Proceedings of the ASME 2022
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE2022
August 14-17, 2022, St. Louis, Missouri

DETC2022-89967

HUMAN HAND MOTION PREDICTION IN DISASSEMBLY OPERATIONS

Hao-yu Liao
Graduate Research Assistant
Environmental Engineering Sciences
University of Florida, Gainesville, FL, 32611
haoyuliao@ufl.edu

Minghui Zheng
Assistant Professor
Mechanical and Aerospace Engineering
University at Buffalo, Buffalo, NY, 14260
mhzheng@buffalo.edu

Boyi Hu
Assistant Professor
Industrial and Systems Engineering
University of Florida, Gainesville, FL, 32611
boyihu@ise.ufl.edu

Sara Behdad*
Associate Professor
Environmental Engineering Sciences
University of Florida, Gainesville, FL, 32611
sarabehdad@ufl.edu

ABSTRACT

The remanufacturing workforce can benefit from the capabilities of robotic technology, where robots can alleviate the labor-intensive nature of disassembly operations and help with handling toxic and hazardous materials. However, operators' safety is an important aspect of human-robot collaboration in disassembly operations. This study focuses on predicting human hand motion to provide advanced information to disassembly robots when collaborating with humans. A prediction framework is proposed, which consists of two deep learning models, including convolutional long short-term memory (ConvLSTM) and You Only Look Once (YOLO). ConvLSTM forecasts the next-frame image using images from the disassembly process, and then the YOLO model identifies the human hand object on the predicted image resulting from ConvLSTM. The disassembly images collected from four desktop computers are used to train the ConvLSTM and YOLO. The results reveal that the combined framework of ConvLSTM and YOLO performs well in predicting human hand motion and locating the hand object. The outcomes highlight the need for developing deep learning models capable of recognizing human motion when working with different designs as often remanufacturing workforce have to deal with a wide range of products from different brands, models, and conditions.

Keywords: human motion prediction, end-of-use products, disassembly, ConvLSTM, YOLO, human-robot collaboration

1. INTRODUCTION

Human-robot collaboration for disassembly processes is receiving attention in recent years. Predicting human motions and enhancing product design are among the strategies for improving the operator's safety when interacting with robots [1][2][3].

The human motion prediction is helpful in different applications such as self-driving cars, human-computer interaction, and robotics; however, since human motion is often a complicated stochastic process involving uncertainties, it is still challenging to describe it accurately [4]. This is particularly important in disassembly operations, where human-robot collaboration is still an emerging field.

Recently, a considerable number of studies have conducted human motion prediction utilizing different methods. To name a few, Martinez et al. [5] built a recurrent neural network to forecast human motion. Ding et al. [6] constructed a hidden Markov model on long-term human motion prediction for safe human-robot interaction. Wang et al. [7] applied the Gaussian process to model human motions. Butepage et al. [8] built deep representation learning by comparing different structure neural networks for human motion prediction and classification. Mainprice et al. [9] created a prediction model based on a gaussian mixture model. Cui et al. [10] built a temporal convolutional generative adversarial network on human motion prediction such as walking, smoking, eating, and discussion.

While previous studies have addressed human motion prediction, there remain limitations as the area of disassembly operations has not been explored well. Moreover, many of the previous studies have been focused

on using data collected from wearable sensors [11]. Gril et al. [12] built a linear tensor regression model for human motion prediction in assembly and disassembly operations. Zhang et al. [13] applied the recurrent neural network to predict motion in human-robot collaboration for assembly actions. Liu et al. [14] combined a convolutional neural network (CNN) and long short-term memory network (LSTM) to categorize human motion tasks on assembly by videos. Still, the literature on human motion prediction in disassembly operations is limited. The application of deep learning models for human motion prediction in disassembly operations needs further attention.

Varied from previous studies, this study targets the disassembly operations without using sensor data and relying on the scene images. The human motion prediction is based on (1) predicting the next-frame image and (2) detecting hands using object detection. First, the convolutional long short-term memory (ConvLSTM) model is used to forecast the next-frame image based on previous image frames. Then, the You Only Look Once (YOLO) model detects human hands in the image predicted by ConvLSTM. The prediction duration is 1,000 ms to be considered long-term prediction [15]. To the best of our knowledge, no study combines the ConvLSTM and YOLO model to predict the next-frame image and object detection for human motion prediction.

Human motion prediction is essential for robot awareness and safety. When human workers and robots are disassembling parts simultaneously, it is necessary to avoid any collision. With enhanced prediction, robots predict the next moment and prevent any danger.

In this study, we aim to use deep learning models to identify different models of desktops, each representing a unique design, and further use deep learning to forecast the next-frame image and identify human hand objects.

The remainder of this paper is organized as follows. Section 2 provides an overview of the proposed framework. Section 3 introduces the dataset used for training the deep learning models. Section 4 discusses the results, and Section 5 concludes the paper.

2. METHODOLOGY

The proposed framework consists of two deep learning models: ConvLSTM and YOLO.

2.1 ConvLSTM to predict the next-frame image

Predicting the next frame of an image sequence is a popular topic in artificial intelligence. To name a few studies, Itagi et al. [16] built generative adversarial networks to forecast the next frame of videos. Liu et al. [17] discussed latent space for video prediction. Fujitake et al. [18] trained representation learning for video prediction and online object detection.

Among many available deep learning models, several recent studies have used ConvLSTM. The ConvLSTM is proposed by Shi et al. [19] in 2015 to address the issues of a spatiotemporal sequence forecasting problem for rainfall prediction and is considered spatiotemporal predictive learning [20]. Lotter et al. [21] used ConvLSTM for video prediction along with unsupervised learning. Finn et al. [22] also used ConvLSTM and video prediction for physical interaction with robot arms.

In this study, the ConvLSTM is built with three layers for resized images of 64x64 pixels with one channel of grayscale images. Each pixel value is divided by 255 to be restricted between 0 and 1. The kernel size is 3 by 3, and the ReLU activation function is implemented. The binary cross-entropy is used for loss function with Adam optimizer with 0.0001 learning rate.

2.2 YOLOv3 for hand object detection

The YOLO model was proposed by Redmon et al. [23] in 2016. YOLO is a popular object detection model in various applications. Lan et al. [24] used YOLO for pedestrian detection. Burić et al. [25] adopted YOLO for the ball and player detection. YOLO has evolved with different versions, where the latest version is YOLOv6. This study uses YOLOv3 to identify hand objects. YOLOv3 uses Darknet-53 network as the backbone [26], and has 106 layers [27]. The details of YOLOv3 loss function can be found in Ref. [28].

2.3 The proposed framework for human hand motion prediction

Figure 1 shows the overview of the proposed framework. We use the existing ConvLSTM [29] and YOLO [30] models in this study. The ConvLSTM uses a sequence of disassembly images extracted from videos to forecast the next-frame image. Prior studies used the last 10 frames as input [20]. Thus, in this study, the input is images from time t-10 to t, and the output is the next-frame image for time t+1. The interval between each disassembly image is 1,000 ms. We only consider the 1,000 ms for a long-term prediction [15]. The optimal time length suitable for providing the safety is not discussed in this study. The study only applies the 2D images, which have less information than the kinematics and kinetic data. In future research, the kinematics and kinetic data can supplement the 2D images to increase the prediction accuracy. After predicting the next-frame image by ConvLSTM, the YOLOv3 model will be used to identify the hand object.

The YOLO model is implemented from [30]. The dataset is wrapped up together before shuffling. The default proportion of training and testing is set to 90% and 10% [30]. After splitting into the training and testing phase from the shuffling samples, YOLOv3 will be trained and tested. Different designs of desktops have different layouts and require complex hand movements. When disassembling different brands of desktops, hands will block the status of each component. If the component states occlude the hands, the YOLOv3 cannot detect hands.

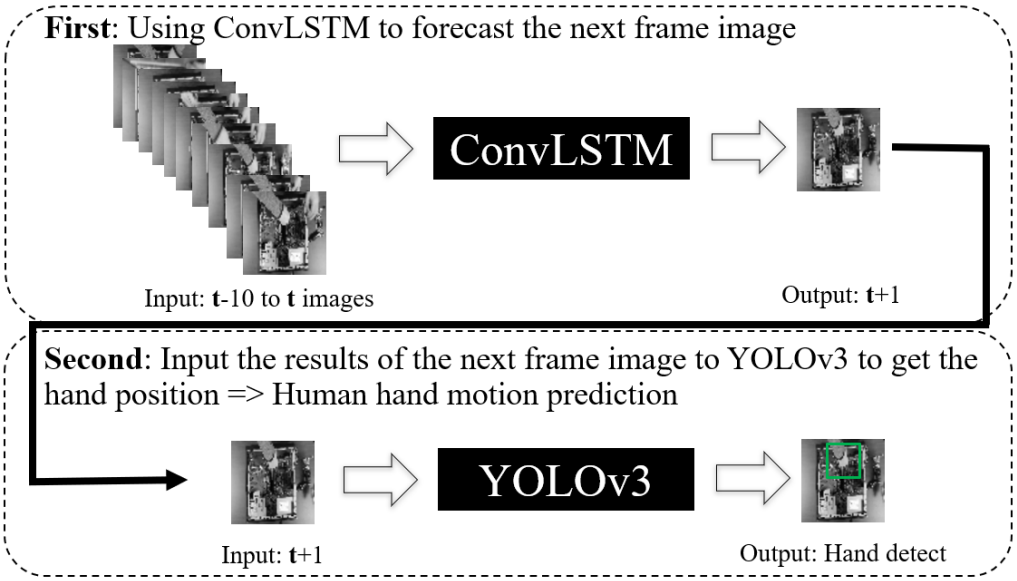

FIGURE 1: The proposed framework consisting of ConvLSTM and YOLOv3

We assume the operator's hands are not occluded in the disassembly process. Further research is needed on identifying overlapping objects and considering hand occlusion covered by component states.

3. BACKGROUND OF DATASET

Four models of desktop computers have been used for data collection and video recording; each desktop is disassembled once. 242, 150, 185, and 200 images have been collected from the disassembly process of Dell XPS, Dell OptiPlex 780, Dell OptiPlex 960, and Dell OptiPlex 990, respectively. The camera resolution on video recordings is 1280 x 720 pixels with 30 fps. The obtained image type is RGB. The distance between the camera and the operator's hand is around 2 ft. The videos are transferred to images. The frame is selected every 30 frames to ensure the interval between each disassembly image is 1,000 ms. The four desktop models have different layouts and positions of components, as shown in Figure 2.

The order of removing components of each desktop is randomly implemented to increase the challenges of model training. The total number of images is 777, and the size of each image is 64 by 64 pixels. Figure 3 shows an example of the disassembly process of Dell OptiPlex 990. The interval between each image is 1,000 ms to reflect the long-term prediction. To make the model training faster, grayscale images have been used.

The images from Dell XPS, Dell OptiPlex 780, and Dell OptiPlex 960 have been used for training and testing. The training and testing proportion is 90% and 10% (519 images for training and 58 for testing).

The images from the three desktops were shuffled before splitting into training and testing. The OptiPlex 990 is used as an unseen dataset to check the model performance. Unlike most studies that combine all datasets to do shuffling before splitting for training models, we want to evaluate if the models can identify an unseen design. Each desktop reflects a different design. The proposed framework should be capable of identifying an unseen design as it is very common in remanufacturing production lines that operators must handle products with different models and conditions.

4. THE RESULTS OF HUMAN HAND MOTION PREDICTION

This section describes the outcomes and elaborates on the limitations of the study.

4.1 Human hand motion prediction

Table 1 shows the loss function values for the training, testing phase, and unseen dataset. To avoid overfitting, the trained parameters are selected such that the training loss and testing loss are close to each other. The result of ConvLSTM is the average loss between the predicted pixel values and ground truth pixel values for a 64x64 image. Considering only 1 pixel for testing, the average loss is 0.60 (2472/64/64). The YOLOv3 combines three types of error: coordinate error, IOU, and classification error to be the loss function [28]. The results of YOLOv3 loss consider all images in contrast to the average loss of an image in ConvLSTM.

To evaluate the prediction errors for ConvLSTM, the mean square error (MSE) is used to calculate the average errors for each pixel. The value of each pixel is between 0 to 255 in the grayscale, and the size of each image is 64 by 64. For example, each image has 4,096 (64x64) pixels, and each pixel is greater than or equal to 0 and less than 255. The MSE for predicted and ground truth images is the summation of square errors for 4,096 pixels divided by 4,096.

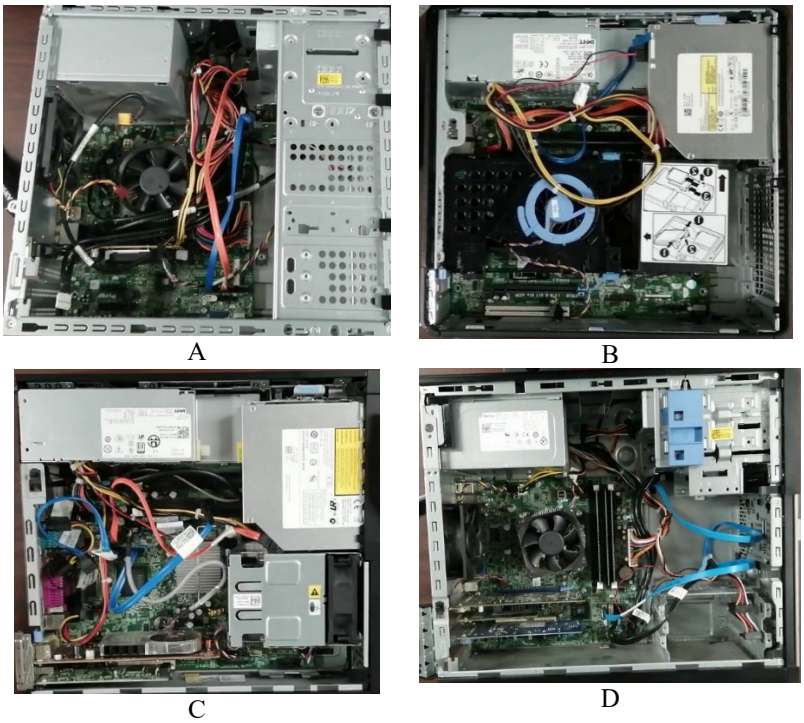

FIGURE 2: The four models of desktops used in the data collection: (A) Dell XPS, (B) Dell OptiPlex 780, (C) Dell OptiPlex 960, and (D) OptiPlex 990.

FIGURE 3: Examples of images of the disassembly process of Dell OptiPlex 990.

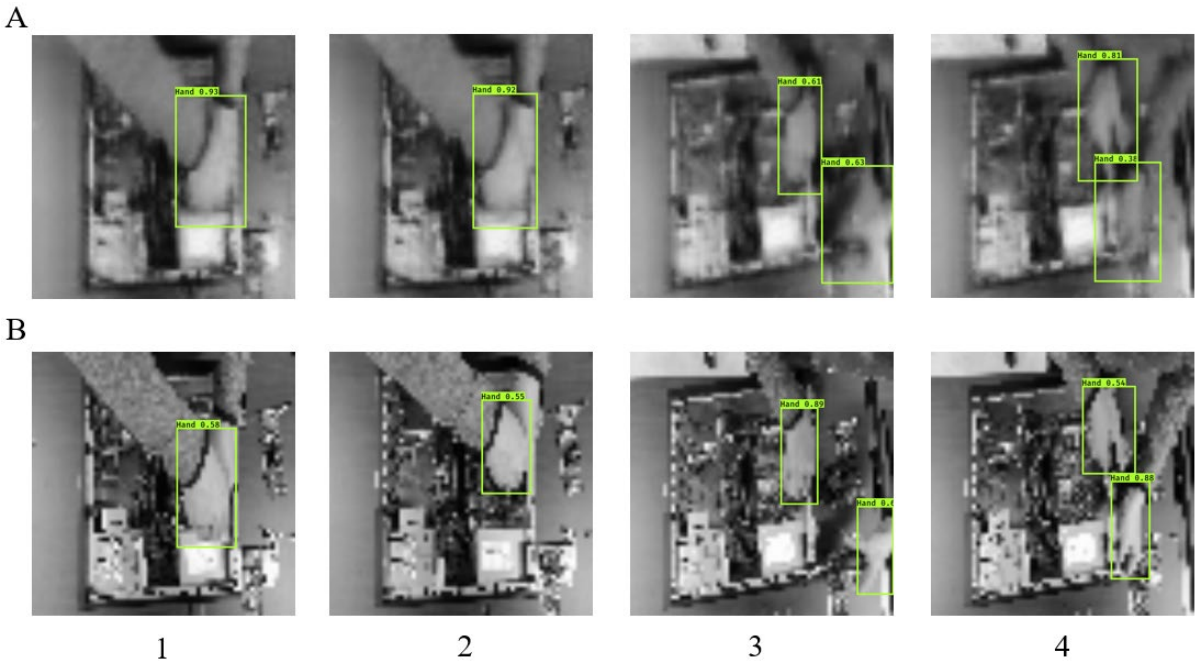

FIGURE 4: Results of OptiPlex 990 as an unseen dataset for the proposed framework on (A) human hand detection in forecasting images and (B) human hand detection on actual images

The training MSE is 0.40, the testing MSE is 0.49, and the MSE of the unseen dataset is 0.41. The maximum MSE is 65,025 (255x255) when the maximum pixel difference value between the predicted and ground truth pixel is 255. The MSE of less than 1 reflects that the predicted images are close to the actual ones.

The overall Intersection over Union (IoU) is used to evaluate the YOLOv3 performance. The IoU of the training and testing phase is 0.79. The IoU greater than 0.5 reflects a reasonable prediction [31]. Figure 4 shows the human hand motion detection for the OptiPlex 990 desktop. Figure 4 (A) is the human hand motion prediction by applying ConvLSTM and YOLOv3, and Figure 4 (B) is only the object detection by YOLOv3 on the actual images.

In Figure 4 (A), although OptiPlex 990 is an unseen dataset, the proposed framework can forecast the next-frame images and locate the hands' position. When both hands are close to each other, YOLOv3 detects them together. When hands are far, YOLOv3 identifies hands separately.

In addition, YOLOv3 has reasonable capabilities to identify hands even though the forecasting images are blurred, as shown in Figure 4 (A)-3.

TABLE 1: The training and testing results of loss function for ConvLSTM and YOLOv3

Phase	ConvLSTM	YOLOv3
Training	2461	15.3
Testing	2472	15.5
Unseen dataset	2538	22.8

4.2 Network structure discussion

The ConvLSTM combines CNN and LSTM. CNN is used for spatial prediction, and LSTM is applied to temporal prediction. The ConvLSTM possesses the above two characteristics for Spatio-temporal prediction. In this study, the disassembly of the different design desktops is a sequence process. The images of the disassembly process have spatial information such as the position of hands and desktop's components and temporal information like the movement of hands. The ConvLSTM has the feature of long short-term memory to remember the historical sequence information and analyze the pictures by the CNN feature. The pattern of human hand motion is regular. Thus, the ConvLSTM structure is suitable for predicting the next frame of images.

The YOLOv3 applies the Darknet-53 network as the backbone for object detection. Except for convolutional layers, the YOLOv3 also has residual neural network (ResNet) layers. The ResNet layers reduce the vanishing gradient problem. The architectures of ConvLSTM and YOLOv3 influence the performance of accuracy and results.

4.3 Limitations of the current study

The proposed framework has several limitations. If the hands' movement is too fast or complicated, the ConvLSTM will produce blurred images, and consequently, YOLOv3 cannot detect the human hands. In addition, the proposed framework is limited in forecasting a larger timescale. Currently, the time interval between each image is 1,000 ms.

The framework cannot predict the motion with a larger timescale, such as the next 10 seconds. The larger the time length gets, the uncertainties of human hand motion prediction increase.

5. CONCLUSION

The study proposes a framework for combining ConvLSTM with YOLOv3 to forecast the human hand motion during disassembly operations. The proposed framework applies deep learning models for enhancing human-robot collaborations in disassembly tasks where remanufacturing operators handle a stream of unknown designs. A dataset of four desktop computers, including Dell XPS, Dell OptiPlex 780, Dell OptiPlex 960, and Dell OptiPlex 990, have been used to evaluate the capabilities of the proposed framework. The results reveal that the proposed framework performs well even with an unseen design, as shown in Figure 4.

The study can be extended in several ways. Other object detection models can be applied for comparison. The resolution and the timescale of prediction can be extended. The uncertainty and complexity of movements can be considered. The current study only considers the next 1,000 ms prediction. Other time lengths can be discussed in further research. Also, other data formats such as kinematic data can be applied. Moreover, further research can apply the active learning approach to other brands and products with more complex miniaturized structures, such as smartphones and medical devices. In addition, the outcomes of the study can be evaluated further by conducting human-robot collaboration experiments.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation–USA under grants #2026276 and #2026533. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] L. Gualtieri, G. P. Monizza, E. Rauch, R. Vidoni, and D. T. Matt, “From design for assembly to design for collaborative assembly-product design principles for enhancing safety, ergonomics and efficiency in human-robot collaboration,” *Procedia CIRP*, vol. 91, pp. 546–552, 2020.

[2] Y. Kim, E. S. Choi, J. Seo, W. Choi, J. Lee, and K. Lee, “A novel approach to predicting human ingress motion using an artificial neural network,” *J. Biomech.*, vol. 84, pp. 27–35, 2019.

[3] J. Wang and L. Shen, “Semi-Adaptable Human Hand Motion Prediction Based on Neural Networks and Kalman Filter,” in *Journal of Physics: Conference Series*, 2021, vol. 2029, no. 1, p. 12091.

[4] Z. Ye, H. Wu, and J. Jia, “Human motion modeling with deep learning: A survey,” *AI Open*, 2021.

[5] J. Martinez, M. J. Black, and J. Romero, “On human motion prediction using recurrent neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.

[6] H. Ding, G. Reißig, K. Wijaya, D. Bortot, K. Bengler, and O. Stursberg, “Human arm motion modeling and long-term prediction for safe and efficient human-robot-interaction,” in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 5875–5880.

[7] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian process dynamical models for human motion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, 2007.

[8] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, “Deep representation learning for human motion prediction and classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6158–6166.

[9] J. Mainprice and D. Berenson, “Human-robot collaborative manipulation planning using early prediction of human motion,” in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 299–306.

[10] Q. Cui, H. Sun, Y. Kong, X. Zhang, and Y. Li, “Efficient human motion prediction using temporal convolutional generative adversarial network,” *Inf. Sci. (Ny).*, vol. 545, pp. 427–447, 2021.

[11] B. Su, “Human motion prediction using wearable sensors and machine Learning.” KTH Royal Institute of Technology, 2021.

[12] L. Gril, P. Wedenig, C. Torkar, and U. Kleb, “A Tensor Based Regression Approach for Human Motion Prediction,” *arXiv Prepr. arXiv2202.03179*, 2022.

[13] J. Zhang, H. Liu, Q. Chang, L. Wang, and R. X. Gao, “Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly,” *CIRP Ann.*, vol. 69, no. 1, pp. 9–12, 2020.

[14] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen, “Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing,” *Procedia CIRP*, vol. 83, pp. 272–278, 2019.

[15] Y. Tang, L. Ma, W. Liu, and W. Zheng, “Long-term human motion prediction by modeling motion context and enhancing motion dynamic,” *arXiv Prepr. arXiv1805.02513*, 2018.

[16] S. Itagi, S. Gowda, T. Udupa, and S. S. Shylaja, “Future Frame Prediction Using Deep Learning,” in *Artificial Intelligence and Technologies*, Springer, 2022, pp. 187–199.

[17] B. Liu, Y. Chen, S. Liu, and H.-S. Kim, “Deep learning in latent space for video prediction and compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 701–710.

[18] M. Fujitake and A. Sugimoto, “Video representation learning through prediction for online object detection,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 530–539.

[19] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[20] Y. Zhou, H. Dong, and A. El Saddik, “Deep learning in next-frame prediction: A benchmark review,” *IEEE Access*, vol. 8, pp. 69273–69283, 2020.

[21] W. Lotter, G. Kreiman, and D. Cox, “Deep predictive coding networks for video prediction and unsupervised learning,” *arXiv Prepr. arXiv1605.08104*, 2016.

[22] C. Finn, I. Goodfellow, and S. Levine, “Unsupervised learning for physical interaction through video prediction,” *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[24] W. Lan, J. Dang, Y. Wang, and S. Wang, “Pedestrian detection based on YOLO network model,” in *2018 IEEE international conference on mechatronics and automation (ICMA)*, 2018, pp. 1547–1551.

[25] M. Burić, M. Pobar, and M. Ivašić-Kos, “Adapting YOLO network for ball and player detection,” in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019, vol. 1, pp. 845–851.

[26] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv Prepr. arXiv1804.02767*, 2018.

[27] S.-C. Lee, H.-E. Tseng, C.-C. Chang, and Y.-M. Huang, “Applying Interactive Genetic Algorithms to Disassembly Sequence Planning,” *Int. J. Precis. Eng. Manuf.*, vol. 21, no. 4, pp. 663–679, 2020, doi: 10.1007/s12541-019-00276-w.

[28] F. Wu, G. Jin, M. Gao, H. E. Zhiwei, and Y. Yang, “Helmet detection based on improved YOLO V3 deep model,” in *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, 2019, pp. 363–368.

[29] R. Panda, “Video Frame Prediction using ConvLSTM Network in PyTorch,” *https://github.com/sladewinter/ConvLSTM*, 2021.

[30] A. Muehlemann, “TrainYourOwnYOLO: Building a Custom Object Detector from Scratch,” *Dispon. on-line https//github. com/AntonMu/TrainYourOwnYOLO (Accedido Diciembre 2020)*, 2019.

[31] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3150–3158.