Automation in Construction 137 (2022) 104193

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Computer vision for anatomical analysis of equipment in civil infrastructure projects: Theorizing the development of regression-based deep neural networks

Mehrdad Arashpour [a,*], Vineet Kamat [b], Amin Heidarpour [a], M. Reza Hosseini [c], Peter Gill [d]

[a] Department of Civil Engineering, Monash University, Melbourne, Australia
[b] Civil and Environmental Engineering, Electrical Engineering and Computer Science, University of Michigan, United States
[c] School of Architecture & Built Environment, Deakin University, Victoria, Australia
[d] Managing Director for Infrastructure, Donald Cant Watts Corke, Melbourne, Australia

ARTICLE INFO

Keywords:
Artificial intelligence (AI)
Cyber physical systems
Error evaluation metrics
Experimental design and testing
Full body pose estimation
Industry and construction 4.0
Machine learning algorithms
Network backbone architectures

ABSTRACT

There is high demand for heavy equipment in civil infrastructure projects and their performance is a determinant of the successful delivery of site operations. Although manufacturers provide equipment performance handbooks, additional monitoring mechanisms are required to depart from measuring performance on the sole basis of unit cost for moved materials. Vision-based tracking and pose estimation can facilitate site performance monitoring. This research develops several regression-based deep neural networks (DNNs) to monitor equipment with the aim of ensuring safety, productivity, sustainability and quality of equipment operations. Annotated image libraries are used to train and test several backbone architectures. Experimental results reveal the precision of DNNs with depthwise separable convolutions and computational efficiency of DNNs with channel shuffle. This research provides scientific utility by developing a method for equipment pose estimation with the ability to detect anatomical angles and critical keypoints. The practical utility of this study is the provision of potentials to influence current practice of articulated machinery monitoring in projects.

1. Introduction

Interactions between equipment and human in construction and civil infrastructure projects create unique and dynamic production environments [1]. Arrays of sensing and vision systems are transforming construction production into cyber physical systems where self-awareness and context-awareness are improved significantly [2]. Equipment operations in contemporary site environments are pivotal to achieving project deliverables and hence the optimization of performance in the new era of industry 4.0 [3]. Modern equipment performance monitoring should be intelligent and analyze the machine anatomy and project contexts in which equipment is operating [4]. Different equipment poses in proximity of overhead powerlines, trenches and moving traffic has impacts on equipment performance with certain implications for safety, productivity, sustainability and quality of operations [5].

A research stream within the construction literature is using object detection techniques to locate heavy equipment in site scenes [6]. The functionality of object detection is limited by assuming equipment as rigid bodies with no relative joint movement. Articulated site equipment such as hydraulic excavators, mobile cranes and loaders cannot be accurately monitored using only bounding boxes [7]. Accurate keypoint detection is necessary for equipment monitoring with the aim of identifying accident risks, productivity bottlenecks and quality/sustainability issues [8]. Previously, marker-based pose estimation has been undertaken by placing AprilTags or Aruco markers on main joints of articulated equipment [9]. Although markers are reasonably detectable in controlled lab experiments, challenging site conditions such as dusty environment with occlusion and poor lighting reduce the efficiency of marker detection for articulated machine pose estimation [10].

To address this gap, vision-based research for keypoint detection and full body pose estimation uses the power of deep neural networks (DNNs) in analyzing site scenes [11]. Vision-based detection initiates by annotating equipment images to provide ground truth labels of main keypoints and features (see Fig. 1). Then image datasets are partitioned

* Corresponding author.
E-mail addresses: mehrdad.arashpour@monash.edu (M. Arashpour), vkamat@umich.edu (V. Kamat), amin.heidarpour@monash.edu (A. Heidarpour), reza.hosseini@deakin.edu.au (M.R. Hosseini), peter.gill@dcwc.com.au (P. Gill).

https://doi.org/10.1016/j.autcon.2022.104193
Received 28 July 2021; Received in revised form 9 January 2022; Accepted 26 February 2022
Available online 5 March 2022

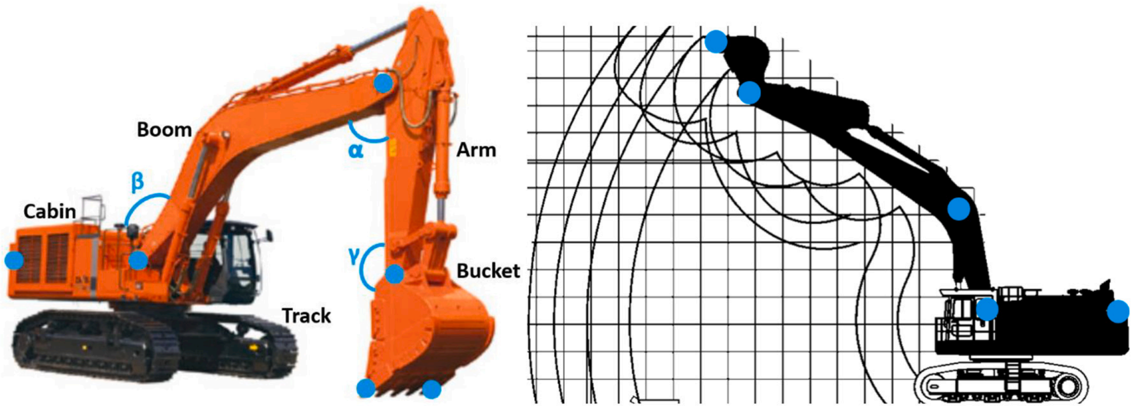

Fig. 1. Anatomical clues of the equipment: critical angels (left) and working ranges (right).

to three subsets for training, validation and testing processes. The processes are run on GPU arrays and many variables including network hyperparameters are finetuned to improve the pose estimation performance [12]. Within the construction literature, networks for human pose estimation are often adopted and the process of transfer learning is implemented to train networks on equipment images and detect keypoints of interest [13]. Adjusting network architectures and finetuning hyperparameters are some of the necessary steps to achieve good quality pose estimation through networks that are not natively developed for construction scenes [14].

The current study develops regression-based networks for anatomical analysis of articulated equipment in civil infrastructure projects. For this purpose, two annotated image libraries of earthmoving equipment are used to train regression-based DNNs. Three backbone architectures are implemented and their performance are evaluated in terms of training time requirements and error metrics such as loss, RMSE and normalized error (NE). Several experiments are performed to yield the optimum pose estimation performance. The proposed workflow and results lay the foundation to conduct anatomical analysis of articulated site equipment.

The structure of the paper is as follows. Section 2 focuses on relevant background research and advancements in equipment pose estimation. Section 3 explains the methods and formulates research questions that govern this study. Section 4 provides experimental results and main findings. Section 5 focuses on interpretation of results and implications for safety, productivity, sustainability and quality of site operations. Conclusions, research limitations and opportunities for future research are presented in Section 6.

2. Background

Civil infrastructure projects are production environments in which machine-worker teams interact to generate intended outputs [15]. Site equipment role in achieving project targets in terms of safety, quality, productivity and sustainability is significant. Previous research has focused on equipment physics and efficient monitoring to enhance performance [16]. Using computer vision and pose estimation techniques are proved to be efficient for equipment monitoring in civil infrastructure projects [10]. The following sections of this paper provide background information on the necessity of monitoring equipment physics, important visual features for use in deep neural networks, and advancements in deep learning for performance analysis of site equipment.

2.1. Performance analysis of site equipment

Civil infrastructure projects are resource-intensive environments in which heavy equipment is an important production agent [17]. Equipment manufacturer's manual (EMM) is the first point of reference for estimating machine performance. However, due to many unique

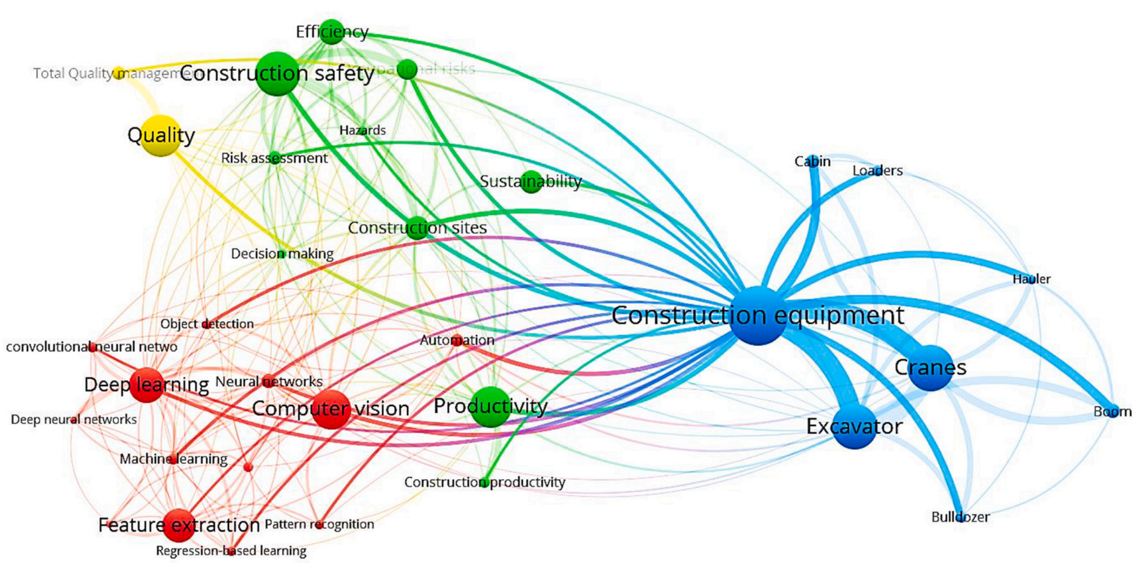

Fig. 2. Map of literature and research links among construction equipment, visual monitoring and project performance.

Table 1
Prior works on pose estimation within the construction context.

Research methods	Description	Author
Synthesized data (unity), Multistage Temporal Convolutional Network (MS-TCN)	Synthesizing Pose Sequences from 3D Assets	Torres Calderon et al. [13]
Gated recurrent unit (GRU)	Predict construction machine poses based on historical motion data and activity attributes	Luo et al. [57]
Stacked hourglass network (HG)	Estimate pose of articulated construction robots	Liang et al. [10,56]
3D pose obtained by matching 2D information across multi- or stereo camera setups	Annotate 2D Imagery with 3D Kinematically Configurable Assets of Construction	Roberts et al. [40,63]
Stacked Hourglass Network (HG), Cascaded Pyramid Network (CPN), ensemble model (HG-CPN)	Estimate full body pose of construction equipment	Luo et al. [11]
Neural network-based estimator	Vision-based estimation of excavator manipulator pose	Xu et al. [70]
Part detection, skeleton extraction method	Part-Based Construction Equipment Pose Estimation	Soltani et al. [55]

variables involved in worksites, performance estimations made through EMM is often misleading [18]. An additional source of inaccuracy is the prevalent assumption in EMMs regarding 100% efficiency in operations [19]. A potential solution is to continuously analyze equipment performance and movements within the context of projects. This is achievable through several techniques such as sensor array deployment [20], tag-based monitoring [21] and vision-based pose estimation [22]. Detection of keypoints using sensors, Aruco mrkers and AprilTags has been the mainstream practice in manufacturing due to the precision and robustness of estimations in controlled factory environments [23]. However, high levels of noise, vibration, occlusion and dust in construction sites reduce efficacy of scene understanding [24]. Consequently, using computer vision techniques is gaining momentum to analyze equipment performance and movements within the context of projects. Such techniques often use image features to undertake keypoint detection and equipment pose estimation.

2.2. Important visual features of site equipment

Site equipment in civil infrastructure projects can be categorized into articulated and rigid body variants [25]. Anatomy of certain equipment such as haulers, drum compactors and pneumatic rollers does not allow relative movements of joints and therefore rigid body models provide some reasonable representations of the equipment. Pose estimation of rigid body equipment can be done by detecting main keypoints on cabin, tracks/tires, front and rear end [26], however, anatomy of articulated equipment allows relative translation and rotation of joints with some degrees of freedom. Anatomical interpretation of articulated equipment requires anatomical angle analysis and detection of keypoints on bucket, arm and boom, as well as cabin, tracks/tires, front and rear ends [27]. Analyzing relative member angles can facilitate inference making on bucket swings and arm/boom status [28], which has significant implications for safety improvement and productivity gains. As can be seen in the map of literature (Fig. 2), mainstream research currently uses computer vision techniques to analyze equipment image features for performance analysis purposes.

As Fig. 2 illustrates, construction safety is an important application area for computer vision. There are many codes of practice and regulations defining safe work requirements for heavy site equipment. For example, occupational safety and health administration (OSHA) demands maintaining buffer zones around powerlines for safe lifting operations using cranes [29]. In this instance, keypoint tracking of boom end and hook block is necessary to avoid entering exclusion zones around overhead transmission lines and the resultant electrocution hazard to equipment operators and bystanders [30]. In addition, productivity improvements can be achieved by keypoint tracking of excavator arm and bucket to detect idleness or interactions with hauler fleets [31]. For a further example, operations above and below ground levels can cause damages to services and equipment. Keypoint tracking can generate timely warning for operators of hydraulic hammers and

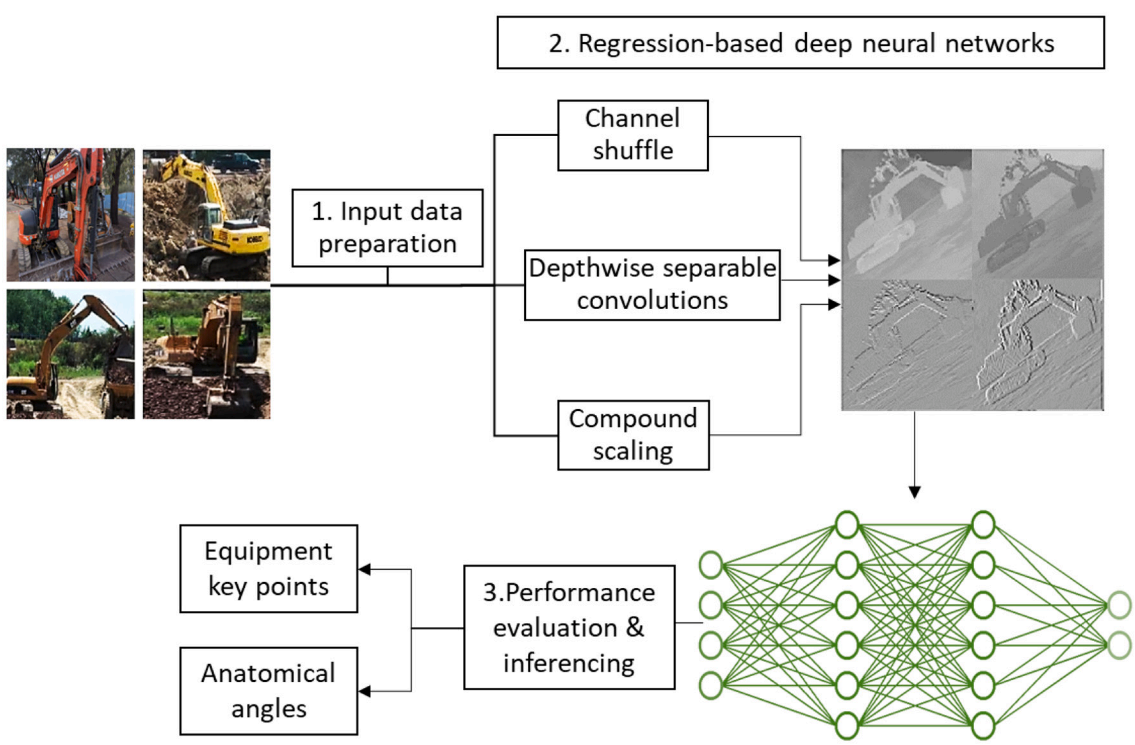

Fig. 3. Workflow of the anatomical analysis of site equipment.

A	B	C	D	E	F	G
image_id	body_end	cab_boom	boom_arm	arm_bucket	bucket_end_left	bucket_end_right
104154_I00000.	40	76	163	148	115	154
104154_I00252.	6	104	234	281	243	255
104154_I01008.	88	78	66	59	80	42
104154_I01260.	65	81	105	93	74	110
104154_I01512.	6	104	233	297	259	259
104154_I01764.	33	76	134	177	226	191
104154_I02268.	185	131	30	67	103	74
104154_I02520.	67	74	91	87	107	64
104154_I02772.	70	74	86	82	101	60
104154_I03276.	85	75	66	54	73	25
104154_I03528.	7	108	231	308	283	292
104154_I04032.	54	76	106	108	98	118
104154_I04788.	84	73	52	49	39	63
104154_I05040.	29	109	184	221	192	217
104154_I05544.	85	71	51	48	38	62
104154_I05796.	8	116	253	328	325	341
104154_I06048.	107	73	35	25	6	37
104154_I06300.	15	108	207	252	221	237
104154_I06804.	154	108	43	23	12	31
104154_I07056	6	114	234	298	277	289

Fig. 4. Sample of input data library including site equipment images and geometric information of main joints.

excavators to avoid quality defects and rework requirements [32].

2.3. *Deep learning for performance analysis of site equipment*

Detection of site entities in previous research has been undertaken using deep CNNs. Pretrained networks are often adopted from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Studies in the civil infrastructure field then customize networks to detect site entities instead of everyday objects [33]. In typical deep CNNs, layers are stacked in specific orders to extract desired prediction outputs from image inputs. Network convolutions can detect patterns using different kernels [14]. An example is detecting personal protection equipment (PPE) for improving safety in civil infrastructure projects with many researchers focusing on detecting hardhat, hi-vis vests or site crews in exclusion zones [34]. Table 1 provides information about previous studies focused on pose estimation of construction equipment.

A second stream of research has used regression networks to analyze tabular site data [35]. Different activation functions enable regression networks to model complexities in project environment and make numerical predictions rather than classifications [36]. Numerous activation functions are used for this purpose including sigmoid, hyperbolic tangent, exponential linear unit and rectified linear unit (ReLU). Regression networks use activation layers to filter optimum features in input data and pooling layers to shortlist the best features [37]. Previous research in this field has used regression-based networks to predict numerical values relative to tangible performance metrics [38]. For instance, Kassem, et al. [35] predicted the excavated soil volume per day using equipment input variables including fuel consumption, bucket payload, travel time and vehicle weight. Main input data format to regression networks is often tabular and networks do not see input

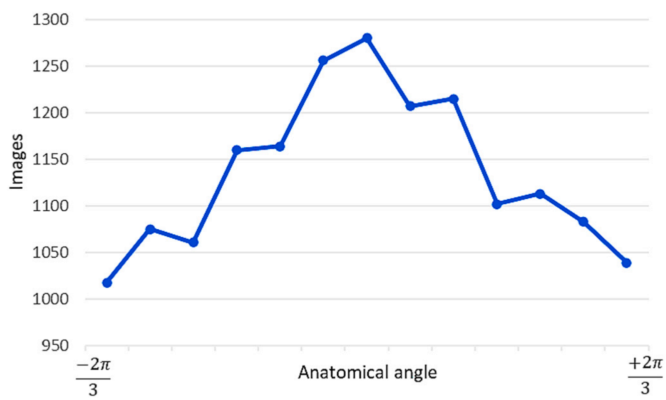

Fig. 5. Distribution of anatomical angles in the training dataset.

images directly. Considering the availability of cameras on most civil infrastructure projects, regression-based networks should be more effectively used [39]. Hence, the current research aims to conduct anatomical analysis of articulated site equipment using regression-based networks that analyze a hybrid of imagery and geometric information of main joints.

3. Methods

To improve performance analysis of site equipment in civil infrastructure projects a workflow is developed to deploy regression-based DNNs and conduct anatomical analysis. The three main research steps are input data preparation, regression-based modeling and performance evaluation/inferencing (see Fig. 3).

3.1. Input data preparation

Two image libraries are generated in which equipment keypoints are annotated and relevant geometric information are recorded in the spreadsheet format. The first image library consists of 1281 images of earthmoving equipment published by Luo, et al. [11]. In this image library, detection of keypoints such as cabin, boom, arm and bucket joints facilitate pose estimation of equipment and analysis of performance. The second image library was built using video footage of earthmoving equipment published by Roberts and Golparvar-Fard [40]. Since required data for regression-based analysis of keypoints were not readily available in this dataset, frames are extracted and annotation of main joints for articulated equipment is done. Image data augmentation is then implemented to increase the size of the two libraries fivefold. Data normalization is conducted on input images to stabilize and also speed up the learning process. Network outputs are also normalized using batch normalization and sigmoid layers. Fig. 4 shows sample images and coordinates of main joints.

Coordinates of main equipment joints assist in obtaining the required ground truth in the training stage. For any three identified joints of the equipment (N_1, N_2, N_3), the corresponding angle is calculated using Eq. (1).

$$Angle = \frac{Arccos\left(N_{12}{}^2 + N_{13}{}^2 - N_{23}{}^2\right)}{2 \times N_{12} \times N_{13}} \tag{1}$$

where N_{12}, N_{13} *and* N_{23} represent the segment length between the two identified joints. For instance, N_{12} can be computed as,

$$N_{12} = \sqrt{(N_{1x} - N_{2x})^2 + \left(N_{1y} - N_{2y}\right)^2} \tag{2}$$

where x, y are image plane coordinates of the detected equipment keypoints. It is worth mentioning that the above equations are only used to obtain the ground truth required for training. Understandably, the trained networks predict anatomical angles directly. Fig. 5 illustrates a breakdown of training data based on anatomical angle distribution. Interestingly, an approximate Gaussian distribution is observed due to the high prevalence of nonextreme angles during construction site works.

The two image libraries along with linked tabular data of critical equipment keypoints and anatomical angles are analyzed using several deep neural networks, with the aim of addressing the first research question in this study on what anatomical features are most suitable for equipment pose estimation in civil infrastructure projects.

3.2. Regression-based modeling by deep neural networks

Previous research shows that pretrained networks for human pose estimation can be adopted and customized to conduct keypoint detection on-site equipment [41]. However, the heatmaps (gaussian maps) that are often used in human pose estimators are disadvantageous on several fronts, including limitations in the prediction of anatomical angles, their use of computationally expensive upsample layers, the necessity of using coordinate refinement to address quantization error, and significant performance drop when using low-resolution imagery [42]. Downsizing the resolution of images is done as a preprocessing step to create a balance between pose estimation performance and pose estimation throughput. Therefore, the current research develops regression-based DNNs using three different backbone blocks of channel shuffle, depthwise separable convolutions, and compound scaling that have shown swift and precise performance for pose estimation (see Fig. 6). The backbone blocks for anatomical analysis of site equipment are repeated in network architectures to extract higher-level features using channel shuffle (172 layers), depthwise separable convolutions (154 layers), and compound scaling (290 layers).

The first backbone block in Fig. 6 consists of group convolution and channel shuffle layers. Group convolutions reduce computational expensiveness of using many dense 1 × 1 convolutions [43]. The resultant problem, however, is division of feature maps in the channel dimension. The lack of information flow among feature maps results in deriving the output from only a portion of input channels [44]. A potential solution is to use transpose and reshape operations to shuffle channels and rearrange the feature maps [45]. The first group convolution (GConv) is a bottleneck layer and squeezes channel numbers while the last GConv layer expands channel numbers to become consistent with the residual connection [46]. In summary, the channel shuffle operation switches output feature maps to minimize computation cost and maximize the detection performance.

The second (middle) backbone block in Fig. 6 consists of depthwise separable convolutions and a projection layer. Depthwise separable convolutions are capable of minimizing computations in shallow layers of DNNs [47]. The projection layer (bottleneck) is a convolution of 1 × 1 that reduces channel numbers. Moreover, a variant of rectified linear unit (ReLU) is used as the activation function with a maximum magnitude of 6 (ReLU6). This activation increases the robustness of detection performance [48]. The residual connection in the backbone architecture improves the gradient flows over the backward pass [49].

The third backbone block in Fig. 6 uses compound scaling to determine number of filters in each layer and number of layers in the network. Larger image input into this architecture results in having more layers and filters to maximize recipient fields for capturing high-level patterns [50]. The MBConv layer in this backbone is an inverted residual block with additional squeeze and excitation (SE) connections. The MBConv layer is routinely used in mobile-optimized DNNs to minimize latency in computer vision tasks such as keypoint detection and pose estimation [51].

Three deep convolutional neural networks are developed using backbone blocks of channel shuffle, depthwise separable convolutions and compound scaling, which all have regression output layers to simultaneously estimate coordinates of main equipment joints and anatomical angles. The three regression-based DNNs can extract higher

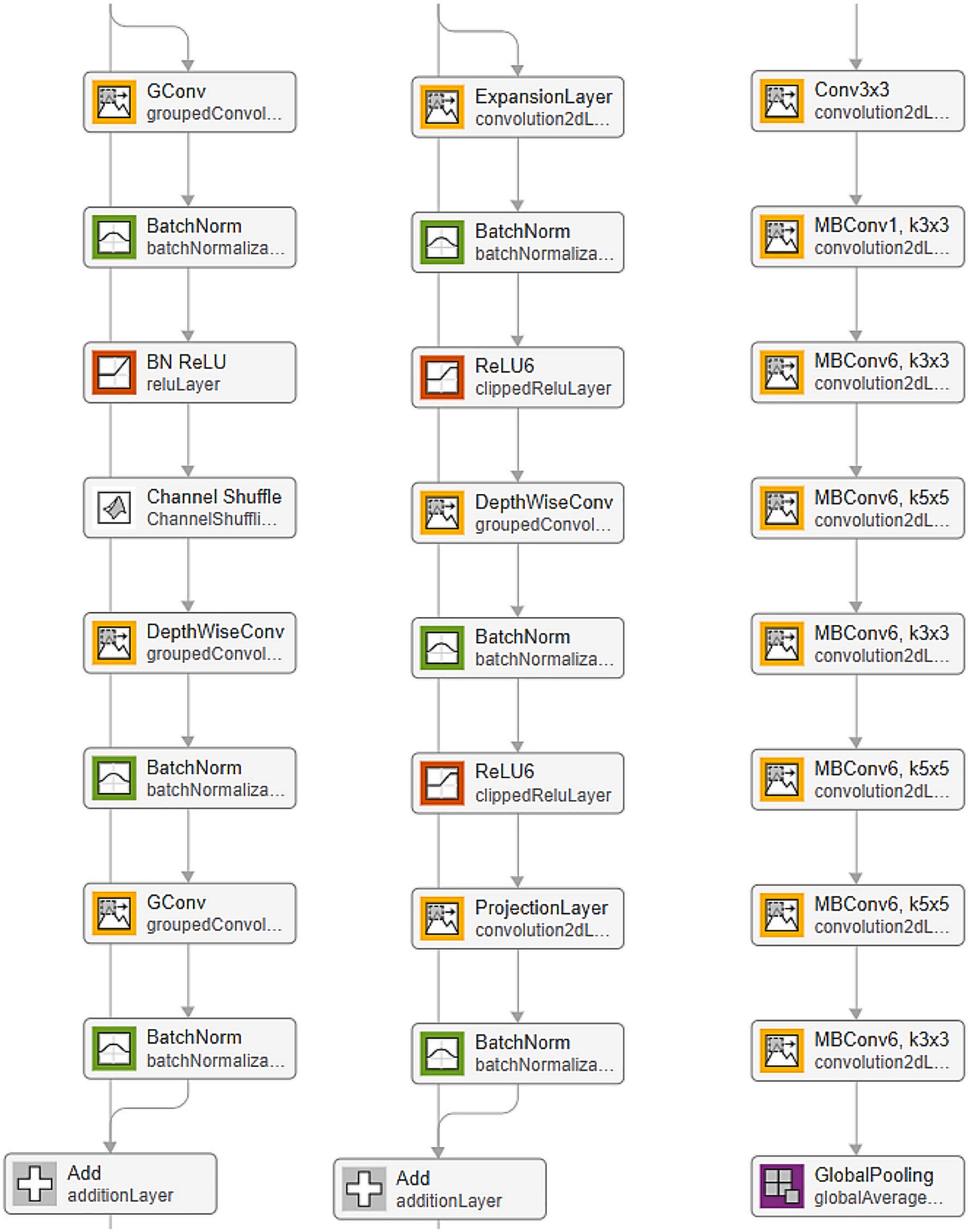

Fig. 6. Backbone architectures for anatomical analysis of site equipment- Channel shuffle (left), depthwise separable convolutions (middle), and compound scaling (right).

level features via activations of different layers. A visualization of network activations can be seen in Fig. 7.

Performance of the three networks were evaluated and compared to address the second research question in this study on whether regression-based deep neural networks can optimally handle equipment anatomical analysis.

3.3. Performance evaluation and inferencing

Several evaluation metrics are used to analyze the performance of the three networks in conducting anatomical analysis of articulated site equipment. For instance, the loss function for the DNN networks is computed using Eq. (3).

$$Loss = \frac{1}{2}\sum_{j=1}^{\varphi}\left(g_j - p_j\right)^2 \tag{3}$$

where φ is the output number, g_j is the ground truth value for keypoint and p_j is the network prediction. Another important evaluation metric for keypoint detection is normalized error (NE).

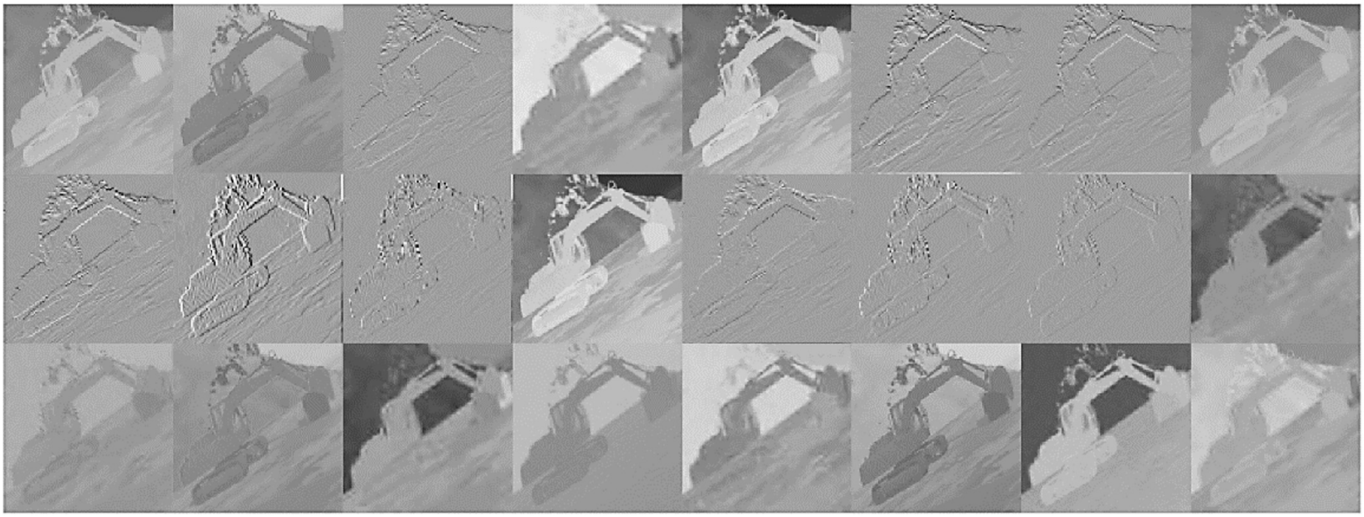

Fig. 7. Display of the activations of different layers- Comparison of the features learned based on areas of activation shows how channels in earlier layers learn simple features and channels in deeper layers learn more complex features.

$$\text{Normalized error }(NE) = \frac{1}{n}\sum_{j=1}^{\varphi}\frac{\lVert g_j - p_j \rVert}{\Delta_j} \tag{4}$$

where Δ_j represents the image diagonal measurement before rescaling. The performance of the three regression-based networks for pose estimation of site equipment is also evaluated using RMSE (Eq. (5)).

$$RMSE = \sqrt{\frac{\sum_{j=1}^{\varphi}\left(g_j - p_j\right)^2}{\varphi}} \tag{5}$$

The fourth important performance metric for regression-based DNNs is the required time to complete training [52]. Training time depends on several variables of the network including the number of floating point operations per second (FLOPS), number of network parameters, and number of memory accesses [53]. Analysis of the above evaluation metrics and prediction errors for the three networks addresses the third research question in this study on how well regression-based DNNs perform when deployed for full anatomical analysis of articulated equipment in civil infrastructure projects.

In addition to using quantitative evaluation techniques, the three developed regression-based models and results were validated using face validation [54]. The models for anatomical analysis and the obtained results were discussed with the industry partner involved. The received feedback confirmed the credibility of the three models and correctness of achieved results.

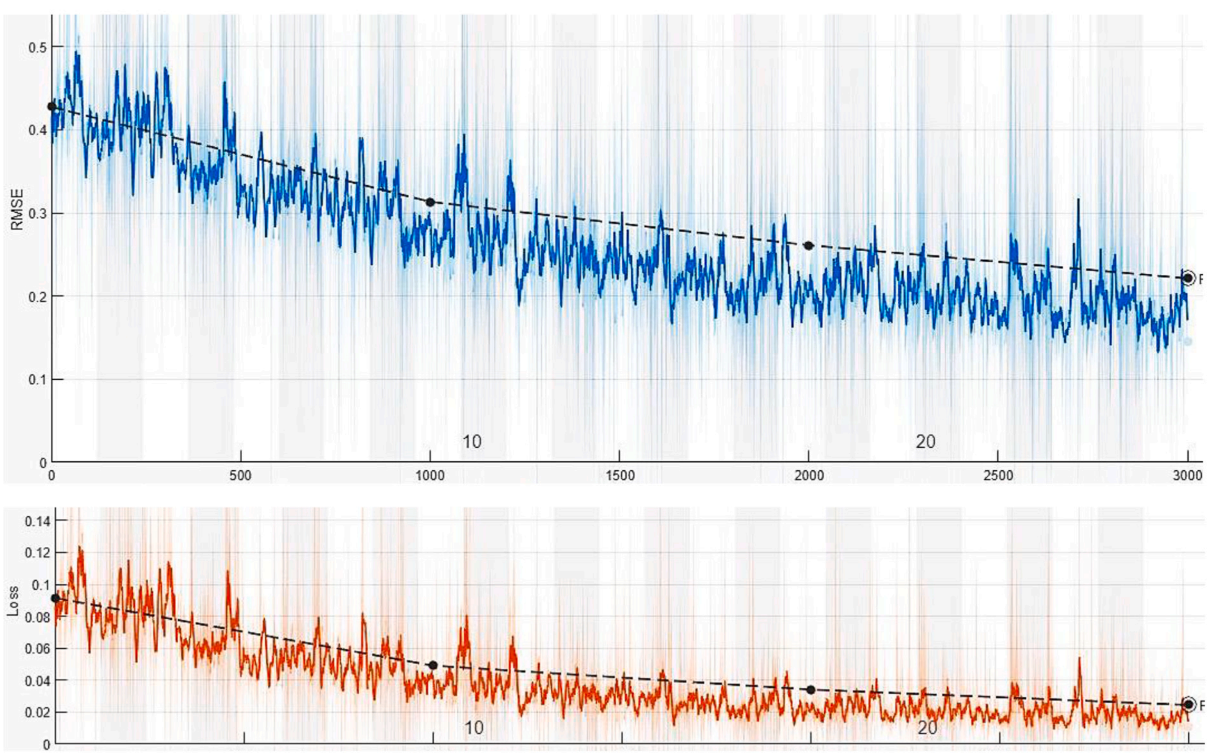

Fig. 8. Performance of equipment keypoint detection using regression-based DNN with channel shuffle.

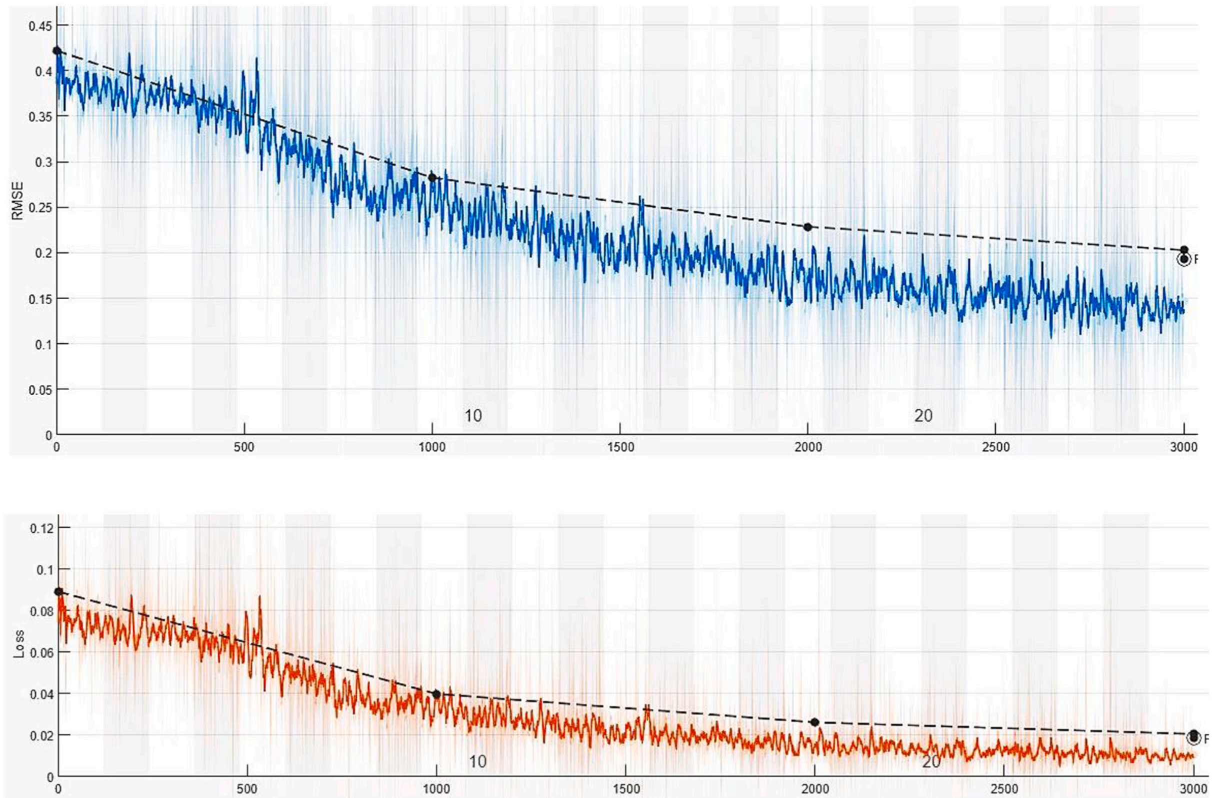

Fig. 9. Performance of equipment keypoint detection using regression-based DNN with depthwise separable convolutions.

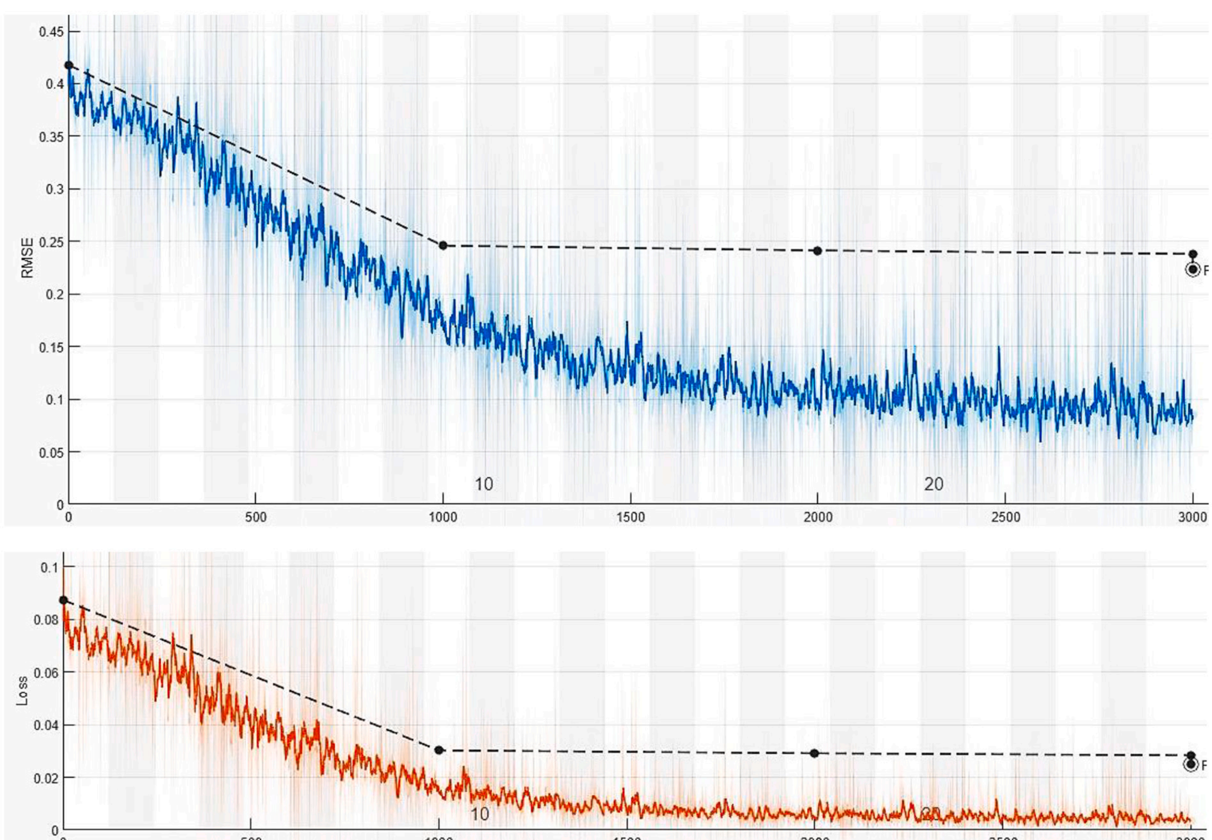

Fig. 10. Performance of equipment keypoint detection using regression-based DNN with compound scaling.

4. Experimental results

The two image libraries are analyzed by the three regression-based networks, for which the backbone architecture blocks are illustrated in Fig. 6. The objective of regression-based analysis is to predict pixel coordinates of equipment main joints and relative angles. The pose estimation technique based on keypoint detection provides the basis for anatomical analysis of site equipment with significant implications for safety, productivity, sustainability and quality of operations in civil infrastructure projects [55]. The first regression-based DNN with channel shuffle architecture is trained using image libraries with data augmentation, which resulted in increasing the number of images five-fold. In training all networks 70% of the hybrid imagery and geometric information is used. Validation and testing are undertaken using 10% and 20% of the input data, respectively. All experiments are performed using an Intel Core i7-10750H @ 5.0GHz, 12 MB and NVIDIA Quadro RTX 3000. The performance of the DNN with channel shuffle architecture can be seen in Fig. 8.

As Fig. 8 illustrates, critical keypoint detection using the first regression-based network yields a decreasing validation RMSE with implementation of piecewise drop to the initial learning rate. The required training time for the regression-model with channel shuffle is a dependent variable to number of model parameters, FLOPS and memory access instances.

The second experimentation on the hybrid imagery and geometric information of main joints is undertaken using the DNN model with depthwise separable convolutions. Understandably, utilization of the ReLU6 activations improves the keypoint detection performance with the validation RMSE decreased further (see Fig. 9). However, this network requires a longer training time than DNN with channel shuffle.

The performance of regression-based DNN with compound scaling is evaluated in the third experiment. As Fig. 10 illustrates, this backbone architecture for anatomical analysis of site equipment yields RMSE values higher than the other two networks. Moreover, the compound scaling model needs the longest training period on the image libraries. Comparison of the three experiments on regression-based DNNs for equipment pose estimation proves the superiority of the network with depthwise separable convolutions in terms of estimation performance. The network with channel shuffle is proved to be the most computationally efficient among the three DNNs.

In the final round of experiments, learned image features are extracted from the three DNNs for anatomical analysis of site equipment. Using features to train image regressors is an efficient method to harness the representational power of DNNs [56,57]. For instance, support vector regression and other regression models can be efficiently trained on the extracted image features. Since DNN networks have hierarchical representation of all input images, deeper layers often contain higher-level features when compared to shallower layers. In order to achieve feature representations of images, activations are applied to deep layers in the three DNNs containing input features over all spatial locations. Then extracted features are used as predictor variables in order to fit the optimum regression model. Performance indicators such as loss and RMSE are comparable to those achieved in previous experiments. Results show the possibility of extracting image features and training regressors for in-situ inferencing in civil infrastructure projects where access to GPU arrays and state-of-the-art hardware is limited.

5. Discussion

Previous studies have used AprilTags and Aruco markers for detection of keypoints on construction site equipment [9]. In addition, vision-based studies have adopted algorithms for human pose estimation and customized them to detect equipment keypoints [11]. This research develops regression-based DNNs for analysis of equipment in civil infrastructure projects with focus on anatomical angles and critical keypoints. Experimental results show that regression-based DNNs are capable of efficiently extracting features from image signals for pose estimation of site equipment. Three backbone architectures are tested for building deep regression networks and their performance is compared using appropriate evaluation metrics such as normalized error in detection of critical keypoints (see Fig. 11).

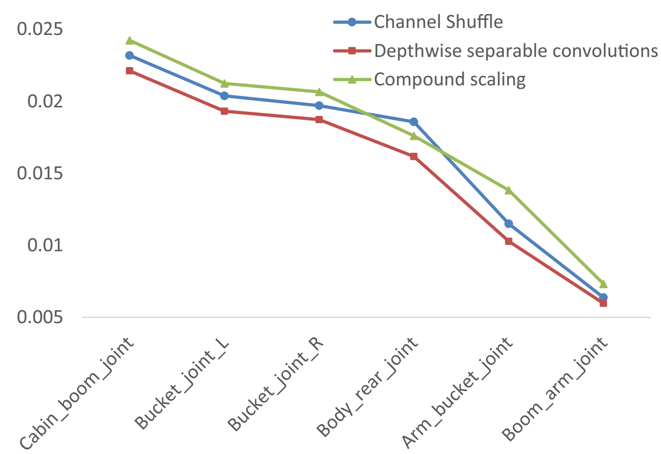

Fig. 11. Normalized error for critical keypoint estimation using regression-based DNNs.

Table 2
Performance of regression-based DNNs for anatomical angle estimation.

DNN backbone architecture	Detection speed (millisecond/image)	Average RMSE
Channel shuffle	**58.44**	0.2231
Depthwise separable convolutions	62.08	**0.1975**
Compound scaling	70.26	0.2247

As can be seen in Fig. 11, the three networks perform well in keypoint estimation of equipment joints especially boom_arm, for which the lowest NE value of 5.97×10^{-3} is achieved using the DNN with depthwise separable convolutions. This is followed by channel shuffle and compound scaling DNNs with NE values of 6.37×10^{-3} and 7.32×10^{-3} respectively. NE values for keypoint detection show an approximate 4% decrease when compared with most recent studies on equipment pose estimation. The highest value of NE is associated with detection of the intersection of equipment boom and cabin (NE value of 24.23×10^{-3} achieved by the DNN with compound scaling). The next highest NE values are associated with the two joints of the equipment bucket (left and right), for which the best performing DNN with depthwise separable convolutions yields NE values of 19.31×10^{-3} and 18.72×10^{-3} respectively. The high values of NE for the previously mentioned joints are associated with the occlusion problem caused by site materials/entities or the equipment itself (self-occlusion) in many images. The performance of the three deep regression-based networks for anatomical analysis is also compared in terms of detection speed and average error (see Table 2).

As can be seen in Table 2 the regression-based DNN with channel shuffle is the most time efficient network with only 58.44 milliseconds spent on detection per image. On the other hand, the DNN with depthwise separable convolutions yields the minimum average error of 0.1975 in estimating the equipment anatomical angles. The regression-based DNN with depthwise separable convolutions is the second swift network with a detection speed of 62.08 milliseconds per image. The following sections discuss practical applications of equipment anatomical analysis using regression-based DNNs.

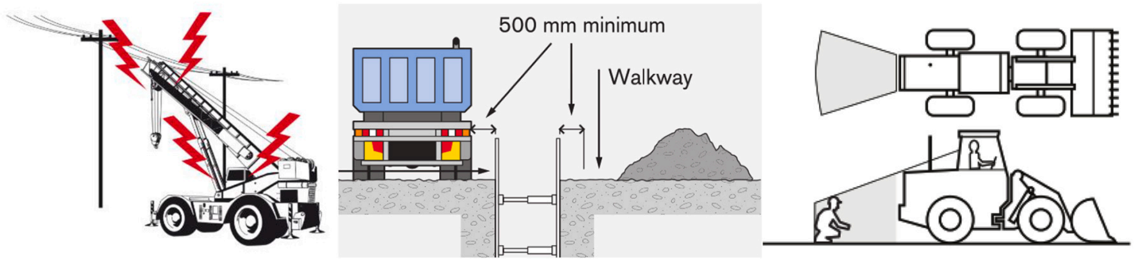

Fig. 12. Anatomical analysis of equipment in site scenes for safety improvement (based on industry standards)- cranes working near overhead powerlines (left), haulers rolling over in trenching operations (middle), and accidents in blind spots (right).

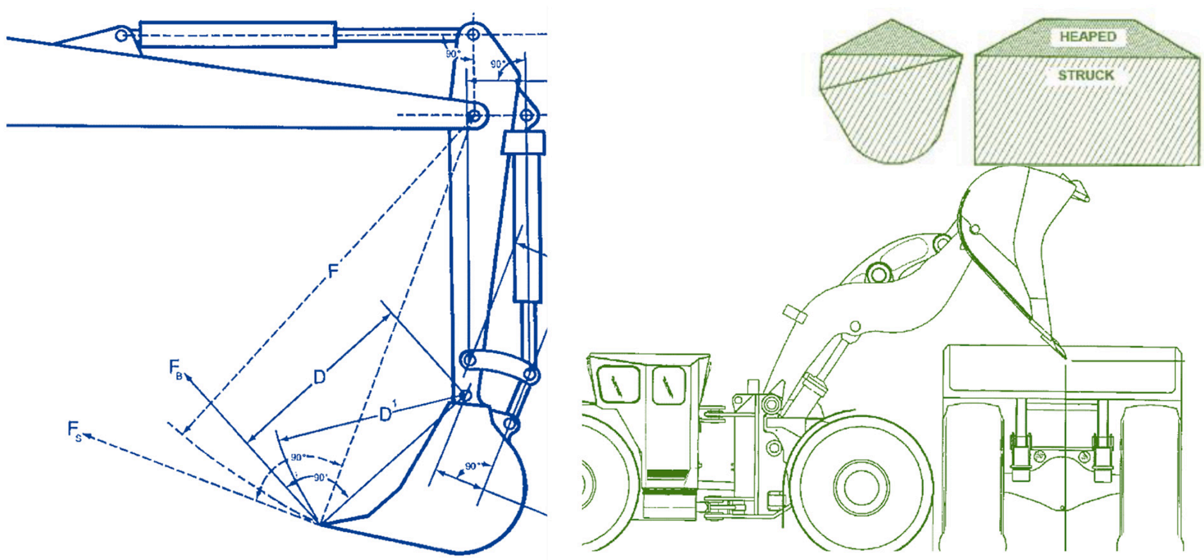

Fig. 13. Anatomical analysis of site equipment and implications for operation productivity and sustainability- Excavator forces (left), loader/hauler cycles (right).

5.1. Safety and quality of site operations

Equipment operations in civil infrastructure projects are often considered as high-risk and must comply with local health and safety legislations [58]. Although site equipment as multipurpose machines can increase efficiency of projects operations, their size and particular anatomy can trigger accident risks [59]. Far-reaching equipment such as telescopic cranes, excavators and scissor lifts can be subject to hazards including rolling over, contacting powerlines, collision with moving vehicles/bystanders, and engaging with underground utilities/networks [60]. Pose estimation and keypoint detection on site equipment have been found effective in improving safety of project operations [61]. As can be seen in Fig. 12, equipment keypoints are associated with multiple risks and their precise analysis is significantly important. For instance, highpoint of equipment boom is associated with electrocution hazards, equipment wheels/tracks are associated with rolling over hazards, and equipment rear points are associated with blind spot collision hazards.

Quality of site operations can also be improved by conducting equipment anatomical analysis with the aim of minimizing defects/rework and double handlings of processes [62]. For instance, underground services are often damaged during construction operations and cause significant delays for project completions [63]. Although preventative measures such as potholing can reveal critical locations surrounding heavy equipment, monitoring critical machine keypoints is necessary to avoid collision [64]. This study shows that regression-based DNNs are useful for keypoint detection of equipment in civil infrastructure projects with high levels of precision and efficiency. This can result in generating warnings to equipment operators and safety crews when there is an imminent risk. Moreover, historical analysis of site images using regression-based DNNs is useful in recording near miss incidents and can facilitate better safety/quality planning and training in future projects.

5.2. Productivity and sustainability of operations

Manufacturers of site equipment usually provide references to predict machinery performance and productivity. It is noteworthy that such references are prepared based on experiments in controlled environments and under certain assumptions [19]. For instance, in estimating the number of bucket passes to load a hauler, the considered parameters are hauler payload, type of soil and bucket payload [65]. Other important variables related to site environments and work dynamics warrant deploying detailed analysis of equipment. This analysis is required for detecting keypoints and anatomical angles, which are directly related to generated forces/emissions and productivity/sustainability of operations (see Fig. 13).

Legislations in the US and Europe mandate manufacturers to use low-emission engines on site equipment to protect environment and improve sustainability [66]. Working conditions, however, differ from site to site and above-average fuel consumption and oil leaks are probable due to high working pressures and overutilization [67]. Detailed analysis of site equipment is required to estimate critical keypoints and angles in harsh working conditions. Anatomical analysis of site equipment will be critical for machine health monitoring and maximizing productivity of operations while minimizing emissions and harm to the environment.

6. Conclusion

Civil infrastructure projects are resource intensive and rely heavily on site equipment among other requirements. The abundance of imagery data has provided unprecedented opportunity for detailed analysis of site equipment with gains regarding safety, quality, productivity and sustainability of operations [68,69]. To this end, the current research focuses on anatomical analysis of site equipment using regression-based DNNs. This is a departure from marker-based pose estimation and adopting human pose estimation networks for analyzing site equipment. Two large image libraries with annotated keypoints are used for training, validation and testing of regression-based DNNs. Data normalization is conducted on input images to stabilize and also speed up the learning process. To estimate equipment pose, three backbone architectures are evaluated with performances cross-compared and validated against ground truth. The regression-based DNNs with channel shuffle are found to be the most computationally time efficient while networks with depthwise separable convolutions yield best estimation performance. This research provides theoretical contributions to the body of knowledge by developing an alternative method for equipment pose estimation with focus on both keypoint and anatomical angle detection. The potential practical contribution of this study is to influence current practice of monitoring heavy equipment in construction and infrastructure projects.

Experiments in this research are conducted on two published datasets of earthmoving operations, which can be considered as a limitation. Computer vision research in the construction and infrastructure requires benchmark imagery datasets for complex tasks. Aerial images of project operations are complementary to site camera photos and facilitate full body pose estimation and detailed anatomical analysis of equipment. Future research can focus on deploying regression-based DNNs for pose estimation of other site equipment such as mobile cranes, scissor lifts and front loaders. There is also a research need to undertake dense keypoint estimation for detailed analysis of equipment and their working environment in the presence of inclement weather conditions such as high wind, heat, rain and humiture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partly funded by the Monash Data Futures Institute (MDFI) grant scheme on "AI and Data Science for Monash Global Challenges. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of MDFI.

The authors acknowledge the work done by Mr. Amin Assadzadeh on preparing the datasets. The assistance of the ASCII Lab members at Monash University is greatly appreciated.

References

[1] D.D. Gransberg, J.A. Rueda-Benavides, Construction Equipment Management for Engineers, Estimators, and Owners, CRC Press, 2020, https://doi.org/10.1201/9781420013993.
[2] C.J. Anumba, A. Akanmu, X. Yuan, C. Kan, Cyber-physical systems development for construction applications, Front. Eng. Manag. 8 (1) (2021) 72–87, https://doi.org/10.1007/s42524-020-0130-4.
[3] A. Sawhney, M. Riley, J. Irizarry, Construction 4.0: An Innovation Platform for the Built Environment, Routledge, 2020. ISBN: 978-0-429-39810-0.
[4] B. Sherafat, et al., Automated methods for activity recognition of construction workers and equipment: state-of-the-art review, J. Constr. Eng. Manag. 146 (6) (2020) 03120002, https://doi.org/10.1061/(asce)co.1943-7862.0001843.
[5] R.L. Peurifoy, C.J. Schexnayder, R.L. Schmitt, A. Shapira, Construction Planning, Equipment, and Methods, McGraw-Hill Education, 2018. ISBN: 9781260108804.
[6] N.D. Nath, A.H. Behzadan, Deep convolutional networks for construction object detection under different visual conditions, Front. Built Environ. 6 (2020) 97, https://doi.org/10.3389/fbuil.2020.00097.
[7] X. Li, H. Wang, L. Yi, L.J. Guibas, A.L. Abbott, S. Song, Category-level articulated object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3706–3715, https://doi.org/10.1109/CVPR42600.2020.00376.
[8] S.B. Atitallah, M. Driss, W. Boulila, H.B. Ghézala, Leveraging deep learning and IoT big data analytics to support the smart cities development: review and future directions, Comp. Sci. Rev. 38 (2020) 100303, https://doi.org/10.1016/j.cosrev.2020.100303.
[9] C. Feng, V.R. Kamat, H. Cai, Camera marker networks for articulated machine pose estimation (in English), Autom. Constr. 96 (2018) 148–160, https://doi.org/10.1016/j.autcon.2018.09.004.
[10] C.J. Liang, K.M. Lundeen, W. McGee, C.C. Menassa, S. Lee, V.R. Kamat, A vision-based marker-less pose estimation system for articulated construction robots (in English), Autom. Constr. 104 (2019) 80–94, https://doi.org/10.1016/j.autcon.2019.04.004.
[11] H. Luo, M. Wang, P.K.Y. Wong, J.C.P. Cheng, Full body pose estimation of construction equipment using computer vision and deep learning techniques (in English), Autom. Constr. 110 (2020), https://doi.org/10.1016/j.autcon.2019.103016. Art no. 103016.
[12] S. Arabi, A. Haghighat, A. Sharma, A deep-learning-based computer vision solution for construction vehicle detection, Comp. Aided Civil Infrastruct. Eng. 35 (7) (2020) 753–767, https://doi.org/10.1111/mice.12530.
[13] W. Torres Calderon, D. Roberts, M. Golparvar-Fard, Synthesizing pose sequences from 3D assets for vision-based activity analysis, J. Comput. Civ. Eng. 35 (1) (2021) 04020052, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000937.
[14] H. Kim, H. Kim, Y.W. Hong, H. Byun, Detecting construction equipment using a region-based fully convolutional network and transfer learning, J. Comput. Civ. Eng. 32 (2) (2018) 04017082, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.
[15] M. Arashpour, A. Heidarpour, A. Akbar Nezhad, Z. Hosseinifard, N. Chileshe, R. Hosseini, Performance-based control of variability and tolerance in off-site manufacture and assembly: optimization of penalty on poor production quality, Constr. Manag. Econ. (2019) 1–13, https://doi.org/10.1080/01446193.2019.1616789.
[16] A. Afanasyev, A. Egoshin, S. Alekseev, The choice substantiation method of road construction equipment for mining enterprises, J. Phys. Conf. Ser. 1399 (3) (2019) 033021, https://doi.org/10.1088/1742-6596/1399/3/033021. IOP Publishing.
[17] N. Gerami Seresht, A.R. Fayek, Dynamic modeling of multifactor construction productivity for equipment-intensive activities, J. Constr. Eng. Manag. 144 (9) (2018) 04018091, https://doi.org/10.1061/(asce)co.1943-7862.0001549.
[18] J.E. Schaufelberger, G.C. Migliaccio, Construction equipment management, Routledge, 2019.
[19] Caterpillar, Performance Handbook, edition 49, Caterpillar Inc, 2019. ISBN: 978-0131152632.
[20] Y. Jiang, X. He, Overview of applications of the sensor technologies for construction machinery, IEEE Access 8 (2020) 110324–110335, https://doi.org/10.1109/ACCESS.2020.3001968.
[21] J. Park, K. Kim, Y.K. Cho, Framework of automated construction-safety monitoring using cloud-enabled BIM and BLE mobile tracking sensors, J. Constr. Eng. Manag. 143 (2) (2017) 05016019, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001223.
[22] Z. Wang, et al., Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase, J. Comput. Civ. Eng. 35 (1) (2021) 04020056, https://doi.org/10.1061/(asce)cp.1943-5487.0000933.
[23] A. Asadzadeh, M. Arashpour, H. Li, T. Ngo, A. Bab-Hadiashar, A. Rashidi, Sensor-based safety management, Autom. Constr. 113 (2020) 103128, https://doi.org/10.1016/j.autcon.2020.103128.
[24] M. Arashpour, T. Ngo, H. Li, Scene understanding in construction and buildings using image processing methods: A comprehensive review and a case study, J. Build. Eng. 33 (2021) 101672, 2021/01/01/, https://doi.org/10.1016/j.jobe.2020.101672.
[25] A.K. Langroodi, F. Vahdatikhaki, A. Doree, Activity recognition of construction equipment using fractional random forest, Autom. Constr. 122 (2021) 103465, https://doi.org/10.1016/j.autcon.2020.103465.
[26] H.I. Moud, M.A. Razkenari, I. Flood, C. Kibert, A flight simulator for unmanned aerial vehicle flights over construction job sites, in: Advances in Informatics and Computing in Civil and Construction Engineering, Springer, 2019, pp. 609–616, https://doi.org/10.1007/978-3-030-00220-6_73.
[27] K.M. Rashid, J. Louis, Times-series data augmentation and deep learning for construction equipment activity recognition, Adv. Eng. Inform. 42 (2019) 100944, https://doi.org/10.1016/j.aei.2019.100944.
[28] Y. Guo, Y. Xu, S. Li, Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network, Autom. Constr. 112 (2020) 103124, https://doi.org/10.1016/j.autcon.2020.103124.
[29] Y. Suh, Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database, Saf. Sci. 142 (2021) 105363, https://doi.org/10.1016/j.ssci.2021.105363.
[30] V.I. Tomakov, M.V. Tomakov, E.G. Pahomova, N.E. Semicheva, N.V. Bredihina, A study on the causes and consequences of accidents with cranes for lifting and moving loads in industrial plants and construction sites of the Russian Federation, J. Appl. Eng. Sci. 16 (1) (2018), https://doi.org/10.5937/jaes16-16478.

[31] A. Roy, Generalizable approaches for tracking, estimating, optimizing, and quantifying uncertainty of fuel use in earthworks operations, University of Toronto, Canada, 2020. Available at: https://www.proquest.com/openview/2afceac124a0356468d2f419c409bf91/1?pq-origsite=gscholar&cbl=18750&diss=y.
[32] W.A. Tanoli, A. Sharafat, J. Park, J.W. Seo, Damage prevention for underground utilities using machine guidance, Autom. Constr. 107 (2019) 102893, https://doi.org/10.1016/j.autcon.2019.102893.
[33] S. Bang, F. Baek, S. Park, W. Kim, H. Kim, An image augmentation method for detecting construction resources using convolutional neural network and uav images, in: M. Al-Hussein (Ed.), 36th International Symposium on Automation and Robotics in Construction (ISARC), 2019, https://doi.org/10.22260/ISARC2019/0085.
[34] J. Wu, N. Cai, W. Chen, H. Wang, G. Wang, Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset (in English), Autom. Constr. 106 (2019) 102894, https://doi.org/10.1016/j.autcon.2019.102894. Art no. 102894.
[35] M. Kassem, E. Mahamedi, K. Rogage, K. Duffy, J. Huntingdon, Measuring and benchmarking the productivity of excavators in infrastructure projects: a deep neural network approach, Autom. Constr. 124 (2021) 103532, 2021/04/01/, https://doi.org/10.1016/j.autcon.2020.103532.
[36] J. Shen, X. Xiong, Y. Li, W. He, P. Li, X. Zheng, Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning, Comp. Aided Civil Infrastruct. Eng. 36 (2) (2021) 180–196, https://doi.org/10.1111/mice.12579.
[37] S. Langer, Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function, J. Multivar. Anal. 182 (2021) 104695, https://doi.org/10.1016/j.jmva.2020.104695.
[38] F. Ning, Y. Shi, M. Cai, W. Xu, X. Zhang, Manufacturing cost estimation based on a deep-learning method, J. Manuf. Syst. 54 (2020) 186–195, https://doi.org/10.1016/j.jmsy.2019.12.005.
[39] R. Khallaf, M. Khallaf, Classification and analysis of deep learning applications in construction: a systematic literature review, Autom. Constr. 129 (2021) 103760, https://doi.org/10.1016/j.autcon.2021.103760.
[40] D. Roberts, M. Golparvar-Fard, End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level, Autom. Constr. 105 (2019), https://doi.org/10.1016/j.autcon.2019.04.006. Art no. 102811.
[41] J. Zhao, Y. Hu, M. Tian, Pose estimation of excavator manipulator based on monocular vision marker system, Sensors 21 (13) (2021) 4478, https://doi.org/10.3390/s21134478.
[42] J. Wang, et al., Deep 3D human pose estimation: a review, Comput. Vis. Image Underst. 210 (2021) 103225, 2021/09/01, https://doi.org/10.1016/j.cviu.2021.103225.
[43] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856, https://doi.org/10.1109/CVPR.2018.00716.
[44] L. Masure, C. Dumas, E. Prouff, A comprehensive study of deep learning for side-channel analysis, in: IACR Transactions on Cryptographic Hardware and Embedded Systems, 2020, pp. 348–375, https://doi.org/10.13154/tches.v2020.i1.348-375.
[45] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116–131, https://doi.org/10.1007/978-3-030-01264-9_8.
[46] T. Li, Z. Zhao, C. Sun, R. Yan, X. Chen, Multi-receptive field graph convolutional networks for machine fault diagnosis, IEEE Trans. Ind. Electron. (2020), https://doi.org/10.1109/TIE.2020.3040669.
[47] A.G. Howard, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint (2017) 1–9. https://arxiv.org/abs/1704.04861v1.
[48] M. Kiyama, M. Amagasaki, M. Iida, Deep learning framework with arbitrary numerical precision, in: 2019 IEEE 13th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), IEEE, 2019, pp. 81–86, https://doi.org/10.1109/MCSoC.2019.00019.
[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.
[50] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, 2019, pp. 6105–6114, *available at:i http://proceedings.mlr.press/v97/tan19a.html*.
[51] M. Tan, R. Pang, Q.V. Le, Efficientdet: scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790, https://doi.org/10.1109/cvpr42600.2020.01079.
[52] L. Chen, M. Zhou, W. Su, M. Wu, J. She, K. Hirota, Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction, Inf. Sci. 428 (2018) 49–61, https://doi.org/10.1016/j.ins.2017.10.044.
[53] S. Han, J. Pool, J. Tran, W.J. Dally, Learning both weights and connections for efficient neural networks, arXiv preprint (2015) 1–9. https://arxiv.org/abs/1506.02626.
[54] A.M. Law, W.D. Kelton, W.D. Kelton, Simulation Modeling and Analysis, McGraw-Hill New York, 2000. ISBN: 9780070592926.
[55] M.M. Soltani, Z. Zhu, A. Hammad, Framework for location data fusion and pose estimation of excavators using stereo vision, J. Comput. Civ. Eng. 32 (6) (2018) 04018045, https://doi.org/10.1061/(asce)cp.1943-5487.0000783.
[56] H. Luo, M. Wang, P.K.-Y. Wong, J. Tang, J.C.P. Cheng, Construction machine pose prediction considering historical motions and activity attributes using gated recurrent unit (GRU), Autom. Constr. 121 (2021) 103444, 2021/01/01/, https://doi.org/10.1016/j.autcon.2020.103444.
[57] C.-J. Liang, K.M. Lundeen, W. McGee, C.C. Menassa, S. Lee, V.R. Kamat, Fast dataset collection approach for articulated equipment pose estimation, in: Computing in Civil Engineering 2019: Data, Sensing, and Analytics, American Society of Civil Engineers Reston, VA, 2019, pp. 146–152.
[58] M. Alkaissy, M. Arashpour, H. Li, S. Alaghmand, A. Nezamian, Quantitative analysis of safety risks and relationship with delayed project completion times, Risk Anal. n/a (n/a) (2021), https://doi.org/10.1111/risa.13778.
[59] H. Luo, M. Wang, P.K.-Y. Wong, J. Tang, J.C. Cheng, Vision-based pose forecasting of construction equipment for monitoring construction site safety, in: International Conference on Computing in Civil and Building Engineering, Springer, 2020, pp. 1127–1138, https://doi.org/10.1007/978-3-030-51295-8_78.
[60] D. Sun, S. Hwang, B. Kim, Y. Ahn, J. Lee, J. Han, Creation of one excavator as an obstacle in c-space for collision avoidance during remote control of the two excavators using pose sensors, Remote Sens. 12 (7) (2020) 1122, https://doi.org/10.3390/rs12071122.
[61] S. Tang, D. Roberts, M. Golparvar-Fard, Human-object interaction recognition for automatic construction site safety inspection, Autom. Constr. 120 (2020) 103356, https://doi.org/10.1016/j.autcon.2020.103356.
[62] D. Roberts, W. Torres Calderon, S. Tang, M. Golparvar-Fard, Vision-based construction worker activity analysis informed by body posture, J. Comput. Civ. Eng. 34 (4) (2020) 04020017, https://doi.org/10.1061/(asce)cp.1943-5487.0000898.
[63] C. Liu, Z. Lei, D. Morley, S.M. AbouRizk, Dynamic, data-driven decision-support approach for construction equipment acquisition and disposal, J. Comput. Civ. Eng. 34 (2) (2020) 04019053, https://doi.org/10.1061/(asce)cp.1943-5487.0000871.
[64] N. Khan, A.K. Ali, M.J. Skibniewski, D.Y. Lee, C. Park, Excavation safety modeling approach using BIM and VPL, Adv. Civ. Eng. (2019), https://doi.org/10.1155/2019/1515808.
[65] T. Kujundžić, M. Klanfar, T. Korman, Z. Briševac, Influence of crushed rock properties on the productivity of a hydraulic excavator, Appl. Sci. 11 (5) (2021) 2345, https://doi.org/10.3390/app11052345.
[66] H.S. Jassim, W. Lu, T. Olofsson, Quantification of energy consumption and carbon dioxide emissions during excavator operations, in: Workshop of the European Group for Intelligent Computing in Engineering, Springer, 2018, pp. 431–453, https://doi.org/10.1007/978-3-319-91635-4_22.
[67] C. Yi, M. Lu, A simulation-based earthmoving fleet optimization platform (SEFOP) for truck/Excavator selection in rough grading project, in: ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction vol. 35, IAARC Publications, 2018, pp. 1–7, https://doi.org/10.22260/ISARC2018/0133.
[68] Y. Pan, L. Zhang, Roles of artificial intelligence in construction engineering and management: a critical review and future trends, Autom. Constr. 122 (2021) 103517, https://doi.org/10.1016/j.autcon.2020.103517.
[69] J. Xu, H.-S. Yoon, Vision-based estimation of excavator manipulator pose for automated grading control, Autom. Constr. 98 (2019) 122–131, 2019/02/01/, https://doi.org/10.1016/j.autcon.2018.11.022.