Realization of the physical to virtual connection for digital twin of construction crane

Enliu Yuan[a c], Jian Yang[b], Mohamed Saafi[a], Fei Wang[c], Jianqiao Ye[a,*]

[a] School of Engineering, Lancaster University, Lancaster, LA1 4YW, UK

[b] Shanghai Key Laboratory for Digital Maintenance of Buildings and Infrastructure, School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, 200240, PR China

[c] Shanghai Waterway Engineering Design & Consulting Co., Ltd Pudong New District Shanghai, 200120，

Abstract

Digital twin is an integrated multi-physics representation of a complex physical entity. This article proposes a framework for the construction of a tower crane digital twin, and develops the physical-to-virtual connection of the digital twin. The main contributions of this paper include development of tower crane monitoring dataset, tower crane detection and tower crane operation mode recognition. By annotating more than 20,000 tower crane images in 583 tower crane videos, a tower crane image recognition dataset and a tower crane operating mode dataset are established. Yolov5x algorithm is used in the tower crane detection, and the test set detection accuracy is 93.85%. After comparing the LSTM and CNN algorithms, 3DResNet algorithm is selected for tower crane operational mode recognition. The dataset is augmented by rotating the image and the final recognition accuracy reaches 87%. These models can be installed on CCTV to monitor operational status of tower crane in real time and transfer relevant information to the virtual model. The tower crane in the virtual space completes the action of the physical tower crane, thereby realizing the physical-to-virtual mapping in the digital twin.

1. Introduction

The concept of digital twin originated from the military and aerospace industries[1], and has been increasingly used in many other applications, such as in manufacturing. In general, Engineering is the field with the slowest progress in digitalization[2], and only recently digital twin technology has started attracting attention of engineering researchers and practitioners. For instance, there are currently insufficient digital twin applications for construction sites and equipment, e.g., tower cranes, which should have benefit enormously from the digital technology.

Tower crane is one of the widely used construction machinery in construction projects[3]. Operational safety of a tower crane is primarily important for any construction company to reduce losses and even casualties[4]. Thus, it is of vital importance to monitor the working status of a tower crane, feedback and analyses the abnormal situation both physically and digitally to ensure the safe operation of the tower crane equipment. At present, the modeling part of digital twin in an engineering application is often by point cloud, laser scanning, BIM, 3D modelling software like Pro/e and Solidworks, and other methods. These models have high fidelity and restoration degree, but normally take a large amount of time for manual processing and cannot offer real-time responses to changes in physical entities. Construction site is a physical entity where experience changes all the time, hence, it is necessary to reflect the changes in real time in a virtual model. On the other hand, traditional tower crane inspection relies more on sensors and daily manual inspections, which may result in a waste of data or needs to deal with a large amount of multi-source heterogeneous data. To address these issues, this study adopts the method of pre-modeling, by which a tower crane model is developed in the virtual space. Our work addresses this gap by introducing a computer vision-based solution that leverages YOLOv5 for crane detection and 3DResNet for operational mode recognition. This enables real-time monitoring and recognition of crane movements within a digital twin environment, where the physical status of cranes is continuously mapped to a virtual model.

Unlike standard object detection studies, which typically focus on static images, our research emphasizes the recognition of dynamic crane operations, such as rotation and movement states, within a construction site. The digital twin environment further differentiates our approach by ensuring that real-time data from CCTV systems is immediately transferred to a virtual crane

model. This allows for immediate analysis, risk assessment, and operational optimization, making it an essential tool for enhancing construction site safety and efficiency.

The structure of this paper is as follows. The second chapter of this paper review the development of digital twin and its application in engineering. Chapter 3 presents the framework of the digital twin for a tower crane. Chapter 4 explains the development of the tower crane dataset and data augmentation methods. Chapters 5 and 6 introduce the method of tower crane image recognition and tower crane operation mode recognition to connect the physical and the virtual crane. The final chapter discusses the findings, limitations, and future work of this study.

2. State of the Art

2.1 Origin of Digital twin

With the development of IT technology in the 1990s, it was increasingly possible to develop virtual models to generate complex physical artificial products and to integrate simulation systems[5]. At the beginning of the 21st century, virtual models of products began to include the definition of product personality[6]. Modelling has gradually become a common tool to solve some problems that are related to manufacturing and engineering. It can be used to check aspects of functionality or the entire production system. The concept of digital twins has also become more and more specific. The concept of digital twins was first proposed by Grieves as digital representation of actual physical products[7]. A digital twin mainly includes physical entities, virtual models, and the connection of the physical and the virtual parts. It is updated by modelling, simulating, and self-optimising the physical entities[8].

2.2 Application of digital twin in construction engineering

The application of digital twin in the civil engineering industry is still relatively vague. At present, most digital twins related to construction are concentrated in a single life cycle stage[9], for instance, design and engineering phase, construction stage, operation and maintenance stage, respectively.

In the design phase, in order to describe the digital twin of a single building, full element building modeling and simulation technology are required, which is different to the traditional method that is based on the building information modeling[10]. BIM does not pay attention to the relationship between the model and the physical entities, but digital twins require the existence of physical entities. The current digital twins are overly pursuing high fidelity and neglect the requirements of modeling. In this case, Zhang proposed the evolutionary concurrent modeling method[11], which is based on traditional modeling and simulation, and can systematically guide the modeling process of the digital twin. For a completed building, Shanbari[12] proposed lidar modeling, and Zhang[13] proposed point cloud modeling to describe the digital twin of the building. Kaewunruen[14] used Revit and Navisworks software to build the BIM model of a railway transportation system. Combined with the digital twin technology, they can manage the entire lifecycle of the railway capacity system, reduce costs

and increase sustainability. Lu[15] used point clouds and labelled point clusters to model concrete bridges as part of the rapid modeling of digital twins. Angjeliu[16] used the point cloud to model the vault of the Milan Cathedral as the first step of connecting the physical model to the virtual model in the digital twin to prepare for the subsequent structural monitoring, operation and maintenance of the building. However, though these methods, which use point clouds, lidars, or commercial software to build BIM models, are able to provide physical-to-virtual connection of digital twins to a certain extent, these connections are not real-time, and cannot instantaneously reflect the physical entities in the digital twin over a period of time.

2.3 State of art of algorithm research

Object detection, an integral part of computer vision, has been widely used in intelligent video surveillance[17, 18], autonomous vehicle[19], manufacturing inspection[19, 20], as well as other fields. With the increases in GPU power, the iteration of convolutional neural algorithms, artificial intelligence technologies, and mature image training datasets, target detection models based on deep learning have made significant progress in the last few years.

Target detection algorithms are divided mainly into categories, i.e., two-stage detection (the process from coarse to fine) and one-stage detection (completed in one step)[21]. Two stage detection convert the target detection into a classification problem, while one stage detection algorithm converts it into a regression problem. The unified network can be used to directly predict bounding boxes and classification categories. In general, the detection accuracy of one stage detection is lower than that of two-stage detection, but one stage detection has a faster training speed. The you-only-look-once series (YOLO) was first proposed by Redmond for one-stage detection and it is widely used because of its speed[22]. Although YOLOv1 has a fast detection speed, it is not as good as the two-stage detection method in terms of detection accuracy. Liu[23] applied the archor mechanism of Faster RCNN to the Yolo algorithm and proposed the Single Shot Detection (SSD) algorithm to achieve good detection speed along with improved accuracy. However, the detection accuracy for small objects is not ideal. Yolov2[24] added batch normalization after the convolutional layer, anchor boxes and multi-scale training, thus improved further both the training speed and accuracy. Yolov3[25] designed Darknet-53 as the backbone model, which is deeper and more complex, and replaced

the softmax loss with logistic loss. As a result, the accuracy of the model was greatly improved, especially for small target detection, without reducing the processing speed. However, the improved model can only predict two bounding boxes in each image grid, thus the model is not as accurate as the two-stage one. Yolov4 is a real-time and high-precision detection model which meets more field needs. It uses mosaic data argument to expand the dataset and decrease CPU, replaces darknet-53 with CSPDarknet 53, and in the neck layer, uses FPN+FAN to make full use of feature fusion[26]. Yolov5 adopts a more lightweight model, the accuracy of which is close to that of Yolov4, while, theoretically, the detection speed is more than twice as fast as that of Yolov4.

It has been recognized that yolov5 has shown poor detection ability for occluded targets when the weighted non-maximum suppression (NMS) is used to filter out the target frame in the post-processing process of target detection. Under these circumstances, Distance-Intersection of Union_non maximum suppression (DIoU_nms) will be used to detect the occluded objects in this experiment. Traditional nms only considers IoU. When two different objects are very close, the IoU value is relatively large and only one detection frame remains after nms processing. However, DIoU_nms considers the distance between the center points of the two prediction frames. This situation will be considered that this is the frame of two objects and another one will not be filter out. In addition, the mosaic data argumentation method mainly focuses on raw images, i.e, four images are randomly cropped and then spliced into one new image to argument the training dataset. To reduce the influences of noises, edge extraction, which can eliminate image noise, emphasizes the edges and outlines of the detected objects on the image, will be used to simplify image information and use edge lines to represent the information carried by the image.

3. Framework of digital twin of tower crane

To develop a digital twin, the first step is to define the physical entity and the physical environment in the virtual model through modelling and other methods. At the same time, the impact of the physical environment on the physical entity to be reflected in the virtual model. Then, by simulating and analyzing the changes of the virtual model in the virtual environment to predict the future condition of the former. The information will be transmitted to the physical entity so that decision-makers can be presented with different options. It is expected that digital twins should facilitate real-time data transmission and autonomous analysis and decision-making. Given the above framework and the requirements,
digital twin should have following seven important features i.e., the physical entity and the physical environment; virtual entities and the virtual environment; the physical to virtual connection; the virtual to physical connection; the digital twin process; the real-time nature of the digital twin; and autonomy of the digital twins. This article trained a tower crane detection algorithm and use this algorithm to detect and segment tower cranes in the videos, thus facilitate the recognition of the operation mode. This information will be transmitted to virtual model through the physical to virtual connection. Therefore, the detection of true tower crane is an important step to realize the digital twin system of tower crane.

Figure 1 is a simple illustration of digital twin of a construction crane, by which the relevant data of the physical entity and environment are transferred to the virtual environment through the real-time and autonomous physical to virtual connections, while the virtual entity analyses the data and then transfer execution commands to the physical entity through the virtual to physical connections.

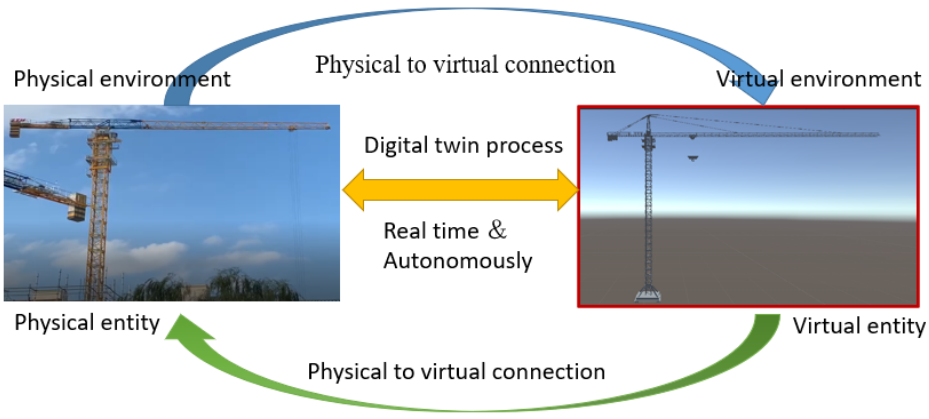

Figure 1. Core concepts of Digital twin

3.1 Physical entity and physical environment

In this research, the physical entity refers to the tower crane. The QTZ80 tower crane, shown in Fig.2, was selected in the case. Other types of tower crane will be modeled separately. The physical environment of the digital twin is the space where the tower crane is located, which includes the external environment and physical processes. When the crane is operating, the physical process is the way the crane behaves, including, e.g., direction and speed of rotation, and operation status. In a model, it is required to quantitatively measure the physical process of the crane. Simultaneously, various external environment factors that may affect the operation of the tower crane are needed to be measured, and as inputs of the virtual twin environment as they change.

Figure 2. Physical entity of the QTZ80 tower crane

3.2 Virtual entities and virtual environment

The virtual entity is the twin representation of the tower crane in the virtual space. According to the CAD drawings of the QTZ80 tower crane, Creo (Pro/E) software is used to draw the components of the tower crane and then assemble them in the 3D unity software. As is shown in Figure 2, the boom and the main body of the tower crane are defined, respectively, as a virtual environment existing in the digital domain as the mirror image of the physical environment, which contains the digital representation of some external sensor data (temperature, humidity, wind speed etc.,)

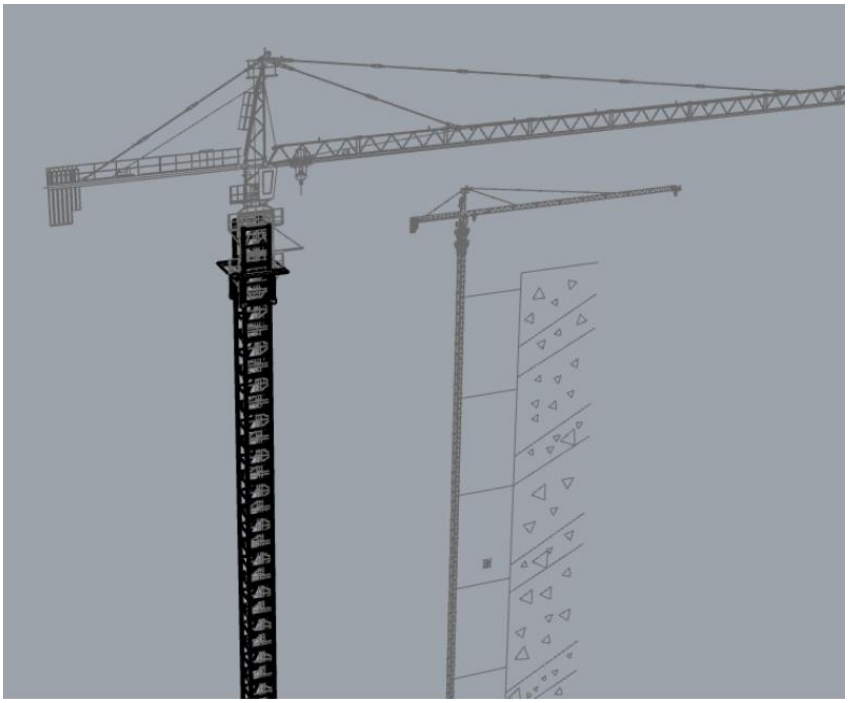

Figure 3. Tower crane boom and main body

3.3 Physical to virtual connection

Digital twins instantaneously reflect changes in the state of physical entities themselves and external physical environment to the virtual entity through the physical to the virtual connection. At the level of technical application, the parameters of the changes in the physical crane can be updated through sensors[7] [27, 28], 5G[29], IOT[10, 30, 31], CPS[32]. In this research, tower crane detection, operational mode recognition and modelling are used to capture the changes of tower crane and update the virtual crane accordingly.

3.4 Virtual to physical connection

A digital twin does not work with only physical-to-virtual connection. Information, such as abnormal operation, etc -should be fed back to the physical entity (management personnel) through the virtual-to-physical connection for optimization[33-35]or to provide decision-making opinions[36-38].

3.5 The digital twin process

The digital twin process is to reflect the detailed information of physical tower crane and its environment to the virtual world using digitalization method, and then simulate the virtual

tower crane to transfer useful information to improve the physical entity. Digital twin of tower crane with physical-to-virtual connections and virtual-to-physical connections can realize the closed-loop “simulation-execution-adjustment” function of the tower crane digital twin and enable the digital twin to continuously update through self-learning.

3.6 The real-time nature of the digital twin

The close-loop connection of the physical and virtual crane will generate a large amount of multi-source heterogeneous data (temperature and humidity sensor data, tower crane operation video and image data and etc.,), these data have many types and fast generation speed. It is necessary to establish a big data storage management platform and ensure the security of data through blockchain technology to support real-time interaction of tower crane digital twin. Digital twin and big-data driven application platform are needed, through the latest technologies to scientifically manage the operation safety of tower cranes, improve the efficiency of tower crane hoisting and distribution, and establish a safety management platform, which can online/offline detect operation conditions, work time.

3.7 The autonomous nature of digital twin

The tower crane digital twin takes the experience of tower crane construction personnel, construction knowledge, historical operation and maintenance data, and real-time data as input to output prediction data, enriches and updates the feature database for different safety problems, and finally forms autonomous intelligent diagnosis and determination and feedback to site managers. At the same time, use big data technology to collect information from sensors, such as tower crane operation status, real-time data on the health status of tower crane components, and digital-related historical data (maintenance records, energy consumption record data), etc. Through the Bayesian cycle, the predicted data and the actual data are compared and analyzed, and the optimization model is continuously learned to realize the autonomous digital twin of the tower crane.

4. Creation of image dataset

This research mainly focuses on realizing the physical-to-virtual connection in the digital twin concept. In order to achieve this step, the pre-modeling method is used, by identifying the state changes of the physical entity of the tower crane, and then reflecting these changes in the model. As shown in Figure 4, this research first uses the object detection algorithm to train the tower crane image recognition model, then uses the best model to process new tower crane videos to detect tower cranes, and the tower crane segmentation algorithm to segment the tower cranes, and finally performs tower crane operational mode recognition. This requires tower crane object detection dataset and tower crane operation mode recognition dataset.

Data is the most critical part of machine learning that is incorporated in the software used for the research. Using high-quality, large-scale image datasets can maximize the efficiency of deep learning, thereby training models with higher quality and accuracy. However, currently, there is no publicly available general-purpose tower crane image dataset with category labels for tower cranes on construction sites. Collecting relevant video and image data and performing a series of optimization processes is a primary task of this research.

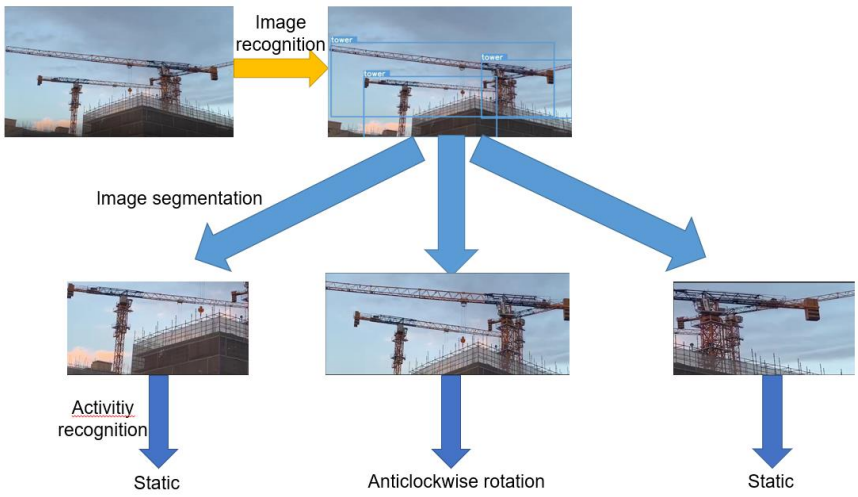

Figure 4. Framework of recognition of tower crane operation process

In this research, the creation and optimization of tower crane image datasets are conducted in the following 5 steps, i.e., (a) video data acquisition to create datasets with a small number and medium quality; (b) data pre-processing through grayscale processing, binarization processing or edge processing to reduce the number of data with high-quality datasets, (c)

image annotation, using Labellmg for tower crane 2D bounding box annotation to create a labelled dataset; (d) tower crane segmentation to connect image recognition deep learning and motion state recognition deep learning; and (e) data augmentation by rotating images, mosaic enhancement and other methods to form larger, high-quality datasets that can be used for deep learning. As a result, the initial data set with a small number and medium quality is turned into a training data set with a larger number and high quality.

4.1 Data collection and pre-procession

The tower crane object detection dataset includes a video set, an image set and an annotation set. The video mainly comes from the combination of Google videos and the tower crane operation video shot on-site. After filtering out some blurred and poor-quality videos, a total of 583 videos with a length of approximately 7-15 seconds have been collected. The videos only contain images of the tower crane in static, clockwise or anti-clockwise rotation. The number of the annotated tower cranes exceeds 20,000.

In the dataset of tower crane operation mode recognition, this research separates the tower cranes in the image through the tower crane segmentation algorithm. For each group of tower crane images, 20 images at a certain interval of frames are saved in different folders. It is required to indicate the motion state of the tower crane captured by the images in each fold, i.e., 0 for static, 1 for clockwise rotation and 2 for anti-clockwise rotation. For example, tower 2 in test 3 rotates clockwise, and this is marked as test 3/tower 2_1. After these processes, 1,373 sets of data were obtained, including 27,460 tower crane pictures, which is divided into 1,167 sets of training set data, 119 sets of dev dataset and 87 sets of test dataset.

Finally, this research obtained a detailed tower crane dataset. Including a video set of 583 tower crane operation videos, an image set of 27460 tower crane pictures, a pre-processed tower crane operation mode recognition dataset with 1373 sets of image data.

4.2 Image annotation

After completing image acquisition and pre-processing, this study uses Labellmg to annotate tower crane images for subsequent deep learning of image recognition. The images of the annotated tower cranes ate shown in Figure 5, where each of the tower cranes is closely

enclosed by a rectangular frame, which covers all parts of the tower crane. Normally, disturbing noises, e.g., tree branches in Figure 5, may also be included.

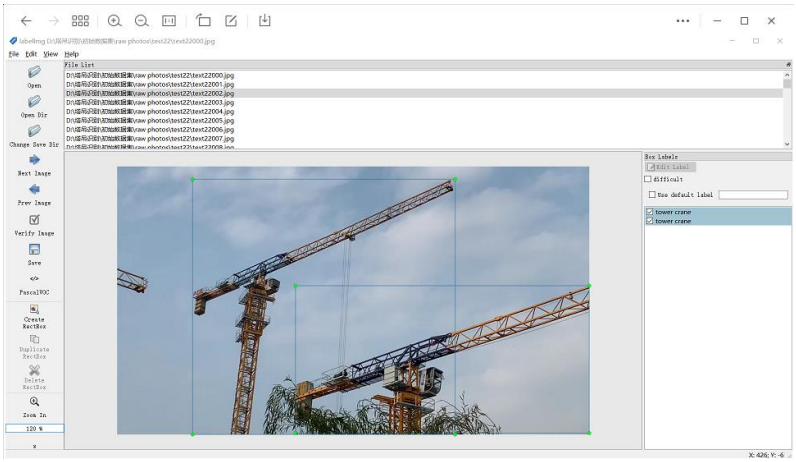

Figure 5：Sample of tower crane annotation

4.3 Data argument

After completing the process of image data collection, pre-processing and labelling, this research adopts several methods of data enhancement and augmentation to expand the tower crane image dataset. Among them, mosaic enhancement is used in the object detection algorithm. Data augmentation is performed on the tower crane image dataset through mosaic data augmentation and traditional single-sample data augmentation methods (flip, rotate, crop, scale, color transform, etc.) After completing data enhancement, the original data and the enhanced data are scrambled and mixed into a new dataset, the volume of the dataset can be increased and the new dataset provides more data to support the parameter optimization in the model, thereby improving the generalization ability of deep learning models.

In the recognition of the tower crane operation model, the data augmentation method of rotating images is used. In the model training of the tower crane operation mode recognition, to increase the training dataset, it is vital to make the dataset as diverse as possible so that the trained model has a stronger generalization ability. Improving the relevant data in the dataset through data augmentation can prevent the network from learning irrelevant features, learn more data-related performance and significantly improve the overall performance. In this research, the tower crane was rotated 10º and 20º clockwise and anti-clockwise. As shown in

the figure below, the tower crane is rotated by 10º, -10º, and the original image of the tower crane is 20º and -20º. Through the above operations, the number of datasets increased. from 1,373 to 6,865, which is five times larger, and improves the overall performance of the subsequent model.

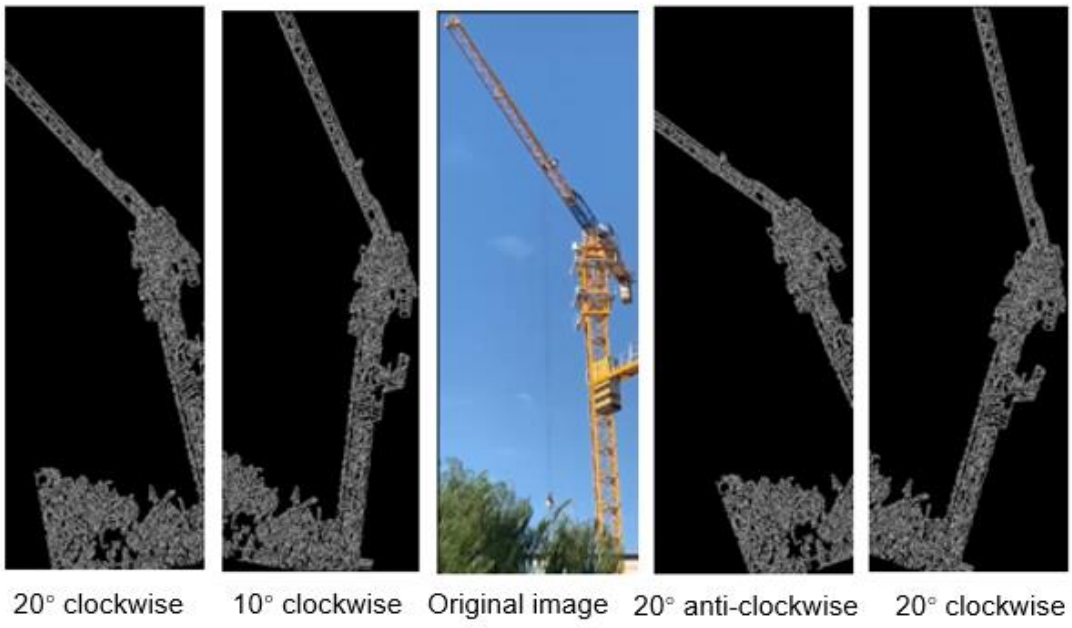

Figure 6. Augmented dataset

5. Tower crane detection based on yolov5

After completing the creation and optimization of the tower crane image dataset, it is necessary to select an appropriate deep neural network model to train and test the dataset images, and continuously adjust the model parameters to obtain a model with low loss and high recognition accuracy. Different types of neural networks and deep learning algorithms have significant differences in recognition accuracy and detection speed, and should be selected according to the requirements of an application. In order to realize the real-time and autonomous nature of the digital twin of tower crane, tower crane object detection algorithm should have the ability to detect the tower crane in a quick response, yolov5 algorithm is a good choice for its lightweight model, high precision and it can be built in a cctv monitor. High-quality and large-scale datasets can effectively improve the accuracy of training models, and large volume datasets also play a positive role in the development of algorithms. Based on the tower crane construction video images collected in this study, Yolov5 passes each batch of training data through the data loader and simultaneously enhances the training data through scaling, color-space adjustment, and mosaic enhancement. In the development of the digital twin of the tower crane, image recognition of the tower crane is the first step from the physical to the virtual end.

5.1 Yolov5 structure of tower crane detection

Yolov5 is the latest version of yolo series algorithms after yolov4, and it has the advantage of fast calculation speed, and smaller model volume. Yolov5s network structure consists of four parts: input, backbone, neck, and prediction. Figure 7 shows the structure of Yolo series algorithm.

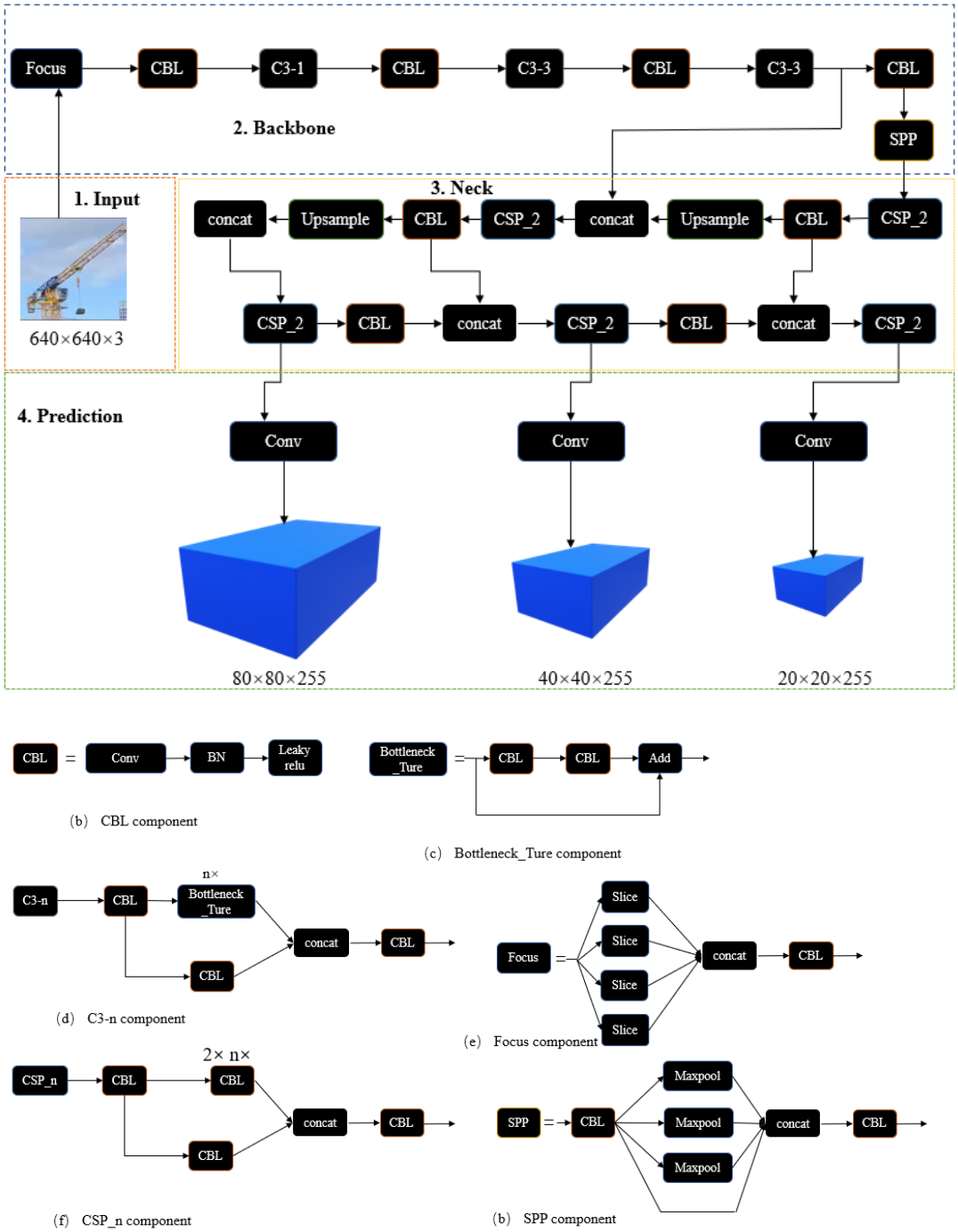

Figure 7: Network structure of Yolo series algorithm

The input component uses mosaic data enhancement, adaptive anchor box calculation and adaptive image scaling. Mosaic data enhancement uses traditional data enhancement methods to process the selected four initial images and combine these four images to form a new image. Adaptive anchor box enhancement uses initial anchor box function to adaptively calculate the best anchor box values for different training datasets. Adaptive image scaling method unify

different input sizes to standard, thereby improving the training and inference speed.

Backbone structure is combined by CPS component and focus component. The CSP structure divides the feature map of the base layer into two parts, and then merges them to achieve rich gradient combinations and solve the problem of heavy computation in previous work. Yolov5 algorithm adds a slice operation in the focus structure, which can double the sampling feature map without losing any information.

In order to better extract the fusion features of the target, Yolov5 inserts the neck layer containing the Feature Pyramid Network(FPN)+Path Aggregation Network(PAN) structure in the backbone layer and the output layer. Figure 8 demonstrates the FPN+PAN structure. Among them, the FPN layer conveys strong semantic features from top to bottom through up-sampling, and the PAN layer conveys strong localization features from bottom to top through subsampling.

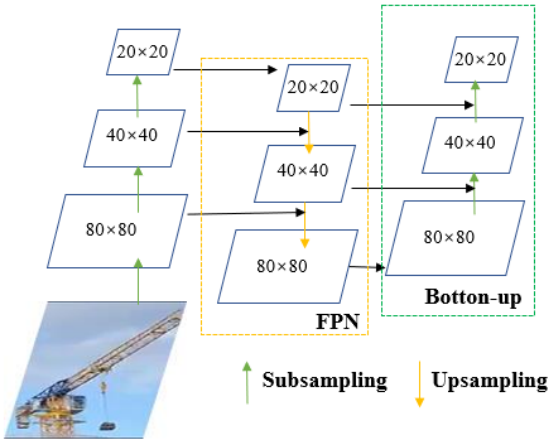

Figure 8：FPN+PAN component

5.2 Experimental environment

5.2.1 Evaluation indicators

Numerous evaluation indicators have been devised by scholars for quantitative analysis of target detection algorithm performance. They all indicate the performance of the detection algorithm to its level to a certain extent. General precision indicators are precision, recall, accuracy and mean average precision (mAP). In the target detection based on the YOLOv5

algorithm series, evaluation indicators such as precision, map, recall and F1 score are introduced to evaluate the training results algorithm accuracy.

Equations (1-3) below are the definition of precision (P), recall (R) and accuracy (A) using TP, FP, TN, FN.

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

$$A = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

where, TP is true positive, FP is false positive, FN is false negative, TN is true negative. Take the tower crane detection as an example, the part detected as the tower crane is *TP*. Other parts of the picture detected as tower cranes are termed *FP*. When the tower crane goes undetected *FN* is indicated. The number of tower cranes images detected is represented by *TP*+*FP*, and the actual number of tower cranes is shown as *TP*+*FN*.

The F1 score, which is the harmonic-mean of precision(P) and recall(R) used in machine learning, is calculated by Eq.(4) . It is useful for evaluating models, especially when dealing with imbalanced dataset.

$$F_1 = \left(\frac{2}{R^{-1}+P^{-1}}\right) = 2 \cdot \frac{P \times R}{P+R} \tag{4}$$

5.2.2 Experiment results

In this study, two groups of experiments are used to compare and test the superiority of the improved yolov5 target detection algorithm. 8,746 annotated tower crane image data are used, with a total of more than 20,000 annotated tower crane data, for algorithm training. In the first set of experiments, 3,265 new tower crane images are used as samples, four yolov5 algorithms with different depths, yolov5l, yolov5m, yolov5s, and yolov5x, were compared, and the advanced and superiority of the improved model in tower crane target detection was analyzed. The second group of experiments selected the best yolov5 algorithm, compared it with the improved yolov5 algorithm with different improvement strategies, and analyzed the impact of the improved strategy on the model performance through ablation experiments. The

Optimizer weight decay is set to 0.0005, and the initial learning rate is 0.01. Among them, for yolov5s and yolov5m, the number of iterations is 300, and the batch size is 32. The number of iterations for Yolov5l is 400, and the batch size is set to 16. Yolov5 has the deepest number of layers and is limited to 24g of video memory of the computer graphics card, so the batch size is set to 8 and the number of iterations is 500 to improve the detection accuracy.

The experimental hardware of this research is introduced as follows: The CPU is Intel(R) Core (TM) i9-11900F@2.50GHz. Memory is 64.0GB; NVIDIA GeForce GTX 3090 24G graphics card. Python 3.7 is used as the programming language, TensorFlow-gpu is used as the deep learning framework, Cuda 10.2 is used for GPU acceleration, and OpenCV4.0 is used for image preprocessing.

This study adopts the idea of transfer learning and uses the pre-trained yolov5. yaml model to improve the convergence speed of the yolov5 target recognition algorithm. The training process adopts the approximate joint method for training. According to the depth of the model, different learning rates and iterations are selected. Since the model depth of yolov5s is the lowest, the number of iterations selected in this study is also relatively small. Considering the large size of the image dataset, the learning rate is also selected to increase the speed of model training.

Method	Precision	Recall	F1 score
Yolov5s	89.64%	98.97%	0.941
Yolov5m	91.68%	99.27%	0.953
Yolov5l	93.01%	97%	0.959
Yolov5x	93.85%	99.12%	0.964

Table 1. Comparison of Resnet algorithm with different depth

From the above table, we can see that the recall values of the four Yolov5 algorithms with different depths are also high, all around 99%. All of them achieved an F-score above 94% (94.1%, 95.3%, 95.9%, 96.4% respectively).

Yolov5x is the deepest in the Yolov5 series algorithms. Due to the large number of convolution kernels, the detection speed of Yolov5x is only 1/3 of that of Yolov5s. Therefore, this part

selects the Yolov5x deep learning target detection model with the highest accuracy as an example. Finally, the training and validation set accuracy of Yolov5x is 93.85%, and the recall is 0.9912. Figure 9 shows the precision, recall, and map curves of the model trained by the Yolov5x algorithm. After 400 epochs, the precision curve tends to be fitting which demonstrate that the training epoch is enough. Figure 10 demonstrates the PR Curve of the Yolov5x algorithm, the PR fitting curve is approaching the upper right corner which shows that the model has an excellent ability to detect tower cranes.

The current research requires a strategy that combines real-time and offline tower crane detection. Therefore, Yolov5s can better meet real-time requirements, consume less computation and be of greater convenience in its ability to deploy on mobile terminals and edge terminals, which is conducive to the landing deployment of products. On the CCTV detection side, we use the built-in Yolov5s algorithm for work deployment. In offline detection, as the requirements for detection speed are less restrictive, we can use the Yolov5x target detection model to detect tower cranes with a higher accuracy.

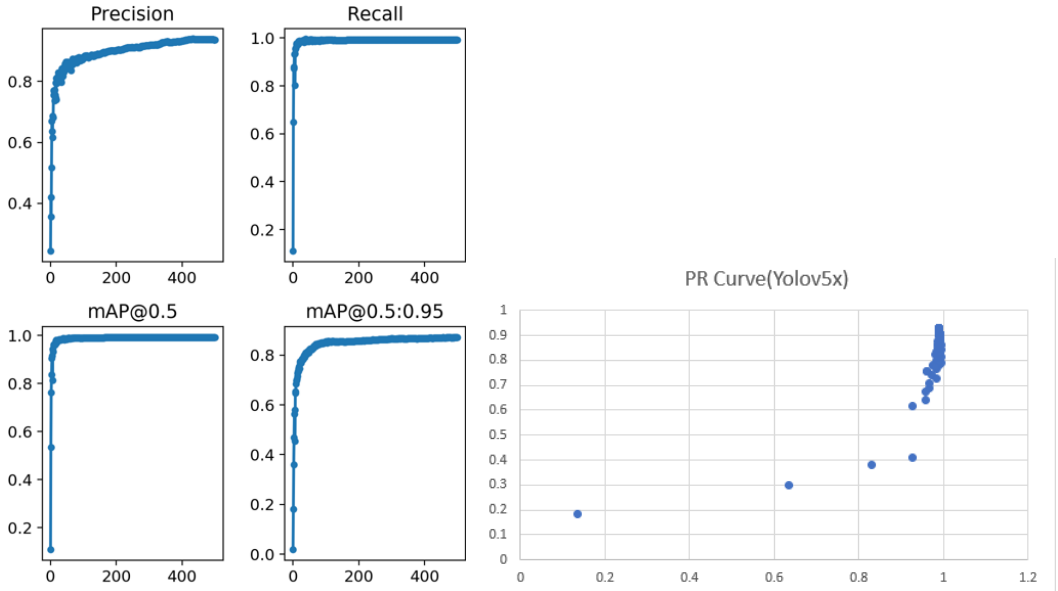

Figure 9: Common evaluation index curve　Figure 10: PR curve（yolov5x）

5.3 Optimization of algorithm

5.3.1 Distance-intersection-over-union (DIoU)_non-maximum suppression (nms) loss function

In the prediction stage of object detection, many candidate anchor boxes will appear near the real box. Some of these anchor boxes overlap and often surround the same target. According to the definition of non-maximum suppression (nms), the nms algorithm is generally used in the post-processing step of target detection to eliminate redundant detection frames. By using nms, similar bounding boxes near the recognized object are merged, and the best bounding box is reserved according to pre-set conditions.

Only the IoU factor needs to be considered in the traditional non-maximum suppression algorithm. Determining the highest-scoring detection frame together with other frames will allow the elimination of all those prediction frames above the nms threshold. In actual situation, following nms processing, there is the possibility of detection failure when two objects are close to one another and only one detection frame remains. Thus, the Distance-Intersection over Union (DIoU_nms) method, which using DIoU as the standard for nms, considers both the overlapping area and the distance between the center point. DIoU_nms is used to decide whether or not to delete a frame by measuring the distance ratio of the two prediction frames.

5.3.2 Edge extraction

Where the local area brightness is considerably different, this part is termed the edge of the image. Where the grayscale changes noticeably to another grayscale value with a major level difference from a buffer area with a small gray value, this can be considered a step change in the gray level profile of this area. Segmentation of the image can be carried out using this feature.

In this research, sobel filter (sobel operator) is used to conduct edge detection. It is a discrete differentiation operator that creates images emphasizing edges. This operator combines Gaussian smoothing, 2-D convolution operator, and differential derivation to calculate the approximate values of the brightness and the darkness of an image. It provides greater accuracy on edge direction information and has a noise smoothing effect. Generally, the Sobel operator tends to be used as an edge detection method when not very high accuracy is the requirement. Figure 11 below shows the tower crane image processed by sobel operator.

Figure 11: Tower crane images using Sobel operator

5.4 Ablation experiment

In order to better analyze and verify the effectiveness of the Yolov5 improvement strategy used in this chapter, this study designed a series of ablation experiments to compare the impact of different improvement strategies on the final tower crane detection results. The target detection results under different improvement strategies are as follows:

	Model 1	Model 2	Model 3	Model 4
DIoU_nms	×	√	×	√
Edge extraction	×	×	√	√
Precision	93.85%	94.23%	95.12%	95.45%
Recall	99.12%	99.25%	99.36%	99.41%
F-score	0.964	0.967	0.972	0.974

Table 2: Detection results of Yolov5 under different improvement strategies

Model 1 represents the original yolov5x model, model 2 represents an improved model that only modifies the weighted nms loss function to DIoU_nms, model 3 represents a model that only performs edge extraction on images, and model 4 represents the improved yolov5 model proposed in this section.

From the detection results of Model 1, it can be seen that the original yolov5 has an accuracy of 93.85% for object detection of tower crane images, the recall is 99.12%, and F1 score is 0.964. In Model 4, the above three indicators are 95.45%, 99.41% and 0.974 respectively, the

improved model proposed in this study can effectively improve the tower crane detection ability of Yolov5 on the tower crane image dataset. Comparing model 1 and model 2, it can be found that after modifying the weighted nms loss function to DIoU_nms, the precision and recall of the model are improved respectively, which shows that the target detection effect can be improved by using DIoU_nms when the training samples and training methods are the same. Comparing model 1 and model 3, the precision has been greatly improved by 1.35%. This shows that edge extraction can obviously reduce the noise of the image, thus allows the algorithm to better identify the tower crane.

Through comparative analysis, it can be found that the yolov5 improvement strategies used in this study can effectively improve the detection accuracy of the model, and at the same time, the comprehensive improvement strategy can also improve the single improvement strategy.

5.5 Tower crane segmentation

Tower crane image recognition is the first and one of the most important steps in tower crane operation mode recognition using real-time video footage. Once a tower crane image recognition process is completed, the best model will have been obtained (the model with the highest accuracy) and trained by the algorithm. The best model (best.pt) trained by the yolov5x algorithm was selected and used in this study and a set of segmentation tower crane algorithms was designed to segment tower cranes in the videos. First, the algorithm splits the video into a series of image frames. Next, tower cranes are identified in the image frames, and these identified tower cranes segmented individually, as shown in Figure 12 below. In this video, each image frame contains three tower cranes, and each tower crane is divided and stored separately as tower1, tower2, and tower3. The resulting hundreds of independent tower crane pictures are prepared consecutively for motion state recognition.

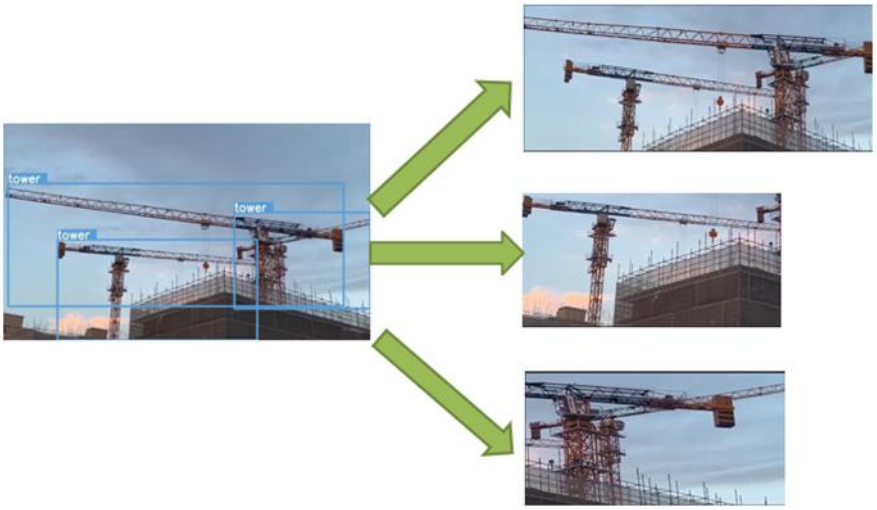

Figure 12: Tower crane segment

However, there are some cases where two tower cranes are in the same image frame, as shown in Figure 13. Often, in such situations, one tower crane is rotating and the other is stationary. If we perform pattern recognition on these image frames, the algorithm often cannot accurately determine the operation mode of the tower crane, resulting in ambiguous responses.

Figure 13: Example of two tower cranes overlapped in one image

Normally, for each group of tower crane datasets, overlaps or duplications will have to be manually filtered out. In this study, however, some samples with two or more tower cranes are retained, as this will increase the robustness of the model. In the video dataset, these videos are usually 5 to 20 seconds and contain hundreds of image frame. In order to create a tower crane operation mode recognition dataset, 20 frames are selected from one video. Tower crane operation mode recognition algorithm trained the model using these 20 frames of images.

6. Recognition of motion mode of tower crane based on 3DResNet

6.1 Selected algorithms

The commonly used motion recognition algorithms include optical flow method, motion recognition classifier, convolutional neural network (CNN) and long-short term memory (LSTM), etc., These algorithms face many issues in dealing with changes in the actual environment, including, failure in obtaining high-quality background model locks during background initialization training, dynamic background shaking (e.g., motion of tree leaves), camera shake, etc. In this study, some of the commonly used operation mode recognition algorithms are tested, we used the combination of LSTM and CNN, 2DResNet and 3DResNet to find the best algorithm with the highest accuracy.

6.1.1 LSTM and CNN

There is a greater information transmission band of cell state in LSTM, relative to the accepted recursion neural network, as this algorithm has increased information memory. The LSTM has four fundamental stages. First, the forget gate discards some earlier information; second, some present information is retained by the input gate; third, past and present memory is melded, and finally, information is outputted by the output gate. The continuous or fixed frame interval images of the tower crane operation mode are recognized in this paper. It is likely that LSTM, where memorizing the past and selecting information, is applicable for deep learning problems with time series.

Convolutional neural network is a commonly used algorithm which uses convolution to simulate the way that human visual system works. Many studies use the method of combining LSTM and CNN to perform classification tasks, time series prediction tasks, etc. Under this circumstance, it is possible to use this method to recognize the operation mode of tower crane, and two different combination method is shown below.

(1) Use CNN as the input of LSTM: First use CNN to extract the local feature of the tower crane operation images, then using LSTM to extract the long-distance feature of these local features, finally, transform these features and input into the fully connected layer to classify the summarized feature.
(2) Use LSTM as the input of CNN: First use LSTM to extract the long-distance features

of tower crane operation images to obtain time series information before and after the fusion, then using CNN to extract the local features, finally input into the fully connected layer after transformation.

6.1.2 Residual network

The Residual Network (ResNet) method was proposed for vanishing gradient problems when using a deepened network[39]. This network is also the first network with a depth of 100-layers. The structure of 2DResNet network used in this section can be seen in Figure 14 (34 layers). One normal convolutional and a max-pooling layer form the first construction layer. The six residual modules make up the second construction layer. The subsequent construction layers of seven, eleven and five residual modules comprise the third, fourth and fifth construction layers, each starting with a down-sampling residual module.

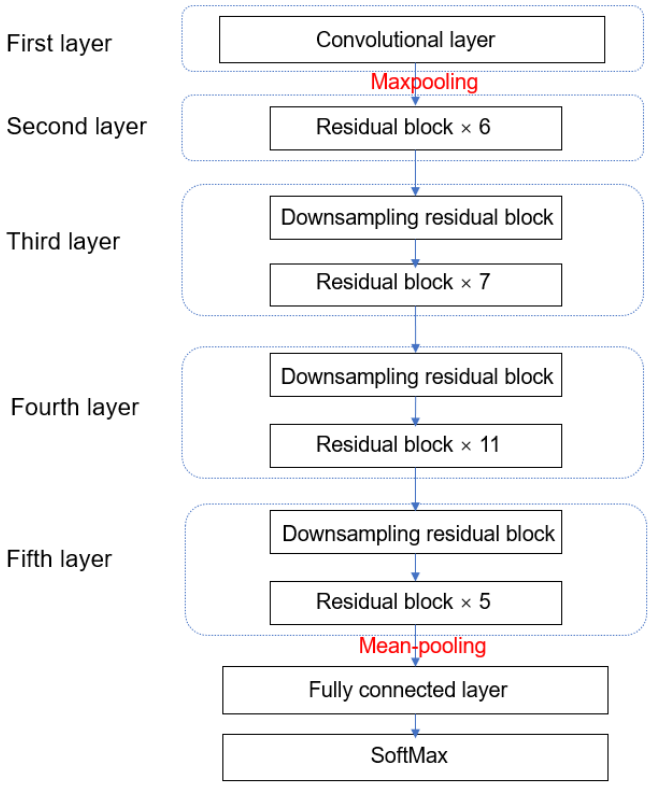

Figure 14: Network structure of ResNet34

6.1.3 3DResNet

The training results using 2DResNet may be overfitting as there are limited number and types of tower cranes. There can be an exponential increase in the 3D convolution kernel network

parameters in comparison to the 2D convolution kernel. Spatiotemporal features can be extracted directly from videos for action recognition using 3D convolution. Each 3DConv module can be connected by using the residual structure of 3DResNet (= ResNet + 3DResConv). The convolution kernel of 3DResNet increases one dimension from 2DResNet by adding an additional parameters T (in channels, out channels, T), which effectively extracts temporal information of the input and features of the diagram by considering timing. From 2DConv to 3DConv, the major difference is that inputs and features have become temporal, adding one dimension. Thus, the time-series information in the feature map can be successfully extracted, following the time-series convolution, thereby allowing the network to extrapolate improved video inputs.

6.2 Training process

The training process involved using a dataset of annotated tower crane images. The dataset was augmented to increase its size and variability, as crane motion can appear in different directions and under different environmental conditions. The following steps were taken during training:

6.2.1 Data Preparation:

- Annotation: Each frame of the video was labeled with one of the three operational modes (static, clockwise, or anticlockwise).
- Data Augmentation: Images were rotated by 10º and 20º to simulate additional motion states. This increased the size of the dataset and made the model more robust to variations in crane operation.

6.2.2 Model Training:

The model was trained using stochastic gradient descent (SGD) as the optimizer, with a learning rate of 0.002. Training was conducted over 200-500 epochs, with the model evaluated on a validation set at each epoch. The cross-entropy loss function was used to optimize the model's classification accuracy.

6.2.3 Evaluation Metrics:

Performance was measured using the Accuracy, Precision and Recall defined in Equation (5).

6.3 Comparison of different algorithms

From the conclusion of section 6.1, we chose the combination of LSTM and CNN, 2DResNet and 3DResNet to find the best algorithm with the highest accuracy. The original dataset with 1373 sets of tower crane operation images was chosen, 70%, 20% and 10% of them are divided into training dataset, test dataset and dev dataset respectively.

Table 3 shows that when using the 2DResnet or CNN+LSTM algorithms to train the tower crane operational image dataset, the accuracy was 40% and 48% respectively (i.e., less than 50%). These algorithms did not, therefore, learn the operational characteristics of the tower crane. The 2DResNet focuses on spatial features, which limited its ability to capture temporal information that is critical in recognizing the crane's operational modes over time.

The training detection accuracy of Model 2 with LSTM+CNN was 57%, indicating that this algorithm learned some of the characteristics of the tower crane's operations, but the error detection and missed detection were quite serious. LSTM (Long Short-Term Memory) is effective for learning time-series data, while CNN (Convolutional Neural Network) excels at recognizing spatial features. When combined, this model provided a moderate accuracy of 57% in our tests. The sequential nature of LSTM was helpful in capturing time-dependent changes in crane operations. However, it struggled with accurately learning complex motion patterns due to the limited dataset size and variability.

Finally, the accuracy rate (75%) of Model 4 (which used 3DResNet) was the highest. The precision of 3DResnet was 1.32 times higher than LSTM+CNN and was therefore selected for further training. The 3DResNet model incorporated an additional temporal dimension in its convolutional layers, allowing it to capture spatiotemporal features directly from video data. This capability made it the most suitable model for tower crane operational mode recognition. After dataset augmentation, the 3DResNet achieved an accuracy of 87%, outperforming both the LSTM+CNN and 2DResNet models. The residual structure of the 3DResNet allowed it to effectively process the time-series data required for recognizing motion patterns such as static, clockwise rotation, and anticlockwise rotation, ensuring that both short-term and long-term operational modes are recognized accurately.

Model	Method	Precision	Recall	Accuracy
Model 1	ResNet34	0.35	0.50	0.40
Model 2	LSTM+CNN	0.54	0.65	0.57
Model 3	CNN+LSTM	0.45	0.55	0.48
Model 4	3DResNet 34	0.72	0.80	0.75

Table 3. Comparison of the candidate algorithm

6.4 Comparison of Resnet algorithm with different depth

In the previous section, we tried different algorithms and compared their accuracy. We decided that among others, 3DRseNet is the best for tower crane mode operation detection. The accuracy of 3DResNet increases as the depth increases, considering that the dataset in this study is relatively small, using a model with more than 152 layers may cause overfitting. In addition to the experiments outlined above, we compared the accuracy of 3DResnet algorithms with different depths, i.e., 3DResNet18, 3DResNet34, 3DResNet50, 3DResNet101. 3DResNet 152 is not chosen as it is too deep, such that the existing datasets may lead to overfitting. Section 4.3 explained how the present study augmented the dataset through edge extraction with the Sobel operator and image rotation, which increased the 1,373 sets of tower crane operational datasets to 6,865 sets. The datasets were used separately in training the model to select the most appropriate depth for optimization. In this section, 3DResNet 50 and 3DResNet 101 are used to train the augmented datasets. Table 4 shows the accuracy and training time (second) of the 3DResNet algorithms with different depths. In these training process, the learning rate is 0.002, the batch size of 3DResNet18 is 64 and the batch size of others is 16. From the table we can see that when the network become deeper, the training time is increasing. What's more, when training with the augmented dataset, it also takes longer training time.

Method	Precision	Recall	Accuracy	Training time
3DResNet18	0.62	0.66	0.64	192
3DResNet34	0.72	0.80	0.75	158
3DResNet50	0.80	0.83	0.82	280

3DResNet50 (augmented)	0.84	0.88	0.87	1024
3DResNet101	0.84	0.88	0.86	293
3DResNet101(augmented)	0.86	0.91	0.87	1117

Table 4. Comparison of ResNet algorithm with different depth

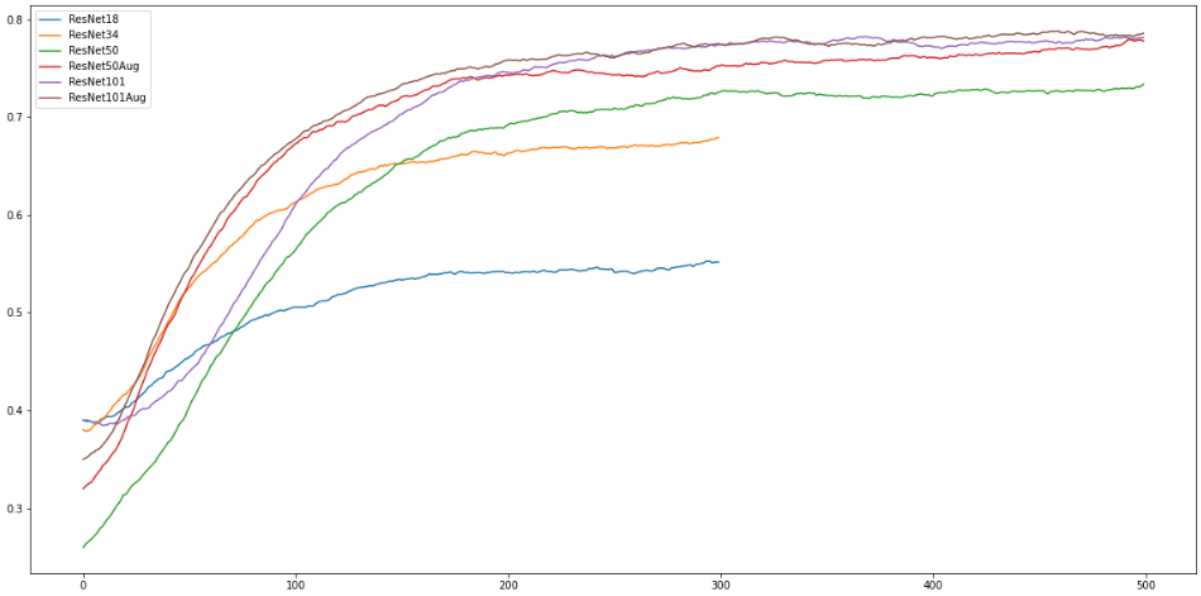

Figure 15. Accuracy curve of different depth of ResNet

Figure 15 shows the fit curve of the accuracy of the above algorithms of different depths. In general, the accuracy of training is fluctuating as the network goes deeper. Here we choose the highest accuracy of a single training as the final accuracy of the model. There are two reasons for the fluctuations: First, the Batch size is limited by the video memory of the computer graphics card. Secondly, the learning rate is chosen to be relatively high in order to reduce model training time. Table 4 shows the achieved accuracy of the ResNet algorithms and their respective depth. From the above accuracy analysis, it can be seen that a deeper 3DResNet structure always results in a higher accuracy, i.e., the recognition accuracy of 3DResnet18 is only 0.64, while the accuracy of 3DResNet 101 reaches 0.86. Table 4 also shows that the recognition accuracy of 3DResNet50 and of 3DResNet101 are increased from 82% to 87% and from 86% to 87%, respectively after using the augmented datasets.

7. Conclusion

This paper proposed a framework for the construction of tower crane digital twin and realized the physical to virtual connection part by creating tower crane datasets, tower crane object detection and tower crane operation mode recognition. It was proposed that tower crane datasets should be generated and optimized using image preprocessing, image annotation, and image dataset augmentation. In addition, an algorithm for tower crane segmentation, combined with a previous object detection model of the highest precision, was developed to separate the tower crane from the image. An improved yolov5 algorithm was proposed for tower crane object detection. DIoU_nms was used to improve the prediction accuracy of the yolov5 algorithm in situations where tower cranes overlap. The concept of edge extraction, which was used to reduce the noise in the tower crane image in the present study, was introduced. Lastly, an ablation experiment was designed to judge the superiority of the two improved methods. The final accuracy rate was improved from 93.85% to 95.45%. In the tower crane operation mode recognition part, it was found that 3DResNet was best placed in identifying the motion state of tower cranes on sorted operating images. After data augmentation process, the accuracy is up to 87%.

This study was confined to the application of digital twin physical-to-virtual connections. Future research is required to add virtual-to-physical connections to the current model and convey the simulation results to management to form a closed digital twin loop. The scope of the modelling should also be extended to include modelling of any experiments involving other tower cranes, thus building a lifecycle digital twin of the construction site as a whole. To improve the accuracy of the current model further, Additional data, such as altitude recognitions and more tower cranes, and modified algorithms are needed. The data collected by the computer vision method was limited. From the perspective of tower crane safety, sensors can be used to capture some abnormal operational status in real-time. Therefore, in future studies, sensors could be added to monitor quantitatively safe operation of tower cranes.

Reference

[1] F. Tao, H. Zhan, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art," (in English), *IEEE Trans. Ind. Inform.,* Article vol. 15, no. 4, pp. 2405-2415, Apr 2019, doi: 10.1109/tii.2018.2873186.

[2] P. Leviakangas, S. M. Paik, and S. Moon, "Keeping up with the pace of digitization: The case of the Australian construction industry," (in English), *TECHNOLOGY IN SOCIETY,* vol. 50, pp. 33-43, AUG 2017, doi: 10.1016/j.techsoc.2017.04.003.

[3] G. Raviv, B. Fishbain, and A. Shapira, "Analyzing risk factors in crane-related near-miss and accident reports," (in English), *SAFETY SCIENCE,* vol. 91, pp. 192-205, JAN 2017, doi: 10.1016/j.ssci.2016.08.022.

[4] W. Zhou, T. S. Zhao, W. Liu, and J. J. Tang, "Tower crane safety on construction sites: A complex sociotechnical system perspective," (in English), *SAFETY SCIENCE,* vol. 109, pp. 95-108, NOV 2018, doi: 10.1016/j.ssci.2018.05.001.

[5] S. C. Y. Lu, D. Li, J. Cheng, C. L. Wu, R. E. S. Int Inst Prod Engn, and R. E. S. Int Inst Prod Engn, "A model fusion approach to support negotiations during complex engineering system design," presented at the CIRP ANNALS 1997 MANUFACTURING TECHNOLOGY, VOLUME 46/1/1997: ANNALS OF THE INTERNATIONAL INSTITUTION FOR PRODUCTION ENGINEERING RESEARCH, 1997.

[6] P. G. Maropoulos and D. Ceglarek, "Design verification and validation in product lifecycle," (in English), *CIRP ANNALS-MANUFACTURING TECHNOLOGY,* vol. 59, no. 2, pp. 740-759, 2010, doi: 10.1016/j.cirp.2010.05.005.

[7] F. Tao *et al.*, "Digital twin-driven product design framework," (in English), *INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH,* vol. 57, no. 12, pp. 3935-3953, JUN 18 2019, doi: 10.1080/00207543.2018.1443229.

[8] H. Zhang, Q. Liu, X. Chen, D. Zhang, and J. W. Leng, "A Digital Twin-Based Approach for Designing and Multi-Objective Optimization of Hollow Glass Production Line," (in English), *IEEE ACCESS,* vol. 5, pp. 26901-26911, 2017, doi: 10.1109/ACCESS.2017.2766453.

[9] D. J. Opoku, S. Perera, R. Osei-Kyei, and M. Rashidi, "Digital twin application in the construction industry: A literature review," (in English), *JOURNAL OF BUILDING ENGINEERING,* vol. 40, AUG 2021, Art no. 102726, doi: 10.1016/j.jobe.2021.102726.

[10] Y. Pan and L. M. Zhang, "A BIM-data mining integrated digital twin framework for advanced project management," (in English), *AUTOMATION IN CONSTRUCTION,* vol. 124, APR 2021, Art no. 103564, doi: 10.1016/j.autcon.2021.103564.

[11] L. Zhang, L. F. Zhou, and B. K. P. Horn, "Building a right digital twin with model engineering," (in English), *JOURNAL OF MANUFACTURING SYSTEMS,* vol. 59, pp. 151-164, APR 2021, doi: 10.1016/j.jmsy.2021.02.009.

[12] H. A. Shanbari, N. M. Blinn, and R. R. Issa, "LASER SCANNING TECHNOLOGY AND BIM IN CONSTRUCTION MANAGEMENT EDUCATION," (in English), *JOURNAL OF INFORMATION TECHNOLOGY IN CONSTRUCTION,* vol. 21, pp. 204-217, 2016.

[13] G. C. Zhang, P. A. Vela, P. Karasev, and I. Brilakis, "A Sparsity-Inducing Optimization-Based Algorithm for Planar Patches Extraction from Noisy Point-Cloud Data," (in English), *COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING,* vol. 30, no. 2, pp. 85-102, FEB 2015, doi: 10.1111/mice.12063.

[14] S. Kaewunruen and Q. Lian, "Digital twin aided sustainability-based lifecycle management for railway turnout systems," (in English), *JOURNAL OF CLEANER PRODUCTION,* vol. 228, pp. 1537-1551, AUG 10 2019, doi: 10.1016/j.jclepro.2019.04.156.

[15] R. D. Lu and I. Brilakis, "Digital twinning of existing reinforced concrete bridges from labelled point clusters," (in English), *AUTOMATION IN CONSTRUCTION,* vol. 105, SEP 2019, Art no. 102837, doi: 10.1016/j.autcon.2019.102837.

[16] G. Angjeliu, D. Coronelli, and G. Cardani, "Development of the simulation model for Digital Twin applications in historical masonry buildings: The integration between numerical and experimental reality," (in English), *COMPUTERS & STRUCTURES,* vol. 238, OCT 1 2020, Art no. 106282, doi: 10.1016/j.compstruc.2020.106282.

[17] A. Shahbaz and K. H. Jo, "Deep Atrous Spatial Features-Based Supervised Foreground Detection Algorithm for Industrial Surveillance Systems," (in English), *IEEE Trans. Ind. Inform.,* vol. 17, no. 7, pp. 4818-4826, JUL 2021, doi: 10.1109/TII.2020.3017078.

[18] X. T. Vo, T. D. Tran, D. L. Nguyen, and K. H. Jo, "Dynamic Multi-Loss Weighting for Multiple People Tracking in Video Surveillance Systems," in *2021 IEEE 19th International Conference*

on Industrial Informatics (INDIN), 21-23 July 2021 2021, pp. 1-6, doi: 10.1109/INDIN45523.2021.9557515.

[19] D. Hou, T. Liu, Y. T. Pan, and J. Hou, "AI on edge device for laser chip defect detection," presented at the 2019 IEEE 9TH ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE (CCWC), 2019.

[20] S. Tang, F. He, X. Huang, and J. Yang, *Online PCB Defect Detector On A New PCB Defect Dataset*. 2019.

[21] X.-T. Vo and K.-H. Jo, "A review on anchor assignment and sampling heuristics in deep learning-based object detection," *Neurocomputing,* vol. 506, pp. 96-116, 2022/09/28/ 2022, doi: https://doi.org/10.1016/j.neucom.2022.07.003.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[23] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, Cham, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016// 2016: Springer International Publishing, pp. 21-37.

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

[25] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 04/08 2018.

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," p. arXiv:2004.10934. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2020arXiv200410934B

[27] F. Tao, J. F. Cheng, Q. L. Qi, M. Zhang, H. Zhang, and F. Y. Sui, "Digital twin-driven product design, manufacturing and service with big data," (in English), *INTERNATIONAL JOURNAL OF ADVANCED MANUFACTURING TECHNOLOGY,* vol. 94, no. 9-12, pp. 3563-3576, FEB 2018, doi: 10.1007/s00170-017-0233-1.

[28] G. Schrotter and C. Hurzeler, "The Digital Twin of the City of Zurich for Urban Planning," (in English), *PFG-JOURNAL OF PHOTOGRAMMETRY REMOTE SENSING AND GEOINFORMATION SCIENCE,* vol. 88, no. 1, pp. 99-112, FEB 2020, doi: 10.1007/s41064-020-00092-2.

[29] J. F. Cheng, W. H. Chen, F. Tao, and C. L. Lin, "Industrial IoT in 5G environment towards smart manufacturing," (in English), *JOURNAL OF INDUSTRIAL INFORMATION INTEGRATION,* vol. 10, pp. 10-19, JUN 2018, doi: 10.1016/j.jii.2018.04.001.

[30] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital Twin: Enabling Technologies, Challenges and Open Research," (in English), *IEEE ACCESS,* vol. 8, pp. 108952-108971, 2020, doi: 10.1109/ACCESS.2020.2998358.

[31] T. H. Deng, K. R. Zhang, and Z. J. Shen, "A systematic review of a digital twin city: A new pattern of urban governance toward smart cities," (in English), *JOURNAL OF MANAGEMENT SCIENCE AND ENGINEERING,* vol. 6, no. 2, pp. 125-134, JUN 2021, doi: 10.1016/j.jmse.2021.03.003.

[32] F. Jiang, L. Ma, T. Broyd, and K. Chen, "Digital twin and its implementations in the civil engineering sector," (in English), *AUTOMATION IN CONSTRUCTION,* vol. 130, OCT 2021, Art no. 103838, doi: 10.1016/j.autcon.2021.103838.

[33] A. K. Ghosh, A. Ullah, R. Teti, and A. Kubo, "Developing sensor signal-based digital twins for intelligent machine tools," (in English), *JOURNAL OF INDUSTRIAL INFORMATION INTEGRATION,* vol. 24, DEC 2021, Art no. 100242, doi: 10.1016/j.jii.2021.100242.

[34] Y. C. Wang, F. Tao, M. Zhang, L. H. Wang, and Y. Zuo, "Digital twin enhanced fault prediction for the autoclave with insufficient data," (in English), *JOURNAL OF MANUFACTURING SYSTEMS,* vol. 60, pp. 350-359, JUL 2021, doi: 10.1016/j.jmsy.2021.05.015.

[35] D. T. Kutzke, J. B. Carter, and B. T. Hartman, "Subsystem selection for digital twin development: A case study on an unmanned underwater vehicle," (in English), *OCEAN ENGINEERING,* vol. 223, MAR 1 2021, Art no. 108629, doi: 10.1016/j.oceaneng.2021.108629.

[36] S. M. Liu, Y. Q. Lu, J. Li, D. Q. Song, X. M. Sun, and J. S. Bao, "Multi-scale evolution mechanism and knowledge construction of a digital twin mimic model," (in English), *ROBOTICS AND COMPUTER-INTEGRATED MANUFACTURING,* vol. 71, OCT 2021, Art no. 102123, doi: 10.1016/j.rcim.2021.102123.

[37] G. Yu, Y. Wang, Z. Y. Mao, M. Hu, V. Sugumaran, and Y. K. Wang, "A digital twin-based decision analysis framework for operation and maintenance of tunnels," (in English), *TUNNELLING AND UNDERGROUND SPACE TECHNOLOGY,* vol. 116, OCT 2021, Art no. 104125, doi:

10.1016/j.tust.2021.104125.

[38] G. H. Zhou, C. Zhang, Z. Li, K. Ding, and C. Wang, "Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing," (in English), *INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH,* vol. 58, no. 4, pp. 1034-1051, FEB 16 2020, doi: 10.1080/00207543.2019.1607978.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.