# Self-supervised Learning Approach for Excavator Activity Recognition Using Contrastive Video Representation

Ghelmani A., Hammad A.
Concordia University, Canada
amin.hammad@concordia.ca

**Abstract.** Detecting construction equipment and classifying their activities are important steps for evaluating their performance and productivity on construction sites. To this end, many automated activity monitoring frameworks based on computer vision have been developed. However, most of the current state-of-the-art activity recognition methods for construction equipment are based on supervised deep learning methods. A major drawback of these approaches is that large, annotated datasets are required for each equipment and activity. To address this problem, this work adapts and customizes the self-supervised, spatiotemporal contrastive video representation learning method for excavator activity recognition. Self-supervised methods use the spatiotemporal information present in video frames as a source of information for pre-training a deep neural network without labels, which is then fine-tuned using a few labeled data. The results show the potential of this method, obtaining the recognition accuracy of 84.6% while using only 10% of the labeled data in an available dataset.

## 1. Introduction

The traditional way of monitoring the activities of various construction equipment is time consuming and labor intensive, especially on large construction sites (C. Chen et al., 2020). Considering that a major step in any large scale construction project is earthmoving operations, excavator activity monitoring would enable site managers to make more informed decisions by providing them with the productivity and work cycle duration information (C. Chen et al., 2020). Examples of such decisions include resource allocation, scheduling, and path planning (Kim et al., 2018).

Recently, with the abundant use of surveillance cameras in construction sites (Kim and Chi, 2019), various automated vision-based activity recognition methods have been proposed (Luo et al., 2020; C. Chen et al., 2020; Kim and Chi, 2019). Convolutional Neural Networks (CNN) are the main building block in all vision-based deep learning methods. For instance, Roberts and Golparvar-Fard (2019) applied a Hidden Markov Model (HMM) on the detection outputs of a CNN network to detect, track, and identify the activities of excavators and dump trucks. Kim and Chi (2020) combined CNN and Long Short-Term Memory (LSTM) architectures for activity recognition of excavators and dump trucks, while Slaton et al. (2020) used CNN and LSTM combination for detecting the sequential activities of excavators and roller compactors.

The main limitation of above-mentioned 2D CNN-based methods is the separate extraction of spatial and temporal features (Li et al., 2020), which limits the efficiency of deep learning models to extract spatiotemporal data simultaneously. 3D CNN-based methods however, incorporate the spatiotemporal data extraction into a single architecture, which alleviates the above limitation. Chen et al. (2020) proposed a three-stage excavator activity recognition method in which excavators are first detected and tracked in consecutive frames. Then, the tracked frames are input into a 3D CNN network for activity recognition. Lou et al. (2020) used the You Only Look Once (YOLOv3) network in a multi-stage framework for worker activity recognition. Recently, Jung et al. (2021) proposed a single-stage architecture by combining 3D

https://doi.org/10.7146/aul.455.c225

CNN with attention mechanism for detecting the activities of multiple construction equipment, while Torabi et al. (2021) used a single-stage architecture called You Only Watch Once 53 (YOWO53) for worker activity recognition. The current state-of-the-art methods for equipment and worker activity recognition, while achieving high performance, are based on supervised deep learning methods, which require large annotated datasets for each equipment/worker and each activity. Particularly, the fact that various types of construction equipment exist on the site, often varying in different phases of the project (Kim and Chi, 2021), poses major challenges for the development of a general supervised activity recognition method.

One possible approach for avoiding the time-consuming and error-prone process of data annotation is using self-supervised methods. In recent years, many self-supervised methods have been proposed for learning visual features from large-scale unlabeled image and video datasets (Jing and Tian, 2021). A popular approach in self-supervised methods is to define a pretext task as the learning objective of the neural network, by which the output label is created from the input data itself (i.e., self-supervision). For instance, some self-supervised video representation learning methods exploit the sampling rate and playback speed information (Cho et al., 2020; Feichtenhofer et al., 2019; Yao et al., 2020), in which a model is trained to maintain the consistency between different representations of the same video at different sampling rates. Another common approach is to use the temporal order of frames or clips (Misra et al., 2016; Xu et al., 2019). Misra et al. (2016) proposed a temporal order verification task in which the model predicts whether or not the input frames have been shuffled or are in the correct order. Xu et al. (2019) extended this task by selecting short clips from an input video and training a model to predict the sequential order of input clips.

Recently, contrastive learning-based self-supervised approaches have obtained the state-of-the-art performance (T. Chen et al., 2020; Grill et al., 2020). In contrastive learning, the pretext task is to have the model produce consistent representations for different augmentations of the same input while increasing the difference with the augmentations of other inputs. Qian et al. (2021) proposed a spatiotemporal contrastive video representation learning (CVRL) method, which uses contrastive learning on spatiotemporally augmented clips extracted from the input video. Compared with supervised methods, self-supervised approaches are able to achieve similar performance (T. Chen et al., 2020) while requiring far less labeled training data.

The main goal of this paper is to adapt and customize the self-supervised CVRL method for the task of excavator activity recognition. To this end, the CVRL method is first trained on a large dataset of unlabeled excavator activities collected from YouTube and local construction sites. Then, it is fine-tuned on a limited annotated dataset. While several works have tried to address the problem of construction equipment detection using limited datasets (Kim and Chi, 2021; Xiao et al., 2021), to the best of authors' knowledge this study is the first attempt to address the problem of excavator activity recognition using self-supervised approaches.

## 2. Methodology

The general framework in which self-supervised methods are applied is comprised of three main phases. The first phase is to use the selected self-supervised learning (SSL) method to pre-train a neural network. The second phase is to use the pre-trained neural network in the downstream task to obtain high performance with a small, labeled dataset. Finally, the third phase is to test the performance of the model trained in previous two phases on the never-before-seen test data. CVRL (Qian et al., 2021) is a self-supervised contrastive video representation learning method. Self-supervised contrastive approaches, in general, try to minimize the distance between the outputs of the model for two different augmentations of the same video, while maximizing their distance with the outputs of the model for augmentations of other videos. In this manner, the model is expected to extract context similarity in the input data without supervision (i.e., labels) (Jing and Tian, 2021). The pre-training phase using the CVRL method is done in five consecutive steps of temporal augmentation, spatial augmentation, clip encoding, projection, and loss calculation (Figure 1). The details of each step are provided in the following paragraphs.

Considering that augmentations are at the core of contrastive approaches, the quality of the learnt representations is determined, in large part, by the choice of the applied augmentations. To this end, and to exploit the spatiotemporal information present in the videos, CVRL proposes a novel temporal augmentation method in which for an input video of length $T$, two clips with the temporal distance of $\Delta t$ are extracted, where $\Delta t$ is sampled from a linearly decreasing distribution over $[0, T]$. The intuition behind using this distribution, which discourages large $\Delta t$ values, is to model the decrease in temporal correlation between extracted clips as the temporal distance between them increases. After sampling two temporally augmented clips from each video, in the second step the spatial augmentations of random cropping, resizing, horizontal flipping, color jittering, gray scaling, and Gaussian blurring are applied in a consistent manner to all of the frames in a clip (e.g., all of them are horizontally flipped).

In the third step, the two augmented clips are input to the selected backbone model to obtain their respective encoding (i.e., representation). CVRL uses the common 3D ResNet (Hara et al., 2018) model as the backbone architecture. In contrastive learning, the model is trained to be invariant to the augmentations applied to be able to produce consistent representations for clips extracted from the same input video. However, information such as the color or object orientation might be important for down-stream tasks. Therefore, in the fourth step, encoding of each extracted clip is passed into a multi-layer projection head to map the encoded representations into an $m$-dimensional space. The use of projection head during training improves down-stream performance by limiting the effect of augmentation invariance to the projection output, on which the contrastive loss is calculated (T. Chen et al., 2020). Projection head is only used during the self-supervised training step and is later discarded. Customizing the projection space dimension plays a crucial balancing role between the quality of the learnt representation during the pre-training phase, and the performance of the model on down-stream tasks. A detailed inspection of the effect of different projection dimensions is presented in Section 3.3.

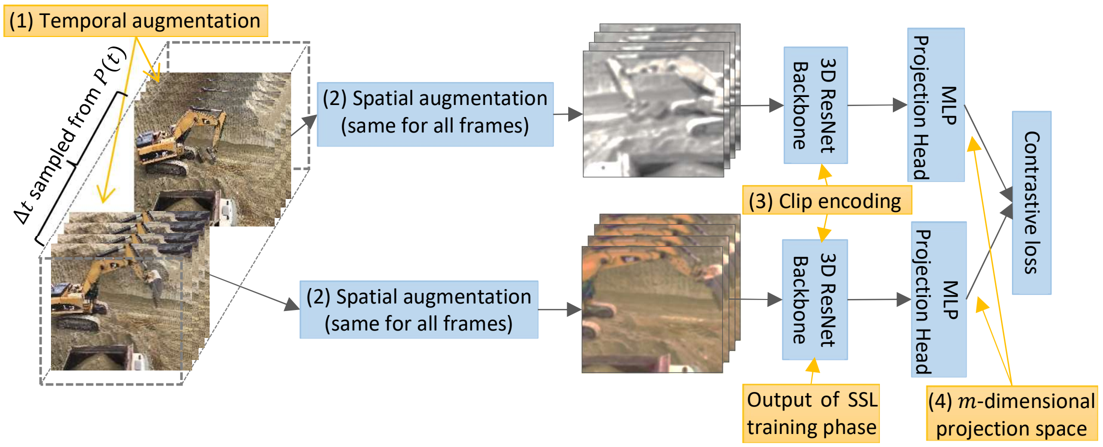

(1) Temporal augmentation
$\Delta t$ sampled from $P(t)$
(2) Spatial augmentation (same for all frames)
3D ResNet Backbone
MLP Projection Head
Contrastive loss
(3) Clip encoding
(2) Spatial augmentation (same for all frames)
3D ResNet Backbone
MLP Projection Head
Output of SSL training phase
(4) $m$-dimensional projection space

Figure 1: Self-supervised pre-training of 3D ResNet backbone.

The final step is to calculate the contrastive loss. CVRL uses the InfoNCE loss (Oord et al., 2019) on the spatiotemporally augmented clips. To apply the contrastive loss, given a batch of $N$ input videos, two augmented clips are extracted from each video, resulting in $2N$ final clips per batch. For a positive pair of clips, i.e., a pair of augmented clips extracted from a single video, the rest of the $2(N-1)$ clips in the batch are treated as negative samples. Assuming $z_i$ and $z_j$ are a positive pair, the contrastive loss for this pair is calculated using Equation (1).

$$\mathcal{L}_i = -\,log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{k\neq i}\,exp(sim(z_i, z_k)/\tau)} \qquad (1)$$

Where, $sim(z_i, z_j)$ represents the cosine similarity between two vectors, and $\mathbb{1}_{k\neq i}$ denotes an indicator function which equals 1 when $k \neq i$ and 0 when $k = i$. $\tau$ is the temperature parameter which improves the performance of the model by increasing or decreasing its confidence in the output prediction (Hinton et al., 2015). Considering that $\tau$ changes the output class probabilities, it affects the calculated loss and subsequently the entire training process. Hence, careful adjustment of this hyperparameter is required for the given dataset and activity classes to obtain the best performance possible, as presented in Section 3.3. Finally, Equation 1 computes the loss for a positive pair of $(i, j)$ clips, the total batch loss is calculated over all positive pairs, both $(i, j)$ and $(j, i)$, in a batch. Considering that in contrastive learning the model learns by comparing the data in each batch, the size of batch plays a crucial role on the quality of the learnt representations (T. Chen et al., 2020). As such, careful customization of this value for the dataset at hand is of paramount importance as shown in Section 3.3.

After self-supervised training, to evaluate the quality of the pre-trained backbone, the most common approach is linear evaluation (T. Chen et al., 2020; Oord et al., 2019). In linear evaluation, the weights of the pre-trained backbone are frozen and a linear classifier is trained on top of the backbone. Considering the limited learning capacity of a linear classifier, the test accuracy in linear evaluation is a proxy for the quality of the representation learnt by the backbone network. In CVRL, backbone network output is $\ell_2$ normalized before being fed into the linear classifier. Prediction loss is then calculated using the output of the linear classifier and the true data label.

Finally, test data are used to obtain the final performance of the model. Following a common practice (Feichtenhofer et al., 2019), during testing ten clips are extracted from each input video by uniformly sampling along the temporal dimension. For each clip, the shorter spatial side is scaled to a size of $n$ pixels and three $n \times n$ crops are then taken along the longer spatial dimension to cover the entire frame, as an approximation to fully convolutional testing. The size $n$, is selected based on the frame size of the videos in the dataset and the input size of the model. Before predicting an activity class for the input video, the output of the softmax layer for the 30 views (10 clips × 3 crops) is averaged over. Afterwards, this averaged softmax value is used to obtain the predicted class of the model and the final test accuracy.

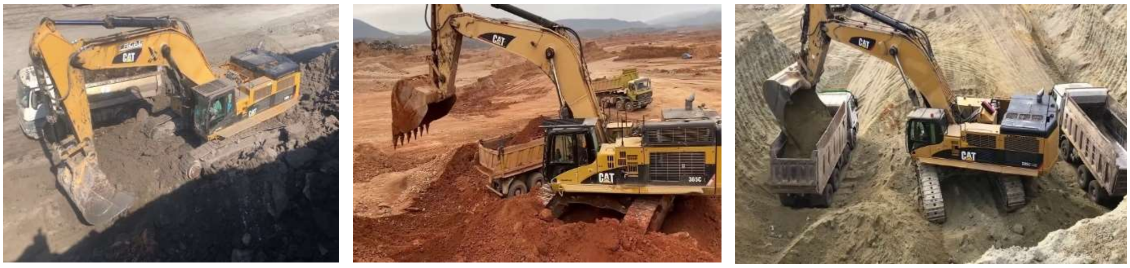
## 3. Experiments

### 3.1 Dataset Description

The experiments were conducted on an excavator video dataset, collected from various sources including local construction sites, YouTube, and videos used in similar research works (Roberts and Golparvar-Fard, 2019). The excavator activities in the dataset include digging, swinging, and loading. The collected videos are from more than 25 different construction sites to add more diversity to the collected dataset by including various site conditions, camera viewpoints, scales, and colors of the excavators. The overall statistics of the dataset are presented in Table 1 and examples of each activity are shown in Figure 2.

Table 1: Statistics of the Collected Excavator Dataset.

| Activity Type | Number of Videos | Number of frames | Average video clip length (sec) |
|---|---|---|---|
| Digging | 295 | 64,436 | 7.28 |
| Swinging | 476 | 51,441 | 3.60 |
| Loading | 321 | 51,632 | 5.36 |
| Total | 1,060 | 163,295 | 5.13 |



(a) Digging

(b) Swinging

(c) Loading

Figure 2: Samples from the Collected Dataset.

### 3.2 Implementation Details

The original CVRL method uses SGD optimization. However, in this work it was found that ADAM optimization algorithm yields better performance. As such, ADAM optimization algorithm with an initial learning rate of 0.16 is used. The learning rate is linearly warmed-up for 5 epochs followed by a half-period cosine learning rate decay strategy. The training is done using two NVIDIA RTX A6000 GPUs. Due to GPU memory limitations, a maximum batch size of 256 videos (512 augmented clips) is considered. During self-supervised training, two 16-frame clips with a temporal stride of 2 are extracted from each video. To increase the variation in the applied spatial augmentations, horizontal flip, color jittering, and gray scaling are only applied 50%, 80%, and 20% of the time, respectively. The initial dimension of the projection space is set to 128 and the self-supervised training is performed for 300 epochs, with the temperature parameter set to 0.1. In contrast to the spatial augmentations applied during the self-supervised pre-training, during linear evaluation, only cropping, resizing, and flipping augmentations are applied.

For linear evaluation, 32-frame clips with a temporal stride of 2 are used, and the training is carried out for 100 epochs with a batch size of 256 videos. The testing is done using the 30 view approach described in Section 2 with a crop size of 256 × 256 per clip. Training, evaluation, and testing data ratios are 80%, 10%, and 10%, respectively. More specifically, during self-supervised pre-training, 80% of the data, without labels, are used to pre-train the 3D ResNet backbone model. During the linear evaluation, the backbone weights are frozen, and the 10% evaluation data are used to train the linear classifier added on top of the backbone. Finally, the performance of the overall framework is evaluated using the 10% test dataset, that was not used in any of the previous two phases, to obtain a more accurate metric for the performance of the trained model.

### 3.3 Experimental Results

The results of linear evaluation on the test dataset for different configurations of the hyperparameters are shown in Table 2 using the common top-1 accuracy metric. In deep learning classification tasks, the final layer of a neural network (softmax), outputs a probability for each class and the final prediction of the network is obtained by finding the class that the network considers to be the most probable (top-1 accuracy). The initial hyperparameters in this work were originally selected in accordance with the hyperparameters suggested in the CVRL work. However, considering the differences between the excavator dataset used in this work and the Kinetics-400 (Kay et al., 2017) and Kinetics-600 (Carreira et al., 2018) human activity datasets used in the CVLR paper, a thorough ablation study on different configurations of hyperparameters was carried out.

Table 2: Linear Evaluation Results.

| Batch size | Temperature | Projection dim. | Learning rate | Top-1 Acc. (%) |
|---|---|---|---|---|
| **32** | 0.1 | 128 | 0.16 | 59.5 |
| **64** | 0.1 | 128 | 0.16 | 66.2 |
| **128** | 0.1 | 128 | 0.16 | 75.5 |
| **256** | 0.1 | 128 | 0.16 | 81.9 |
| 256 | **0.05** | 128 | 0.16 | 80.3 |
| 256 | **0.1** | 128 | 0.16 | 81.9 |
| 256 | **0.5** | 128 | 0.16 | 67.7 |
| 256 | 0.1 | **64** | 0.16 | 78.1 |
| 256 | 0.1 | **128** | 0.16 | 81.9 |
| 256 | 0.1 | **256** | 0.16 | 73.8 |
| 256 | 0.1 | 128 | **0.32** | 70.1 |
| 256 | 0.1 | 128 | **0.16** | 81.9 |
| 256 | 0.1 | 128 | **0.1** | **84.6** |
| 256 | 0.1 | 128 | **0.08** | 83.3 |

As mentioned in Section 2, batch size plays a crucial role in the final performance of contrastive learning-based methods, due to the comparison done between positive and negative samples in each batch. As such, batch size is the first considered hyperparameter to customize for the available excavator dataset and activity classes. It can be seen from the results that the performance of CVRL steadily improves with increasing batch size as has been shown in other contrastive methods (Oord et al., 2019; Tian et al., 2020). Given the role of temperature parameter in the loss calculation and consequently, the entire training process, the temperature parameter is the next considered hyperparameter to customize. The importance and role of temperature parameter is also evident by the large variations in the final performance of the model for different values. It can be seen in Table 2 that the best performance is obtained for temperature parameter of 0.1 while for the value of 0.5 the performance drops significantly.

The third considered hyperparameter is the projection dimension. As explained in Section 2, the trade-off between the quality of the pre-trained model and down-stream performance, is determined by careful selection of the projection dimension. To this end, different dimensions for the projection space were also tested, with the projection dimension of 128 obtaining the best performance. Finally, changing the learning rate shows that the best performance is achieved with the learning rate of 0.1, which differs from the learning rate suggested by the CVRL work (0.32) by a large margin. Considering the significant difference between the performance of the model for learning rate values of 0.16 and 0.32 and the iterative approach of hyperparameter customization in deep learning methods, the results for the customization of batch size, temperature, and projection dimension hyperparameters are reported using the learning rate of 0.16. However, during the final customization of the learning rate hyperparameter, it is observed that the learning rate of 0.1 yields the best performance.

It should be noted that as described in Section 3.2, only 10% of the dataset is used during linear evaluation phase, while the results shown in Table 2 are obtained using the never-before-seen test dataset. Hence, the CVRL model is able to obtain a high activity recognition accuracy of 84.6% while reducing the labeled data requirement by a factor of 10. More generally, the results in Table 2 demonstrate the efficiency and applicability of using self-supervised approaches in the construction domain.

## 4. Conclusion and future work

In this work, the spatiotemporal Contrastive Video Representation Learning (CVRL) method was adapted and customized to the task of excavator activity recognition on construction sites. Extensive linear evaluation of the model and ablation analysis of various hyperparameters demonstrated the promising performance of the CVRL method, by obtaining the top-1 activity classification accuracy of 84.6% while using only 10% of the labeled data from the available dataset. More generally, the results show the applicability of using self-supervised methods for equipment activity recognition in the construction domain.

Considering that the main advantage of the self-supervised approach is reducing the number of required labeled data, the future work includes addressing the problem of recognizing the activities of multiple equipment, which can enhance the applicability of vision-based monitoring during multiple phases of construction projects. Furthermore, while only 10% of the labeled dataset was used in this paper, a more detailed analysis of the ratio of labeled data used and the final classification accuracy of the model can further clarify the advantages and limitations of self-supervised approaches.

## References

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A. (2018). A Short Note about Kinetics-600, arXiv:1808.01340.

Chen, C., Zhu, Z., Hammad, A. (2020). Automated excavators activity recognition and productivity analysis from construction site surveillance videos, Autom. Constr. 110, 103045. https://doi.org/10.1016/j.autcon.2019.103045

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. PMLR, pp. 1597–1607.

Cho, H., Kim, T., Chang, H.J., Hwang, W. (2020). Self-supervised spatio-temporal representation learning using variable playback speed prediction, ArXiv Prepr. arXiv:2003.02692.

Feichtenhofer, C., Fan, H., Malik, J., He, K. (2019). SlowFast Networks for Video Recognition. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised Learning, arXiv:2006.07733.

Hara, K., Kataoka, H., Satoh, Y. (2018). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555.

Hinton, G., Vinyals, O., Dean, J. (2015). Distilling the Knowledge in a Neural Network, ArXiv Prepr. arXiv:1503.02531.

Jing, L., Tian, Y. (2021). Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey, IEEE Trans. Pattern Anal. Mach. Intell. 43, 4037–4058. https://doi.org/10.1109/TPAMI.2020.2992393

Jung, S., Jeoung, J., Kang, H., Hong, T. (2021). 3D convolutional neural network-based one-stage model for real-time action detection in video of construction equipment, Comput.-Aided Civ. Infrastruct. Eng. 37, 126–142. https://doi.org/10.1111/mice.12695

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A. (2017). The Kinetics Human Action Video Dataset, arXiv:1705.06950.

Kim, J., Chi, S. (2021). A few-shot learning approach for database-free vision-based monitoring on construction sites, Autom. Constr. 124, 103566. https://doi.org/10.1016/j.autcon.2021.103566

Kim, J., Chi, S. (2020). Multi-camera vision-based productivity monitoring of earthmoving operations, Autom. Constr. 112, 103121. https://doi.org/10.1016/j.autcon.2020.103121

Kim, J., Chi, S. (2019). Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles, Autom. Constr. 104, 255–264. https://doi.org/10.1016/j.autcon.2019.03.025

Kim, J., Chi, S., Seo, J. (2018). Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks, Autom. Constr. 87, 297–308. https://doi.org/10.1016/j.autcon.2017.12.016

Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., Sebe, N. (2020). Spatio-Temporal Attention Networks for Action Recognition and Detection, IEEE Trans. Multimed. 22, 2990–3001. https://doi.org/10.1109/TMM.2020.2965434

Luo, X., Li, H., Yu, Y., Zhou, C., Cao, D. (2020). Combining deep features and activity context to improve recognition of activities of workers in groups, Comput.-Aided Civ. Infrastruct. Eng. 35, 965–978. https://doi.org/10.1111/mice.12538

Misra, I., Zitnick, C.L., Hebert, M. (2016). Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. In: Computer Vision – ECCV 2016. Presented at the European Conference on Computer Vision, Springer, Cham, pp. 527–544. https://doi.org/10.1007/978-3-319-46448-0_32

Oord, A. van den, Li, Y., Vinyals, O. (2019). Representation Learning with Contrastive Predictive Coding, ArXiv Prepr. arXiv:1807.03748.

Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., Cui, Y. (2021). Spatiotemporal Contrastive Video Representation Learning. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6964–6974.

Roberts, D., Golparvar-Fard, M. (2019). End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level, Autom. Constr. 105, 102811.

Slaton, T., Hernandez, C., Akhavian, R. (2020). Construction activity recognition with convolutional recurrent networks, Autom. Constr. 113, 103138. https://doi.org/10.1016/j.autcon.2020.103138

Tian, Y., Krishnan, D., Isola, P. (2020). Contrastive Multiview Coding. In: Computer Vision – ECCV 2020, Lecture Notes in Computer Science. Springer International Publishing, pp. 776–794. https://doi.org/10.1007/978-3-030-58621-8_45

Torabi, G., Hammad, A., Bouguila, N. (2021). Joint Detection And Activity Recognition Of Construction Workers Using Convolutional Neural Networks. Presented at the 2021 European Conference on Computing in Construction, pp. 212–219. https://doi.org/10.35490/EC3.2021.197

Xiao, B., Zhang, Y., Chen, Y., Yin, X. (2021). A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation, Adv. Eng. Inform. 50, 101372. https://doi.org/10.1016/j.aei.2021.101372

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y. (2019). Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10334–10343.

Yao, Y., Liu, C., Luo, D., Zhou, Y., Ye, Q. (2020). Video Playback Rate Perception for Self-Supervised Spatio-Temporal Representation Learning. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6548–6557.