3D Full-body Pose Estimation of Construction Machines Using Deep Neural Network and Stereo Vision

Han Luo[1], Mingzhu Wang[2,*], Peter Kok-Yiu Wong[3], Pak Him Leung[3], Jingyuan Tang[4], Jack C. P. Cheng[3]

[1] China Three Gorges Investment Management Co., Ltd, China
[2] Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China
[3]Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Hong Kong, China
[4] School of Management Science and Real Estate, Chongqing University, Chongqing, China
*corresponding author
han.luo@connect.ust.hk; m.wang@cityu.edu.hk; kywongaz@connect.ust.hk; phleungae@connect.ust.hk; jtangap@cqu.edu.cn; cejcheng@ust.hk

Abstract –

The movement and posture variability of construction machines is a significant contributor to safety hazards on construction sites. Even when a machine's location is fixed, its moving parts may collide with on-site personnel or objects, leading to injuries or production loss. Accurate estimation of 3D full-body poses of machines can enhance safety by providing more precise spatial information. This paper proposes a framework to estimate 3D full-body poses of construction machines using deep neural networks (DNNs) and stereo vision. The proposed framework employs an entropy-based active learning method to select informative images for fine-tuning the DNN model for 2D pose estimation. 3D poses are estimated through stereo camera calibration, coarse-to-fine stereo matching, and triangulation. Experimental validation using an excavator model achieved an average error percentage (AEP) of 12.11%, demonstrating the framework's feasibility for enhancing safety management.

Keywords –

Computer vision; Construction machine; Deep learning; Deep active learning; Pose estimation, Stereo vision; Construction safety

1 Introduction

Construction sites are inherently hazardous environments, with a significant portion of accidents caused by heavy machinery. In the U.S., the construction sector employed nearly 7.5 million workers in 2018 [1], contributing to 19.77% of all occupational fatalities. Similarly, Hong Kong reported 3,726 work-related injuries in the construction sector in 2018 [2]. Heavy machinery movement is a major source of these hazards, and the interaction between workers and machines exacerbates the risk. Therefore, monitoring machine movements is critical for improving safety. Although surveillance cameras are increasingly used, safety monitoring is often still performed manually, which can be inconsistent and subject to human limitations. Automated monitoring offers a promising solution, but most current systems focus on 2D pose estimation, which is insufficient for evaluating 3D safety conditions. For example, 2D pose estimation cannot accurately capture the spatial relationships between machine components and nearby workers or objects, leading to potential safety hazards. In real-world construction sites, machines like excavators have complex 3D movements, and their moving parts (e.g., booms, arms, and buckets) can pose significant risks to workers even when the machine itself is stationary. The current reliance on 2D monitoring fails to provide the comprehensive spatial information needed to assess these risks accurately. This limitation increases the likelihood of accidents, injuries, and production losses in the construction industry. Therefore, there is a critical need for 3D full-body pose estimation to enhance safety monitoring and mitigate these risks.

Previous studies have utilized devices such as Inertial Measurement Units (IMUs) [3] and Global Positioning Systems (GPS) [4], along with surveillance cameras [5-7], to estimate machine poses, yielding satisfactory results. However, systems relying on pre-installed devices often require multiple installations, making them impractical. In contrast, surveillance cameras can be installed once and provide continuous monitoring. While computer vision techniques are becoming popular for machine pose estimation, most existing studies focus on 2D poses, which are insufficient for evaluating 3D safety conditions. 3D pose estimation can offer critical spatial information about the machine's working zone, which can then be used to assess potential accident risks more accurately. However, research in 3D pose estimation for construction machines remains limited.

This study proposes a framework for 3D full-body machine pose estimation using deep neural networks (DNNs) and stereo vision to provide comprehensive information for safety assessment. The framework uses deep active learning to fine-tune a pre-trained model, minimizing the need for large annotated datasets. The system estimates 2D full-body poses through stereo vision, and 3D poses are derived via stereo matching and triangulation. The proposed method is validated using an excavator model, demonstrating its potential for enhancing safety monitoring on construction sites.

2 Related Works

The safety of construction sites has garnered increasing attention due to its significant potential for spatial-temporal safety analysis and proactive hazard warning. Research in this area focuses on various techniques, including machine pose estimation, transfer learning, and stereo vision, to enhance the monitoring of construction machines and workers.

2.1 Pose Estimation of Construction Machines

Pose estimation is critical for determining the precise working areas of construction machinery, which is essential for safety monitoring. Traditionally, sensors such as Inertial Measurement Units (IMUs) [3] and Global Positioning Systems (GPS) [4] have been employed for estimating machine poses. Tang et al. [8] used IMUs to estimate 3D poses by measuring angles between machine components. While these approaches provide useful data, they are limited by the need for additional hardware installation, increasing costs. GPS and Ultra-Wide Band (UWB) systems [9] offer some advantages but also require extra infrastructure.

In contrast, computer vision has emerged as a cost-effective alternative, utilizing cameras already installed on construction sites. Early computer vision approaches focused on 2D pose estimation. Azar and McCabe [10] proposed a method based on surveillance video analysis using support vector machines (SVMs) and histogram of oriented gradients (HOG). This was later improved by integrating machine skeleton extraction [5] and real-time location systems (RTLS) [11], achieving more accurate 2D pose estimations. However, these methods relied on handcrafted features that required careful design, which limited their generalization capability. More recently, deep neural networks (DNNs) have been used for 2D pose estimation due to their ability to handle complex patterns in data [7]. But it relied solely on 2D image processing, lacking the ability to provide depth information for safety monitoring.

3D machine pose estimation is essential for spatial-temporal safety analysis, as it provides information about the machine's working zone. While stereo vision and DNNs have been combined to extend pose estimation to 3D, existing studies have primarily focused on partial poses, such as the boom, bucket, and stick of excavators [11,12]. These methods fail to provide a comprehensive understanding of the full-body poses of machines and their dynamic working regions, which are crucial for safety evaluations.

2.2 Transfer Learning for On-Site Monitoring

A challenge in machine pose estimation is the need for large, annotated datasets to train reliable DNN models. Manual annotation is time-consuming and costly, particularly for dynamic environments like construction sites. To address this, transfer learning has been explored as a method to apply knowledge gained from one domain to another, reducing the amount of new data required. Transfer learning has been used in construction safety applications, such as Kolar et al.[13], who applied VGG-16 for safety guardrail detection, and Kim et al. [14], who utilized pre-trained models for machine pose estimation.

However, transfer learning approaches often overlook the quality of the annotated data. A smaller high-quality dataset may yield better results than a large poorly annotated one. To improve data efficiency, deep active learning (DAL) has been introduced. DAL allows the model to iteratively select the most informative images for manual annotation, thereby minimizing the amount of data required. Kim et al. [15] demonstrated the effectiveness of DAL in on-site machine detection, showing that it improves DNN performance while reducing manual labeling efforts. While DAL has shown promise, its application in construction machine pose estimation remains underexplored.

2.3 Stereo Vision for On-Site Monitoring

Traditional single-camera systems often struggle with capturing 3D information due to projection and occlusion effects. Stereo vision, which uses multiple cameras, offers a promising solution by reconstructing 3D information from 2D images. Stereo vision has been applied in construction monitoring to estimate the 3D locations of workers and machines [16]. However, previous studies have only provided rough estimations of 3D locations, often ignoring variations in poses during machine operation. Han and Lee [17] explored 3D worker pose estimation using stereo vision, and Yuan et al. [12] applied stereo vision for 3D machine pose estimation by detecting joints and triangulating the corresponding keypoints.

Despite its potential, stereo vision often relies on conventional computer vision techniques, such as HOG-based detection, which have limitations in processing complex 2D images and extracting useful features. Recent advancements suggest that DNNs can overcome these limitations, as demonstrated by researchers in the robotics field [18], who used DNNs for 3D movement tracking of robotic arms. The integration of stereo vision and DNNs for construction machine pose estimation is a promising avenue for improving 3D safety monitoring, but further research is needed to fully explore this approach.

3 The Proposed Framework for 3D Full-body Pose Estimation of Construction Machines Using Deep Neural Network and Stereo Vision

This study proposes a framework for automatically estimating 3D full-body poses of construction machines using deep neural networks (DNNs) and stereo vision, based on images and videos. As illustrated in Figure 1, the framework consists of two sequential modules: (1) 2D machine pose estimation and (2) 3D machine pose estimation.

The process begins by capturing stereo images (e.g., *N* images) using a multi-camera system. Each image is processed independently by a DNN-based 2D pose estimation model to extract keypoints of the machine's pose, along with the confidence probability for each keypoint. Given that collecting a large annotated dataset is time-consuming, deep active learning is utilized to select images with the most informative pose keypoints for fine-tuning the pre-trained 2D pose estimation model.

After fine-tuning, the model estimates 2D machine poses, which are then passed to the 3D pose estimation module. This module first performs stereo camera calibration to obtain intrinsic and extrinsic parameters for triangulation. Using the results from 2D pose estimation, coarse-to-fine stereo matching is applied to refine pose keypoints. Finally, triangulation calculates the 3D coordinates of the machine's poses based on keypoint correspondences and stereo camera calibration results.

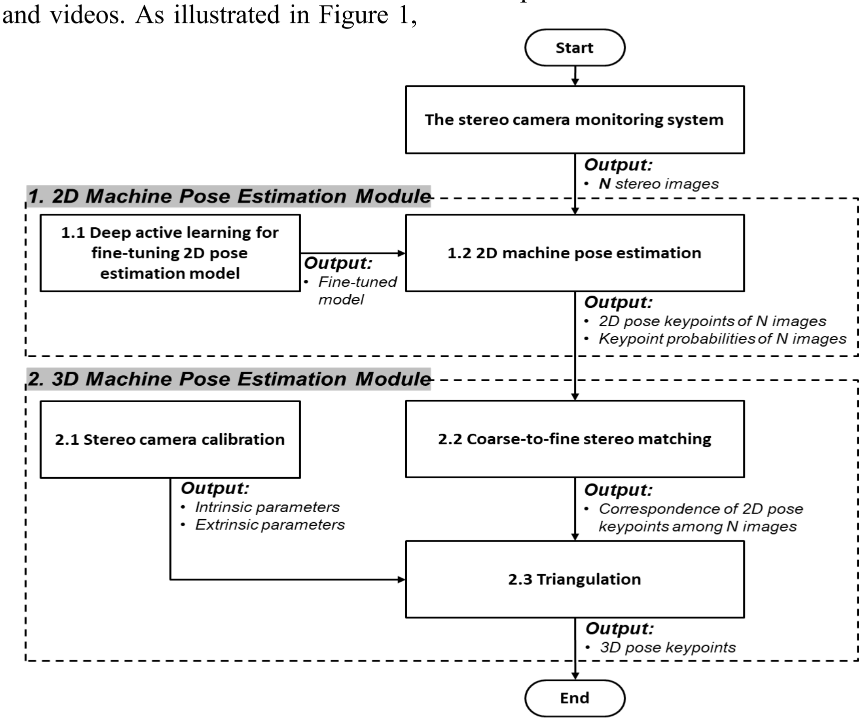

Figure 1. The overall framework for 3D full-body pose estimation of construction machines using DNN and stereo vision.

3.1 2D Machine Pose Estimation Module Based on Deep Active Learning

The goal of this step is to fine-tune a DNN model for 2D machine pose estimation using a small set of manually annotated images collected from new construction sites. Deep active learning (DAL) is employed to reduce the number of training images required by selecting the most informative images for manual annotation. This iterative process involves transferring knowledge from a pre-trained model, which is initially trained on a general dataset of construction sites, to new construction scenes. Figure 2 presents the typical flow of deep active learning for this process.

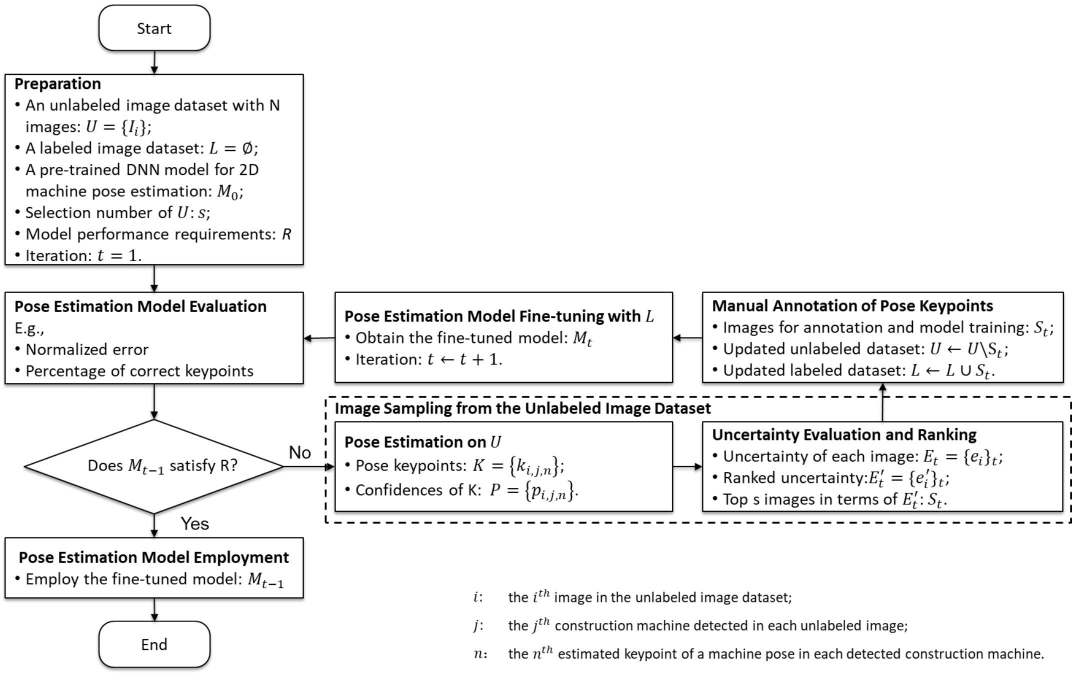

Figure 2. Workflow of deep active learning for fine-tuning the 2D pose estimation model.

Preparation. The first step involves preparing an unlabeled image dataset (*U*), containing images collected from a new construction site. The dataset should include a variety of machine operations to ensure the model can generalize to new images. The pre-trained DNN model (denoted as M_0) is used as the starting point to avoid training from scratch. Pre-defined performance evaluation metrics, such as Normalized Error (NE) and Percentage of Correct Keypoints (PCK), are also defined for model performance assessment.

Pose Estimation Model Evaluation. The performance of the current model is evaluated using the NE and PCK metrics. If the model does not meet the required performance thresholds, the process moves to the next step.

Pose Estimation on U. The 2D pose estimation model is applied to all images in the unlabeled dataset, and the model estimates keypoint coordinates along with the confidence score for each keypoint.

Uncertainty Evaluation and Ranking. The uncertainty of each image is evaluated using entropy-based uncertainty evaluation, where higher entropy indicates more uncertainty. Images with higher uncertainty are considered more informative and are selected for manual annotation.

Manual Annotation of Pose Keypoints. Expert annotators manually label the selected images, providing keypoint annotations. These annotated images are then incorporated into the labeled dataset.

Pose Estimation Model Fine-Tuning with L. The labeled dataset is used to fine-tune the DNN model. The model weights are updated using forward inference and backpropagation.

Pose Estimation Model Employment. Once the fine-tuned model satisfies the performance requirements, it is employed for 2D pose estimation on images and video frames from the target construction site, completing the deep active learning process.

3.2 3D Pose Estimation of Construction Machines Using Stereo Vision

The 2D poses estimated in Section 3.1 are used in the 3D pose estimation module. Stereo camera calibration is firstly performed to obtain necessary camera parameters, and then coarse-to-fine stereo matching is applied to the 2D pose keypoints. This module generates 3D poses by triangulating the keypoints from the stereo images.

3.2.1 Stereo Camera Calibration

Stereo camera calibration is an offline process that provides intrinsic and extrinsic parameters for the triangulation process. The intrinsic parameters describe internal camera characteristics, such as focal length and image center, while extrinsic parameters describe the relative position and orientation of the cameras in the stereo system.

3.2.2 Coarse-to-Fine Stereo Matching

The goal of stereo matching is to find corresponding points from images with different viewpoints. The proposed coarse-to-fine stereo matching strategy includes:

(1) *Keypoint-Based Coarse Stereo Matching:* This step matches pose keypoints based on semantic information (e.g., keypoint type) from 2D pose estimation, as shown in Figure 4.

(2) *Patch-Based Stereo Matching Refinement:* This step refines the stereo matching by using a sliding window approach to compare template patches with search regions to find the best match for each keypoint and therefore improve keypoint correspondence accuracy (as shown in Figure 5).

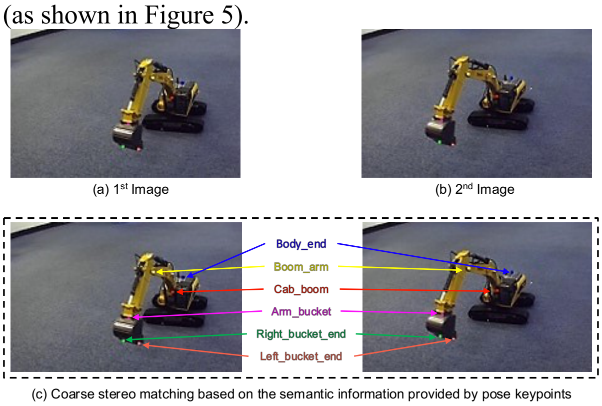

Figure 4. Keypoint-based coarse stereo matching using two stereo images as an example.

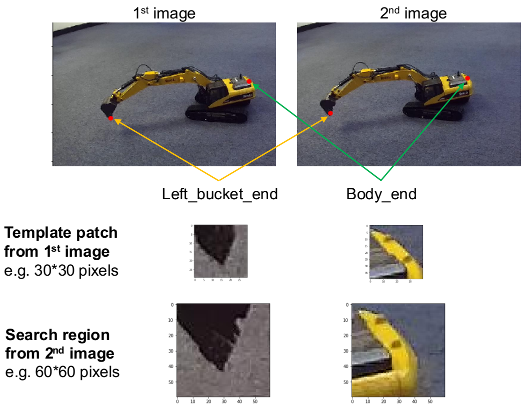

Figure 5. Illustrative examples of template patches and search regions.

3.2.3 Triangulation

The triangulation process uses the refined stereo matching results and camera parameters to compute the 3D position of each keypoint. Linear triangulation is used to solve the equation and determine the 3D coordinates.

4 Validation of the Proposed Framework

To validate the feasibility of the proposed framework, this study uses excavators as illustrative examples due to their numerous movable components and complex poses during construction tasks. The pose keypoints of excavators are defined as shown by the dark dots in Figure 6.

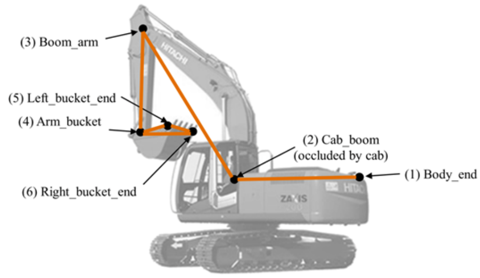

Figure 6. Pose keypoints of construction machines defined in [7].

4.1 Experiment Data Preparation

Due to the lack of a dataset for 3D machine pose estimation using stereo vision, an excavator model was used to create a controlled dataset. The excavator model, a 1:14 scale full-metal replica, performs various tasks like digging, dumping, and swinging during earthmoving operations. The actual lengths of the movable components were measured (see Table 1). A ZED stereo camera was used to capture stereo images, as illustrated in Figure 7. For real-world deployment, cameras can be mounted on fixed infrastructure to ensure a stable and unobstructed view of the construction site. The height and angle of the cameras should be optimized to minimize occlusions and maximize coverage of the working area.

Table 1. Component lengths of the adopted scaled-down excavator model

Component name	Cab	Boom	Arm	Bucket
Length (mm)	180	300	155	80

Figure 7. ZED camera fixed in a holder to capture stereo images of working construction machines.

During data collection, the excavator model mimicked real-world activities such as digging, swinging, and dumping. The dataset consists of 1069 pairs of stereo images, which was split into test and non-test classes using stratified random sampling, with 20% (214 images) allocated for testing and the remaining 80% (855 images) for fine-tuning the DNN models.

4.2 Implementation Details

Two main experiments were conducted to evaluate the framework's performance:

(1) 2D Machine Pose Estimation: This experiment fine-tunes a pre-trained DNN-based model for 2D pose estimation using deep active learning. The model was trained using stereo images captured from Camera 1. The DNN model was fine-tuned using two different sampling strategies: deep active learning (as outlined in Section 3.1) and random sampling.

(2) 3D Full-Body Machine Pose Estimation: After the 2D model was fine-tuned, 3D poses were estimated using both stereo images from Camera 1 and Camera 2. Stereo camera calibration was performed manually, with an overall calibration mean error of 0.051 pixels. The 3D poses were estimated using the proposed coarse-to-fine stereo matching and triangulation method.

The experiments were implemented using Python on Ubuntu 16.04, with an Intel Core i7-6700 CPU @ 3.40 GHz×8 and a NVIDIA GeForce GTX 1080 GPU. The DNN model was built using PyTorch.

4.3 Evaluation Metrics

2D Machine Pose Estimation: The performance of the 2D machine pose estimation model was evaluated using Normalized Error (NE), which calculates the average normalized distance between estimated and actual positions of pose keypoints. The NE formula is defined by Equations (1) and (2).

$$NE = \frac{1}{N}\sum_i \frac{\sum_n \left(\frac{d_{i,n}}{s_i} \cdot \delta(v_{i,\text{n}} \geq 0)\right)}{\sum_n \left(\delta(v_{i,\text{n}} \geq 0)\right)} \tag{1}$$

$$s_i = \max(h, w) \tag{2}$$

where, s_i is the distance normalization parameter in the i^{th} image; h is the height of the bounding box of a detected construction machine in the i^{th} image; w is the width of the bounding box of the detected machine in the i^{th} image; N is the total number of test images as defined in Figure 1; $d_{i,n}$ is the Euclidean distance between the estimated and the actual position of the n^{th} keypoint in the i^{th} image; $v_{i,n}$ means the visibility of the n^{th} keypoint in the i^{th} image.

3D Machine Pose Estimation: The accuracy of 3D pose estimation was evaluated using Average Error Percentage (AEP), comparing the estimated and actual component lengths using Equation (3). A smaller AEP indicates better accuracy in 3D pose estimation.

$$AEP = \frac{1}{N}\sum_i \sum_{mc} \left(\frac{\left|L_{i,mc}^{Estimated} - L_{i,mc}^{Actual}\right|}{L_{i,mc}^{Actual}}\right) \tag{3}$$

where, $L_{i,m}$ is the estimated or actual component length of the m^{th} component of a detected construction machine in the i^{th} image; mc is the indexing of machine component, which can be Cab, Boom, Arm and Bucket.

4.4 Experiment Results and Discussion

4.4.1 2D Pose Estimation Using Deep Active Learning

Table 2 shows the NE results of the 2D pose estimation DNN model, which was fine-tuned using deep active learning. With 25 iterations and 750 images, the fine-tuned model achieved a 77.82% reduction in average NE compared to the pre-trained model. This method significantly reduced the number of images needed for training, demonstrating its efficiency for on-site safety monitoring.

The experiment also analyzed the impact of different sampling strategies for deep active learning. As shown in Figure 8, the NE value decreases significantly as more images are used for fine-tuning. According to Figure 9, the entropy-based uncertainty evaluation method outperforms random sampling in most cases, showing a more effective strategy for selecting images for training.

Table 2. Normalized error (NE) results of the adopted DNN model for 2D pose estimation

Keypoint	Body_end	Cab_boom	Boom_arm	Arm_bucket	Left_bucket_end	Right_bucket_end	Average
Pre-trained (10^{-3})	88.8644	51.1234	48.5361	59.4094	46.8722	43.9860	56.4653
Final fine-tuned (10^{-3})	17.0047	20.1196	6.2324	8.8859	10.6549	12.2473	12.5241
NE reduction	80.86%	60.65%	87.16%	85.04%	77.27%	72.16%	77.82%

Table 3. Results of the proposed 3D pose estimation framework

No	Keypoint-based coarse stereo matching	Patch-based stereo matching refinement		Average error percentage (AEP) (%)					Estimation speed (ms/sample)
		Basic setting (template patch from a random camera)	Confidence-based template patch selection	Cab	Boom	Arm	Bucket	Avg.	
1	√	√	√	**12.86**	10.64	**8.68**	16.26	**12.11**	133.21
2	√	√	×	14.13	10.22	9.16	**15.80**	12.33	132.60
3	√	×	×	18.25	**9.12**	9.71	19.91	14.24	131.61

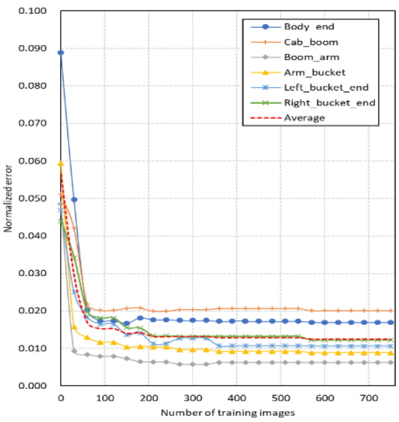

Figure 8. NE curves during deep active learning for 2D pose estimation with different pose keypoints.

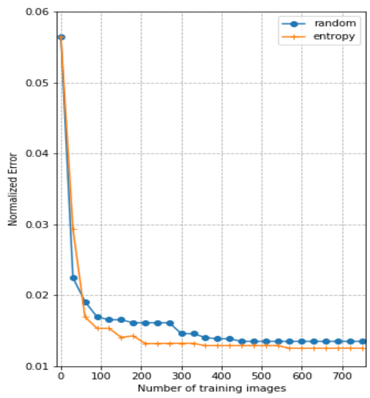

Figure 9. NE curves during deep active learning comparing entropy-based sampling and random sampling.

Table 4. Comparison between the results of the proposed framework and those of previous studies.

No	Method	Experimental condition	Average error percentage (AEP) (%)					Estimation speed (ms/sample)
			Cab	Boom	Arm	Bucket	Avg.	
1	Stereo vison for 3D pose estimation (Ours)	Controlled environment	12.86	10.64	8.68	16.26	**12.11**	133.21
2	Deep learning for 3D pose estimation [6]	Real-world construction site	-	39.10	30.70	58.80	42.80	832.00
3	Stereo vision with location data fusion [11]	Real-world construction site	-	9.50	30.00	-	19.75	≈2000
4	Stereo vision without location data fusion [11]	Real-world construction site	-	36.38	74.21	-	55.29	≈2000

4.4.2 3D Pose Estimation Based on Stereo Vision

Table 3 summarizes the results of 3D pose estimation using the proposed framework. The average AEP for all machine components is 12.11%. The Boom length achieved the smallest error (8.68%), while the Bucket length had the highest error (16.26%). The estimation speed was 133.21 ms per stereo image sample.

The results indicate that the combination of keypoint-based coarse stereo matching and patch-based stereo matching refinement yields the best AEP. Notably, the Boom length was more accurately estimated in the trial with no refinement, likely due to occlusion issues with the Cab_boom keypoint. Regarding the significant error of bucket length, it is computed using Arm_bucket keypoint and the middle point between Left_bucket_end and Right_bucket_end. The 2D pose estimation results of the three keypoints may not be accurate simultaneously, and such errors will be transmitted to 3D machine pose estimation. Hence, the error may exceed expectations, especially when the 2D estimated result of one of the three keypoints deviates from its actual position.

Table 4 presents a comparative analysis of the proposed framework against prior studies, including Ref. [11], which integrates 2D estimation with stereo vision and location data fusion for 3D pose estimation. This comparison serves as a valuable benchmark for evaluating 3D pose estimation performance in the context of construction machinery. The proposed framework demonstrates a substantial improvement in AEP over earlier methods. Specifically, it achieves an AEP of 12.11%, significantly surpassing previous studies that reported AEPs of 42.8%, 19.75%, and 55.29%. Moreover, the estimation speed of the proposed framework is much faster compared to previous studies.

5 Conclusions

Construction site safety is increasingly important due to the hazardous and labor-intensive nature of the industry. Monitoring 3D full-body poses of heavy construction machines is essential for mitigating potential hazards, as it provides comprehensive spatial information about the area occupied by on-site machines. This study proposes a framework for 3D full-body pose estimation of construction machines using deep neural networks (DNNs) and stereo vision. By fine-tuning a pre-trained model with deep active learning and using a coarse-to-fine stereo matching approach, the framework efficiently estimates 3D machine poses based on 2D images. Key contributions include:

(1) Achieved an average AEP of 12.11% for estimating full 3D poses, offering a more comprehensive safety evaluation than partial poses in prior studies;

(2) Applied DNNs to extract features for triangulation, improving efficiency over traditional methods;

(3) Developed a deep active learning strategy with entropy-based uncertainty evaluation, reducing the need for large training datasets;

(4) Proposed a coarse-to-fine stereo matching strategy

to minimize 2D pose estimation errors in 3D estimation.

The framework shows strong potential for jobsite application. However, its real-world deployment faces challenges such as visual complexities, occlusions, lighting variations, and environmental factors. Future work will focus on scaling up the framework for real-world deployment through pilot testing, developing data collection protocols for heavy machinery, and enhancing robustness with techniques like multi-view systems. The framework will undergo rigorous testing in complex jobsite environments and be integrated with real-time safety alerts to detect unsafe actions of construction machines and notify relevant workers and managers instantly.

Acknowledgments

The authors are grateful for the financial support provided by the National Nature Science Foundation of China (Grant number: 72301043).

References

[1] U.S. Bureau of Labor Statistics, "*Graphics for economic news releases - employment levels by industry,*" 2020, Available at: https://www.bls.gov/charts/employment-situation/employment-levels-by-industry.htm (accessed: 13 January 2021).

[2] Occupational Safety and Health Branch, "*Occupational Safety and Health Statistics 2018,*" 2019, Available at: http://www.labour.gov.hk (accessed: 13 January 2021).

[3] J. Tang, H. Luo, P. K.-Y. Wong, and J. C. Cheng, "Study of IMU installation position for posture estimation of excavators," in *Proceedings of the International Conference on Computing in Civil and Building Engineering*, 2020, pp. 980-991: Springer.doi:10.1007/978-3-030-51295-8_68

[4] Leica, "Leica iCON iXE3 - 3D system," 2019.

[5] M. M. Soltani, Z. Zhu, and A. Hammad, "Skeleton estimation of excavator by detecting its parts," *Automation in Construction,* vol. 82, pp. 1-15, 2017.

[6] C.-J. Liang, K. M. Lundeen, W. McGee, C. C. Menassa, S. Lee, and V. R. Kamat, "A vision-based marker-less pose estimation system for articulated construction robots," *Automation in Construction,* vol. 104, pp. 80-94, 2019.

[7] H. Luo, M. Wang, P. K.-Y. Wong, and J. C. Cheng, "Full body pose estimation of construction equipment using computer vision and deep learning techniques," *Automation in Construction,* vol. 110, p. 103016, 2020.

[8] J. Tang, H. Luo, W. Chen, P. K.-Y. Wong, and J. C. Cheng, "IMU-based full-body pose estimation for construction machines using kinematics modeling," *Automation in Construction,* vol. 138, p. 104217, 2022.

[9] F. Vahdatikhaki, A. Hammad, and H. Siddiqui, "Optimization-based excavator pose estimation using real-time location systems," *Automation in Construction,* vol. 56, no. aug., pp. 76-92, 2015.

[10] E. R. Azar and B. Mccabe, "Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos," *Automation in Construction,* vol. 24, no. Jul., pp. 194-202, 2012.

[11] M. M. Soltani, Z. Zhu, and A. Hammad, "Framework for location data fusion and pose estimation of excavators using stereo vision," *Journal of Computing in Civil Engineering,* vol. 32, no. 6, p. 04018045, 2018.

[12] C. Yuan, S. Li, and H. Cai, "Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes," *Journal of Computing in Civil Engineering,* vol. 31, no. 1, p. 04016038, 2017.

[13] Z. Kolar, H. Chen, and X. Luo, "Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images," *Automation in Construction,* vol. 89, no. MAY, pp. 58-70, 2018.

[14] H. Kim, H. Kim, Y. Hong, and H. Byun, "Detecting construction equipment using a region-based fully convolutional network and transfer learning," *Journal of Computing in Civil Engineering,* vol. 32, no. 2, 2018.

[15] J. Kim, J. Hwang, S. Chi, and J. O. Seo, "Towards database-free vision-based monitoring on construction sites: A deep active learning approach," *Automation in Construction,* vol. 120, 2020.

[16] Y. J. Lee and M. W. Park, "3D tracking of multiple onsite workers based on stereo vision," *Automation in Construction,* vol. 98, no. FEB., pp. 146-159, 2019.

[17] S. U. Han and S. H. Lee, "A vision-based motion capture and recognition framework for behavior-based safety management," *Automation in Construction,* vol. 35, pp. 131-141, 2013.

[18] M. Abdelaal, R. M. A. Farag, M. S. Saad, A. Bahgat, H. M. Emara, and A. El-Dessouki, "Uncalibrated stereo vision with deep learning for 6-DOF pose estimation for a robot arm system," *Robotics and Autonomous Systems,* vol. 145, 2021.