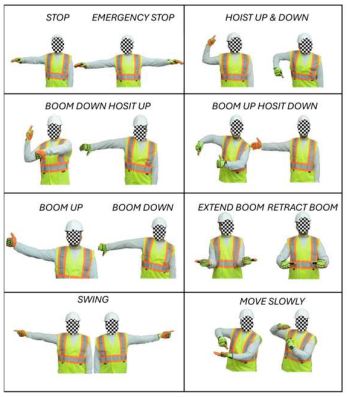


Figure 1. Hand signals for crane operations

3 System Overview

This section outlines the proposed multimodal crane hand-signal recognition system. The framework combines helmet-mounted binocular vision and wearable sensing, with visual and sensor data processed by dedicated recognition models. As shown in Figure.2, dual-view visual predictions are first fused and augmented by a CLIP-based few-shot visual prior, followed by cross-modal fusion with sensor outputs using an optimal transport formulation with a task-informed cost matrix.

3.1 Hardware Design

The proposed system employs a multimodal hardware setup that integrates a dual-view vision module with a wearable sensor device to capture complementary visual and motion information during crane hand-signal

execution.

As shown in Figure 3, a pair of cameras is mounted on the helmet to provide first-person-view left and right observations of the signaler's hands. Each camera has a diagonal field of view of 120° and a horizontal field of view of 102°. The dual-camera setup forms an effective horizontal field of view of approximately 186°, ensuring full coverage of crane hand-signal motions. The first-person, helmet-mounted binocular configuration reduces gesture occlusion in real construction scenarios and maintains a fixed viewing geometry relative to the hands, which simplifies visual recognition and improves robustness.

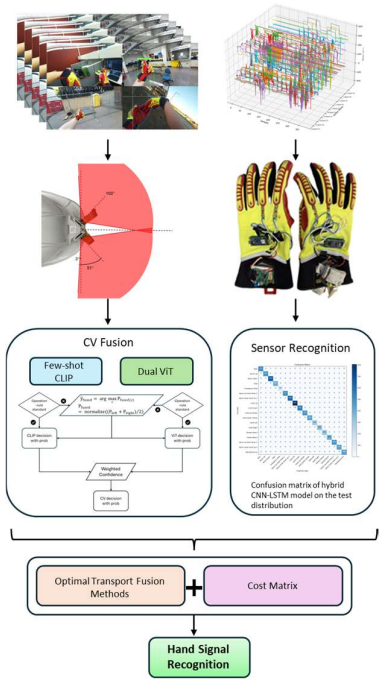
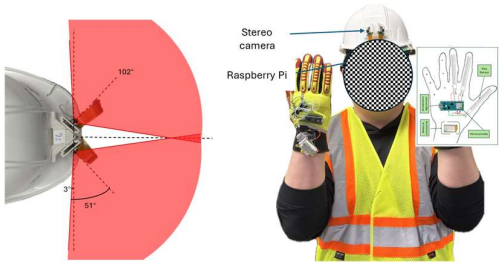


Figure 2. Multimodal crane signal recognition workflow.

In parallel, a wearable glove equipped with inertial measurement units (IMUs) and flex sensors is used to record hand motion and finger articulation signals. The IMU is mounted on the dorsum of the hand to capture overall hand movement, while flex sensors are placed on the thumb, index finger, and middle finger to measure finger bending during gesture execution.

A Raspberry Pi is used as the central edge node for system coordination and data relay. Visual data captured by the helmet-mounted cameras are stored locally on the Raspberry Pi, while the wearable glove transmits sensor data to the same node via Bluetooth from an embedded microcontroller. The Raspberry Pi synchronizes and aggregates multimodal data streams and subsequently uploads the collected data to a cloud server for offline analysis and model evaluation.



(a) (b)

Figure 3. Multimodal hardware setup with helmet-mounted stereo cameras and wearable sensing gloves.

3.2 Data Structure and Preprocess

The visual and sensor modalities are processed independently during training. Strict temporal synchronization is applied only in the test dataset, where each incoming sensor data sample triggers a simultaneous capture from the left and right cameras, forming a synchronized multimodal test instance. Data were collected from 13 volunteers (3 females and 10 males) under approval of the University of Alberta Research Ethics Board (Study ID: Pro00154164). For training, data were acquired under varying conditions depending on volunteer availability, including visual-only samples, sensor-only samples, and partially synchronized samples. Accordingly, the visual and sensor models are trained separately, and multimodal alignment is performed only at the testing stage for fusion evaluation.

For the vision modality, the left and right camera streams are treated as two independent image datasets during training. The right-view dataset contains 7,519 images, while the left-view dataset contains 7,748 images; both datasets are split into 70% for training, 20% for validation, and 10% for testing. Visual data were collected across multiple environments, including office spaces, campus-scale mock construction settings, and a real construction site. No explicit temporal alignment between the two camera views is required during training, as fusion is performed at the inference stage.

For the sensor modality, time-series signals from the IMU and flex sensors are segmented into fixed-length samples using a sliding-window strategy to preserve temporal dynamics [23]. Each glove records nine sensor channels per time step, including gyroscope readings(G_x, G_y, G_z), accelerometer readings(A_x, A_y, A_z), and flex sensor measurements from the thumb, index finger, and middle finger$(flex_1, flex_2, flex_3)$. Sensor samples are constructed using a window size of five

consecutive time steps at a sampling rate of 30 Hz. Data from the left and right gloves are combined, resulting in a 90-dimensional feature vector per sample. In total, 41820 sensor samples are obtained using a semi-automatic annotation procedure. Specifically, a custom software interface developed in Visual Studio is used to synchronize data acquisition between the wearable device and the host computer. During data collection, participants are instructed to continuously perform different gestures rather than holding fixed poses. At the moment a gesture is initiated, the corresponding label is manually triggered through the interface and appended to the incoming sensor data stream. This process enables the recording of complete gesture transitions, including intermediate motion states, allowing the sequences to capture both global hand motion and finger articulation.

3.3 Visual Recognition

The visual recognition module is built upon Vision Transformer (ViT) models to classify crane hand signals from helmet-mounted camera images. The left-view and right-view camera streams are treated as two independent visual inputs during training, and separate ViT models are trained for each camera. This design accounts for viewpoint-specific appearance differences and avoids imposing strict synchronization constraints between the two camera views at the training stage.

Each ViT model adopts a ViT-Base architecture with a patch size of 16 and an input resolution of 384 × 384. Raw RGB images are resized to 384 × 384 and normalized using a mean and standard deviation of 0.5 for all channels before being fed into the network. The classification head outputs camera-specific gesture classes, which are subsequently mapped to a unified gesture label space during inference.

Model training is conducted using standard supervised learning procedures. Specific training configurations, including the loss function, optimizer, learning rate, batch size, and training epochs, are summarized in Table 1. By decoupling visual model training from cross-view fusion, the proposed design maintains a simple and scalable training pipeline while enabling robust late fusion under real-world deployment conditions where partial visibility or asymmetric hand occlusion may occur.

Table 1. Configuration of the Vision Transformer (ViT) models for visual hand-signal recognition

Configuration	
Backbone architecture	ViT-Base
Batch size	256
Number of output class	18
Loss Function	Cross-Entropy Loss
Optimizer	Adam
Learning rate	0.0001
Training epoch	25
Early stop	3

3.4 Few-shot CLIP-based Visual Prior

To complement the supervised ViT models, a CLIP-based few-shot recognition module is introduced to enhance generalization under limited or distribution-shifted visual data. Instead of conventional fine-tuning, this module adopts a prototype-based inference scheme built upon pretrained vision–language representations.

For each camera view, a small reference set is constructed, where 10 labeled images per gesture class (10-shot) are collected and organized by class. These images are encoded using a pretrained CLIP visual encoder, and the resulting feature vectors are L2-normalized. A class prototype is then obtained by averaging the normalized embeddings of the corresponding 10 samples, followed by normalization. This process yields one prototype per gesture class for each camera view. The configuration of the CLIP few-shot recognition module is summarized in Table 2.

Table 2: Configuration of the CLIP-based few-shot recognition module.

Component	Setting
CLIP backbone	ViT-L/14
Input image size	336*336
Shots per class	10
Prototype construction	Mean of normalized embeddings
Similarity metric	Cosine similarity
SoftMax temperature	10

During inference, a test image is encoded by the same CLIP encoder and classified by measuring cosine similarity to all class prototypes. The similarity scores are scaled by a temperature parameter and converted into a probability distribution using a SoftMax function. The resulting CLIP-based prediction is incorporated as an auxiliary enhancement branch and combined with the ViT-based visual output at the probability level in the subsequent fusion stage.

3.5 Sensor-Based recognition

For sensor-based gesture recognition, a hybrid CNN–LSTM architecture [23] is adopted to jointly model local feature patterns and temporal dependencies in hand motion signals. As illustrated in Figure 4, the network employs a dual-head one-dimensional convolutional structure, in which two parallel Conv1D branches extract complementary feature representations from the same input sequence. Each branch consists of a convolutional

layer followed by batch normalization and max pooling, enabling multi-scale feature extraction while improving training stability and robustness.

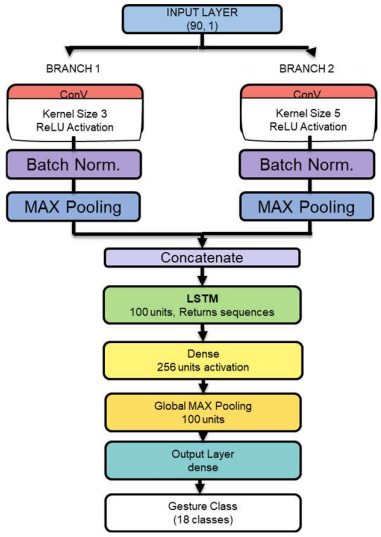


Figure 4. Architecture of proposed hybrid CNN-LSTM Model

The features generated by the parallel CNN branches are subsequently integrated and fed into an LSTM layer to capture temporal dependencies across the gesture execution process. To further emphasize salient temporal patterns, a Global Max Pooling layer is applied before the fully connected layers. The final classification is performed using a SoftMax layer to produce probability distributions over 18 predefined crane hand-signal classes. The hyperparameters used for training the CNN–LSTM model are summarized in Table 3

Table 3. Hyperparameters of the sensor-based hand gesture recognition model.

Component	Setting
Optimizer	Adam
Loss function	Categorical cross-entropy
Learning rate	0.001
Early stop patience	10
Batch size	64
Max epoch	300
Classes	18

4 Fusion Strategy

This section presents the fusion strategy used to combine dual-view visual predictions and wearable sensor outputs at the probability level for robust crane hand-signal recognition.

4.1 Visual-level Fusion

At the visual level, gesture recognition is performed independently for the left and right helmet-mounted cameras, with each view producing a discrete class prediction and a corresponding class probability distribution. The binocular visual outputs are then fused using a rule-driven strategy based on crane hand-signal operation standards. When the predicted gesture pair conforms to a valid operational combination, a semantically consistent fused gesture, including composite gestures, is generated directly. If the predicted combination does not match any predefined operation rule, the fusion falls back to probability aggregation, where the final visual prediction is determined by averaging the class probabilities from both views and selecting the most probable class.

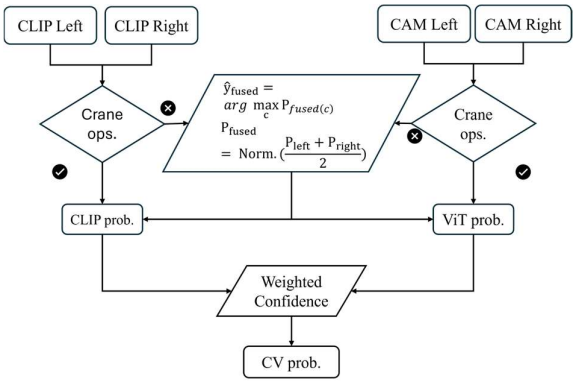


Figure 5. Visual fusion workflow.

In parallel, a few-shot CLIP-based visual branch performs prototype-based classification for each camera view, and its left–right outputs are fused using the same operation-rule-driven mechanism. The fused CLIP probabilities are subsequently integrated with the ViT-based visual probabilities at the probability level via weighted averaging (0.6 for ViT and 0.4 for CLIP), producing a unified visual representation that is forwarded to the multimodal fusion stage. The overall visual recognition and fusion workflow is illustrated in Fig. 5.

4.2 Cross-Modal Fusion via Optimal Transport

To integrate visual and wearable sensor predictions, the final multimodal fusion is formulated using an Optimal Transport (OT) framework. Let p^v and p^s denote the class probability distributions produced by the visual recognition module and the

sensor-based model, respectively, where C is the number of gesture classes.

Rather than directly averaging probabilities, the fusion is defined as the computation of a Wasserstein barycentre [24] that minimizes the transport cost between the fused distribution and each modality:

$$P^* = arg \min_{P \in \Delta^C} \left(\sum_{m \in \{v,s\}} w_m W_{CM}(p, p^m) \right) \quad (1)$$

where Δ^C denotes the probability simplex, w_v and w_s are modality weights, and $W_{\mathrm{CM}}(\cdot,\cdot)$ is the entropy-regularized Wasserstein distance defined over a predefined cost matrix C

The cost matrix C_M is constructed using a rule-based semantic decomposition of crane hand signals. Each gesture is represented along three dimensions: primary motion, hand motion pattern, and direction. The transportation cost is reduced by 0.2 for each shared dimension. For dual-hand gestures, shared sub-motions are treated as part of the motion dimension and are not considered as an additional reduction.

For safety-critical considerations, the transportation cost between *Stop*/*Emergency stop* and all other gestures is strictly set to 1, as misclassification of these signals is unacceptable in crane operations. The pair (*Stop*, *Emergency stop*) is then assigned a reduced cost of 0.3 to reflect their nearly identical operational meaning while preserving their distinction. Similarly, gesture variants that convey the same operational intent (e.g., *Boom up hoist down 1/2*) are assigned a cost of 0.2.

The resulting barycentre P^* represents a consensus distribution that balances visual and sensor evidence while respecting domain-specific gesture semantics. The final prediction is obtained by selecting the class with the maximum probability in P^*.

5 Result

5.1 Experiment setup

During testing, sensor and visual data were strictly synchronized at the sample level. Each fixed-length sensor window triggered a simultaneous capture from the dual helmet-mounted cameras, and multimodal inference was performed on the aligned sensor–image pairs.

To evaluate generalization, all test images were collected in previously unseen environments with entirely new backgrounds. In addition, 50% of the test images were subjected to low-light processing to simulate challenging visual conditions commonly encountered on construction sites.

5.2 Result and discussion

Table 4 summarized the recognition performance of different models under test distribution, unseen background, and unseen background with low-light conditions.

Table 4: Overall accuracy comparison under different conditions

Model	Test distribution	New background	New background with low light
ViT (L)		91.22%	90.69%
ViT (R)		91.02%	81.98%
ViT (Dual cameras)	--	82.20%	66.87%
CLIP	--	72.73%	68.73%
CV Fusion	--	84.73%	68.73%
Sensor	98.94%	92.41%	92.41%
Multimodal Fusion	--	92.93%	92.73%
Sensor Degradation	98.94%	87.81%	87.81%
Fusion (sensor degraded)	--	89.55%	--

The proposed multimodal fusion framework effectively withstands interference from challenging conditions, maintaining 92.93% accuracy in unseen background scenarios and 92.73% under combined unseen background and low-light conditions. These results highlight the effectiveness of integrating complementary visual and sensor modalities for robust crane hand-signal recognition in maintaining reliable performance when visual quality degrades.

To further evaluate the reliability of proposed fusion framework under sensor-degradation conditions, we conduct an additional experiment using a rapid gesture transition scenario, where frequent motion transitions are intentionally introduced by requiring participants to perform each gesture immediately followed by a *Stop* gesture, completing all gestures within one minute. This scenario reflects realistic operational disturbances where sensor signals are less stable.

Compared with prior RGB-D-based approaches (e.g., 93.3% reported in [5] using a third-person setup), the proposed method achieves comparable performance with a more lightweight and occlusion-robust design, benefiting from low-dimensional sensor inputs and a first-person perspective with reduced hand occlusion.

Purely vision-based methods exhibit noticeable performance degradation when evaluated on new environments and under low-light constraints. Notably, the fusion of CLIP and ViT within the visual pipeline improves recognition accuracy by 2.53% compared to ViT alone, demonstrating that large language models can

enhance visual generalization when combined with task-specific visual models.

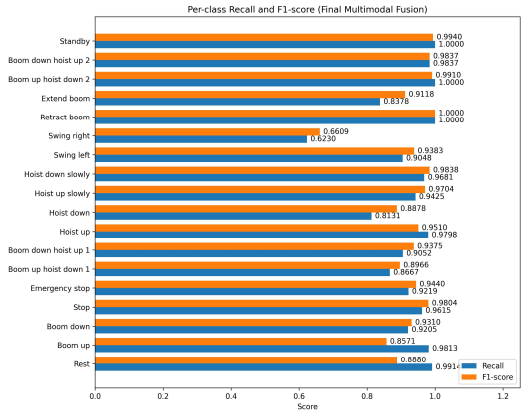


Figure 6. Per-class recall and F1-score of the multimodal fusion model under unseen background and low-light conditions

Figure 6 presents the per-class recall and F1-score of the final multimodal fusion model under unseen background and low-light conditions. High and balanced performance is observed across all 18-crane hand-signal classes. The fusion strategy goes beyond confidence averaging by incorporating an optimal transport–based formulation with a task-informed cost matrix, which accounts for gesture similarities and asymmetric misclassification risks, enabling more reliable decision-making when visual and sensor predictions disagree.

6 Conclusion and Future work

This paper presents a robust multimodal crane hand-signal recognition system that integrates dual-view visual perception and wearable sensor sensing through a principled late-fusion strategy. By combining Vision Transformer–based visual recognition, CLIP-based few-shot visual priors, and a CNN–LSTM sensor model, the proposed framework effectively addresses the challenges posed by new environment backgrounds and low illumination. The experimental results demonstrate that the multimodal fusion approach consistently outperforms single-modality methods, maintaining high recognition accuracy under challenging conditions.

A key contribution of this work lies in introducing a domain-informed multimodal fusion paradigm for crane hand-signal recognition. By embedding standardized operational knowledge into the fusion process, the proposed framework goes beyond generic data-driven aggregation and enables more structured and interpretable decision-making in complex construction environments.

Future work will focus on further improving system robustness and practical deployment. On the visual side, more efficient binocular vision strategies, such as joint training or feature-level fusion, will be explored to better exploit complementary viewpoints while controlling computational cost. On the sensor side, challenges related to user-specific gesture styles, long-term IMU drift, and interference from whole-body motion will be investigated. Additional adverse conditions, including motion blur and rapid gesture transitions, will also be considered. Through continued advances in visual modeling, sensor reliability, fusion strategies, and system-level extensibility (e.g., environmental perception and safety monitoring such as hazard detection and warning for workers entering dangerous zones), future work aims to build a more reliable crane hand-signal recognition system for real construction environments.

Acknowledgement

This work was supported by China Scholarship Council (CSC), the authors sincerely thank CSC for their financial support.

References

[1] A. Mansoor, S. Liu, G. M. Ali, A. Bouferguene, and M. Al-Hussein, “A Deep-Learning Classification Framework for Reducing Communication Errors in Dynamic Hand Signaling for Crane Operation,” *J. Constr. Eng. Manag.*, vol. 149, no. 2, p. 04022167, Feb. 2023, doi: 10.1061/JCEMD4.COENG-12811.

[2] J. Shin *et al.*, “Korean Sign Language Recognition Using Transformer-Based Deep Neural Network,” *Appl. Sci.*, vol. 13, no. 5, p. 3029, Jan. 2023, doi: 10.3390/app13053029.

[3] Y. Zhang, C. Cao, J. Cheng, and H. Lu, “EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition,” *IEEE Trans. Multimed.*, vol. 20, no. 5, pp. 1038–1050, May 2018, doi: 10.1109/TMM.2018.2808769.

[4] C. K. Tan, K. M. Lim, R. K. Y. Chang, C. P. Lee, and A. Alqahtani, “HGR-ViT: Hand Gesture Recognition with Vision Transformer,” *Sensors*, vol. 23, no. 12, p. 5555, Jan. 2023, doi: 10.3390/s23125555.

[5] X. Wang and Z. Zhu, “Vision-based hand signal recognition in construction: A feasibility study,” *Autom. Constr.*, vol. 125, p. 103625, May 2021, doi: 10.1016/j.autcon.2021.103625.

[6] M. Kurmanji and F. Ghaderi, “Hand Gesture Recognition from RGB-D Data using 2D and 3D Convolutional Neural Networks: a comparative

study," *J. AI Data Min.*, vol. 8, no. 2, pp. 177–188, Apr. 2020, doi: 10.22044/jadm.2019.7903.1929.

[7] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug. 2012, pp. 1975–1979.

[8] C. Hubert, N. Odic, M. Noel, S. Gharib, S. H. H. Zargarbashi, and L. S\' eoud, "MuViH: Multi-View Hand gesture dataset and recognition pipeline for human–robot interaction in a collaborative robotic finishing platform," *Robot. Comput.-Integr. Manuf.*, vol. 94, p. 102957, Aug. 2025, doi: 10.1016/j.rcim.2025.102957.

[9] H. Zhang, K. Yang, G. Cao, and C. Xia, "ViT-LLMR: Vision Transformer-based lower limb motion recognition from fusion signals of MMG and IMU," *Biomed. Signal Process. Control*, vol. 82, p. 104508, Apr. 2023, doi: 10.1016/j.bspc.2022.104508.

[10] A. Mansoor, S. Liu, A. Bouferguene, and M. Al-Hussein, "Crane Signalman Hand-Signal Classification Framework Using Sensor-Based Smart Construction Glove and Machine-Learning Algorithms," *J. Constr. Eng. Manag.*, vol. 150, no. 8, p. 04024094, Aug. 2024, doi: 10.1061/JCEMD4.COENG-14458.

[11] Y. Liu, X. Li, L. Yang, G. Bian, and H. Yu, "A CNN-Transformer Hybrid Recognition Approach for sEMG-Based Dynamic Gesture Prediction," *IEEE Trans. Instrum. Meas.*, vol. 72, p. 2514816, 2023, doi: 10.1109/TIM.2023.3273651.

[12] S. Wei, Y. Zhang, and H. Liu, "A Multimodal Multilevel Converged Attention Network for Hand Gesture Recognition With Hybrid sEMG and A-Mode Ultrasound Sensing," *IEEE Trans. Cybern.*, vol. 53, no. 12, pp. 7723–7734, Dec. 2023, doi: 10.1109/TCYB.2022.3204343.

[13] W. Chen, L. Feng, J. Lu, and B. Wu, "An Extended Spatial Transformer Convolutional Neural Network for Gesture Recognition and Self-Calibration Based on Sparse sEMG Electrodes," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 6, pp. 1204–1215, Dec. 2022, doi: 10.1109/TBCAS.2022.3222196.

[14] A. A. Neacsu, G. Cioroiu, A. Radoi, and C. Burileanu, "Automatic EMG-based Hand Gesture Recognition System using Time-Domain Descriptors and Fully-Connected Neural Networks," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2019, pp. 232–235. doi: 10.1109/TSP.2019.8768831.

[15] S. Ortega-Avila, B. Rakova, S. Sadi, and P. Mistry, "Non-invasive optical detection of hand gestures," in *Proceedings of the 6th Augmented Human International Conference*, Singapore Singapore: ACM, Mar. 2015, pp. 179–180. doi: 10.1145/2735711.2735801.

[16] L. Zheng *et al.*, "Dynamic Hand Gesture Recognition in In-Vehicle Environment Based on FMCW Radar and Transformer," *Sensors*, vol. 21, no. 19, p. 6368, Jan. 2021, doi: 10.3390/s21196368.

[17] C. Wang, X. Zhao, and Z. Li, "DCS-CTN: Subtle Gesture Recognition Based on TD-CNN-Transformer via Millimeter-Wave Radar," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 17680–17693, Oct. 2023, doi: 10.1109/JIOT.2023.3280227.

[18] B. Jin, X. Ma, Z. Zhang, Z. Lian, and B. Wang, "Interference-Robust Millimeter-Wave Radar-Based Dynamic Hand Gesture Recognition Using 2-D CNN-Transformer Networks," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 2741–2752, Jan. 2024, doi: 10.1109/JIOT.2023.3293092.

[19] T. Zhao, J. Liu, Y. Wang, H. Liu, and Y. Chen, "Towards Low-Cost Sign Language Gesture Recognition Leveraging Wearables," *IEEE Trans. Mob. Comput.*, vol. 20, no. 4, pp. 1685–1701, Apr. 2021, doi: 10.1109/TMC.2019.2962760.

[20] X. Wang, D. Veeramani, and Z. Zhu, "Wearable Sensors-Based Hand Gesture Recognition for Human–Robot Collaboration in Construction," *IEEE Sens. J.*, vol. 23, no. 1, pp. 495–505, Jan. 2023, doi: 10.1109/JSEN.2022.3222801.

[21] Y. Jiang, L. Song, J. Zhang, Y. Song, and M. Yan, "Multi-Category Gesture Recognition Modeling Based on sEMG and IMU Signals," *Sensors*, vol. 22, no. 15, p. 5855, Jan. 2022, doi: 10.3390/s22155855.

[22] J. Zhang, Q. Wang, Q. Wang, and Z. Zheng, "Multimodal Fusion Framework Based on Statistical Attention and Contrastive Attention for Sign Language Recognition," *IEEE Trans. Mob. Comput.*, vol. 23, no. 2, pp. 1431–1443, Feb. 2024, doi: 10.1109/TMC.2023.3235935.

[23] Z. Wei, G. M. Ali, and X. Li, "Dynamic Hand-Signal Recognition and Prediction Framework for Crane Operations Using Deep Learning," *Int. Symp. Autom. Robot. Constr. ISARC Proc.*, vol. 2025 Proceedings of the 42nd ISARC, Montreal, Canada, pp. 89–96, Jul. 2025, doi: 10.22260/ISARC2025/0013.

[24] M. Agueh and G. Carlier, "Barycenters in the Wasserstein Space," *SIAM J. Math. Anal.*, vol. 43, no. 2, pp. 904–924, Jan. 2011, doi: 10.1137/100805741.