A Dissertation Submitted to Lancaster University for the Degree of Doctor of Engineering

Digital twin of construction crane and realization of the physical to virtual connection

Candidate: Enliu Yuan

Student ID: 35119940

Supervisor: Professor Jianqiao Ye
Professor Mohamed Saafi

Specialty: PhD Engineering

Degree-Conferring-Institution: Lancaster University

Abstract

Digital twin is an integrated multi-physics representation of a complex physical entity. This article constructs the digital twin of the construction crane, proposes a framework for the construction of the tower crane digital twin, and realizes the connection from physical to virtual in the concept of digital twin. The main contributions are divided into three parts: development of tower crane monitoring dataset, tower crane detection and tower crane operation mode recognition. By using labellmg to annotate 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 selected in the tower crane detection. Edge extraction is used to improve the quality of the raw dataset and distance-intersection-over union non-maximum suppression is used to replace the traditional non-maximum suppression part in the Yolov5x algorithm to improve the detect accuracy when some tower cranes are overlapping. The final test set detection accuracy rate is 93.85%. After comparing the LSTM and CNN algorithms, 3DResNet algorithm is selected for tower crane operational mode recognition. The raw dataset is augmented by rotating the image by ±10° and ±20°, and the augmented dataset enlarges five times. Using these methods, the final recognition accuracy of tower crane operation mode reaches 87%. These models can be installed on the cctv to monitor the running status of the 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.

Key words

Digital twin, construction crane, object detection, Deep Neuron Network.

Acknowledgements：

Time flies, and the scene of doctoral admission seems to be still vivid in my memory, but at this moment, when the dissertation is about to complete, which means that my doctoral life is coming to an end. Looking back on the past few years of studying for a PhD, I was bewildered and confused, sad and entangled, excited and rewarded, but more grateful. Here, I would like to express my sincere thanks to every teacher, classmate, friend and family member who helped, supported and cared about me.

Words cannot express my gratitude to Professor Jianqiao Ye, Professor Mohamed Saafi and Professor Jian Yang for their invaluable patience and feedback. They are generous, kind, and extremely responsible to their students. The rigorous and upright, truth-seeking and pragmatic work style deeply shocked and influenced me. They devoted a lot of effort into my scientific research, provided careful guidance and help from the determination of scientific research direction to the development of research work, to the revision and review of graduation dissertation.

I am also grateful to my office mates, for their coding help, late-night feedback sessions, and moral support. They are Zewen Gu, Xiaoxuan Ding, Yueqi Wu, Xinger Wang, Chenjun Zhao, Wuyue Xiong and etc.

Many thanks to my friends who have always encouraged and supported me. They are Minxuan Ni, Yueyang Gao, Tianyu Wang, Qiyang Gu, Yingdan Chen, Kaixin Cheng, Yatao Ge, Xiangyu Gao, etc. After the study and work, they always had lunches, dinners and fine dining with me, we explored gastronomy in many countries and cities. Among them, Noma, Xinrongji and Jinsha ting were so delicious. Of course, the Cornish Bakery in Lancaster town centre is also quite good, and I have spent many breakfasts there.

Lastly, I would be remiss in not mentioning my family, especially my parents, my father Yuanchun shen, my mother Jiawei Xu, and also my grandfather Baotong Xu. Thanks to my grandfather for encouraging me to start the PhD journey. Additionally, this endeavor would not have been possible without the support from my parents during my PhD study, so that I can better devote myself to my study and work. Their belief in me has kept my spirits and motivation high during this process.

Contents

List of figures

List of tables

1. Introduction

1.1 Research background and significance

The development of the building information modelling (BIM) sector has promoted the progress of digitalization in the construction industry. In 2011, the Ministry of Housing and Urban-Rural Development of China released the "2011−2015 Construction Industry Informatization Development Outline," which included BIM for the first time, and released the national standard "Construction Application Standard of Building Information Modelling" GB/T51235-2017. This standardized and directed the application of BIM in the design, construction, operation, and maintenance of various engineering projects, supported the implementation of engineering construction informatization, and improved the efficiency and effectiveness of information application. European and American countries implemented BIM in the construction industry earlier. The United States was the first country to engage in research on construction industry informatization. On 2003, the General Services Administration (PSA) began a real-time 3D-4D-BIM project to achieve technological transformation and improve the economic efficiency, safety, and aesthetics of buildings. It involved examining the application of BIM across the entire project life cycle. In May 2011, the British government released the "Government Construction Strategy," which required enterprises to achieve full coordination of 3D-BIM by 2016. Meanwhile, the South Korean government demanded that BIM applications in all public projects be implemented by 2016. The Singapore Construction Management Agency requires that all government projects use BIM models and encourages universities to offer BIM-related courses.

During the construction process, tower cranes are used to transport materials, so operational safety is of the utmost importance. Tower cranes and booms often collapse and cause accidents because they have not been serviced and/or have not been subject to reasonable and standardized operational and maintenance measures. This can pose a serious threat to the safety of construction sites and personnel. Common modelling methods comprise BIM, laser, and point cloud scanning. Enterprises have been required to develop BIM models in the project preparation stage, but these are often vague and incomplete. The week-long workload of a construction project may be presented in just a few seconds in a BIM model, and the focus is more on a visual representation rather than displaying the dynamics of the entire construction

platform or the tower crane. This does not help towards improving safety. While many construction sites in mainland China use smart technology; for example, video equipment may be installed to collect real-time data, but these are then used for display and monitoring only. This entails a great deal of communication traffic but little potential for analysis. All in all, current research on the dynamic modelling of tower cranes is scarce, and the results thereof do not make it possible to reflect attitude changes dynamically and in real-time. At the same time, the transformation and upgrading of the construction industry in recent years and the new concept of the digital twin have led to increasing demands for the dynamic modelling of physical entities. The data collected by smart construction site projects, which should be used to show the operational mode of tower cranes, have no real practical purpose currently.

The digital twin concept originated in the military and aerospace industries and has since been used in other contexts, for example, manufacturing. The engineering sector has arguably been the slowest to digitalize, and only recently has digital twin technology begun to attract the attention of engineering researchers and practitioners. For instance, there are currently insufficient digital twin applications for construction sites and equipment (such as tower cranes), even though these could benefit enormously from their use. Tower cranes are one of the most widely used pieces of construction equipment. As has been noted, their operational safety is fundamental to the industry. It is therefore vitally important to monitor their status, obtain feedback, and analyses any kind of malfunction both physically and digitally. The digital twin idea can be used to overcome the limitations of existing detection technology and improve safety inspection procedures and work efficiency by realizing independent innovation and reducing the detection capital threshold for a traditionally large number of sensor arrangements.

The present study uses computer graphics, modelling, deep learning, and machine learning, and adopts a method of pre-modelling by which a tower crane model can be developed in the virtual space. The operational mode of the tower crane is reviewed and assessed through computer vision algorithms; the operational data are transmitted to the virtual model through physical to virtual connections to construct a digital twin and thus help managers to monitor the status of the tower crane. The digital twin concept is therefore of great theoretical significance and practical value in construction terms.

1.2 Research purpose and object

The present study aims to propose a digital twin of a tower crane with real-time physical-to-virtual connection, allow for the collection, preprocessing, and data enhancement of operational tower crane data for deep learning, and develops reliable software and experiential support. The study is based on previous research work on deep learning algorithms and digital twin components as presented in the literature. This makes it possible to establish a framework for a tower crane digital twin. By improving upon and optimizing the Yolov5 deep learning model, detecting and identifying tower crane operations from videos and images are realized. In addition, the tower crane in each video frame can be segmented using a bespoke algorithm. Finally, the physical-to-virtual connection in the digital twin is achieved through a residual network that allowed for the recognition of the operational mode of the tower crane from real-time monitor videos.

The main research object is the QTZ80 tower crane. It is selected for the following reasons: (1) The tower crane is key to the operation of the construction site. It is used to transport steel bars, formwork, masonry, equipment, and other materials. The safety of the tower crane is therefore paramount. (2) The QTZ80 tower crane is one of the common tower cranes used in construction projects. Its arm length is 55 meters; the maximum lifting capacity at 55 meters is 1.3 tons; the independent height is 45 meters; and the maximum lifting capacity is 8 tons. Because it is so widely used, the QTZ80 was considered ideal for in-depth exploration.

1.3 Research innovation

(1) Since there is currently no open access, well-labelled tower crane image dataset, this research sets out to build one based on video recordings. Including a tower crane video image recognition dataset and a tower crane motion mode recognition dataset. This work provides a data reference for the selection, design, and optimization of subsequent deep learning algorithms.

(2) Building a digital twin framework of the tower crane, which comprises the tower crane entity itself; the tower crane working background; the tower crane in the virtual model; the virtual background; the physical-to-virtual connection between the physical tower crane and

the virtual tower crane; the virtual-to-physical connection; and the digital twin process of information interaction.

(3) An improved Yolov5 model is proposed that makes the improved one-stage detection algorithm neural network more suitable for situations in which the detection tasks of multiple targets and tower crane distances are short. Using distance-intersection-over-union to replace the original loss function improved the detection of overlapping multiple tower cranes. Edge extraction is used to eliminate the noise in the image and thus improve the recognition ability of the yolov5 network.

(4) Several common pattern recognition algorithms are compared, and the superiority of 3DResNet in sequential object motion recognition is established through comparative experiments. At the same time, an improved image rotation augmentation dataset using the method of combining edge extraction is proposed. It is now possible to determine the operation of the tower crane through video monitoring and respond to the operation of the model, realizing the physical-to-virtual connection of the digital twin concept.

1.4 Organization of the study

1.4.1 Organization

The study comprises seven chapters:
The first chapter introduces the research objectives and innovations. The remainder of the chapter discusses the organization of the study, the technical elements, and its limitations.

The second chapter discusses the concept of the digital twin, summarizes its history and current research status in light of scholarly findings, and presents its five components (i.e., the physical entity and environment; the virtual entity and environment; the physical-to-virtual connection; the virtual-to-physical connection; and the digital twin process) and two characteristics (i.e., real-time and autonomy). The chapter also reviews and summarizes the five principal digital twin applications—health monitoring, smart manufacturing, industry optimization, autonomous prediction and control, and engineering—and discusses technologies that might enable the realization of interdisciplinary digital twins. Finally,

suggestions are made for future research.

The third chapter introduces the framework of the tower crane digital twin. It is based on the five components and the two characteristics referred to above, applies it to a specific situation, and explains how the virtual model of the tower crane was constructed using modelling software.

The fourth chapter proposes the creation and optimization method of the tower crane image dataset. Through low-cost data collection and a series of image preprocessing methods, the background noise of the image was reduced, and the peripheral frame structure of the tower crane image was obtained. In addition, the images of tower cranes were annotated so that the neural network can learn them through training and establish the logical association between image features and image ontology. In other words, the model can learn the characteristics of tower crane images and achieve the target of training image recognition. In describing the connection between tower crane object detection and operational mode recognition, the chapter also introduces the application of tower crane segmentation, which includes separating the tower crane image from the larger picture, thereby further reducing the background noise of the image and concentrating the image dataset on the tower crane itself, thus improving the accuracy of subsequent deep learning. Finally, the initial dataset is augmented by data enhancement methods to avoid overfitting caused by a lack of data during training.

The fifth chapter studies the object detection method based on tower crane images. First, the state of the art of target detection algorithms is reviewed. Computer vision, deep learning, the current status of algorithm research, and the basic convolutional neural network (yolov-5) are introduced (including the related content, framework, loss function, and non-maximum suppression of the yolov5 algorithm). The limitations of the original yolov5 when training the tower crane image dataset (e.g., failure to detect multiple targets that were very close to one other and detection accuracy problems) are discussed and a series of network structure optimization methods are proposed. These include distance-intersection-over-union and edge extraction to improve the ability of the neural network model in detecting tower cranes. Finally, the performance of the improved Yolov5 algorithm is verified through the design of a series of ablation experiments.

The sixth chapter discusses research on tower crane operational mode recognition based on the image dataset. First, several candidate neural networks featured in the present study are introduced, namely long-short term memory, a convolutional neural network, and a residual network. Here, LSTM and CNN, 2DResNet and 3DResnet, and the optical flow method are combined and used in preliminary attempts. The neural network 3DResNet is finally derived through the training of the above algorithm. The limitations of the original 3DResNet model in processing tower crane image datasets (e.g., the model depth problem and the small quantity of data) are recognized, and 3DResNet neural networks of four different depths (18, 34, 50, 101 layers,) are trained to compensate for these. The best two depths of 3DResNet were employed to test its effect on the augmented dataset (using edge extraction and image rotation augmentation). The effects of the two data enhancement methods are examined by designing an ablation experiment. Finally, the 3DResNet50 trained on the augment dataset is established as the best algorithm for tower crane operational mode recognition.

The seventh chapter summarizes the article and discusses prospects for future research.

1.4.2 Technical route

According to the content arrangement of the article in Section 1.4.1, the technical route of this research is shown in Figure 1.1.

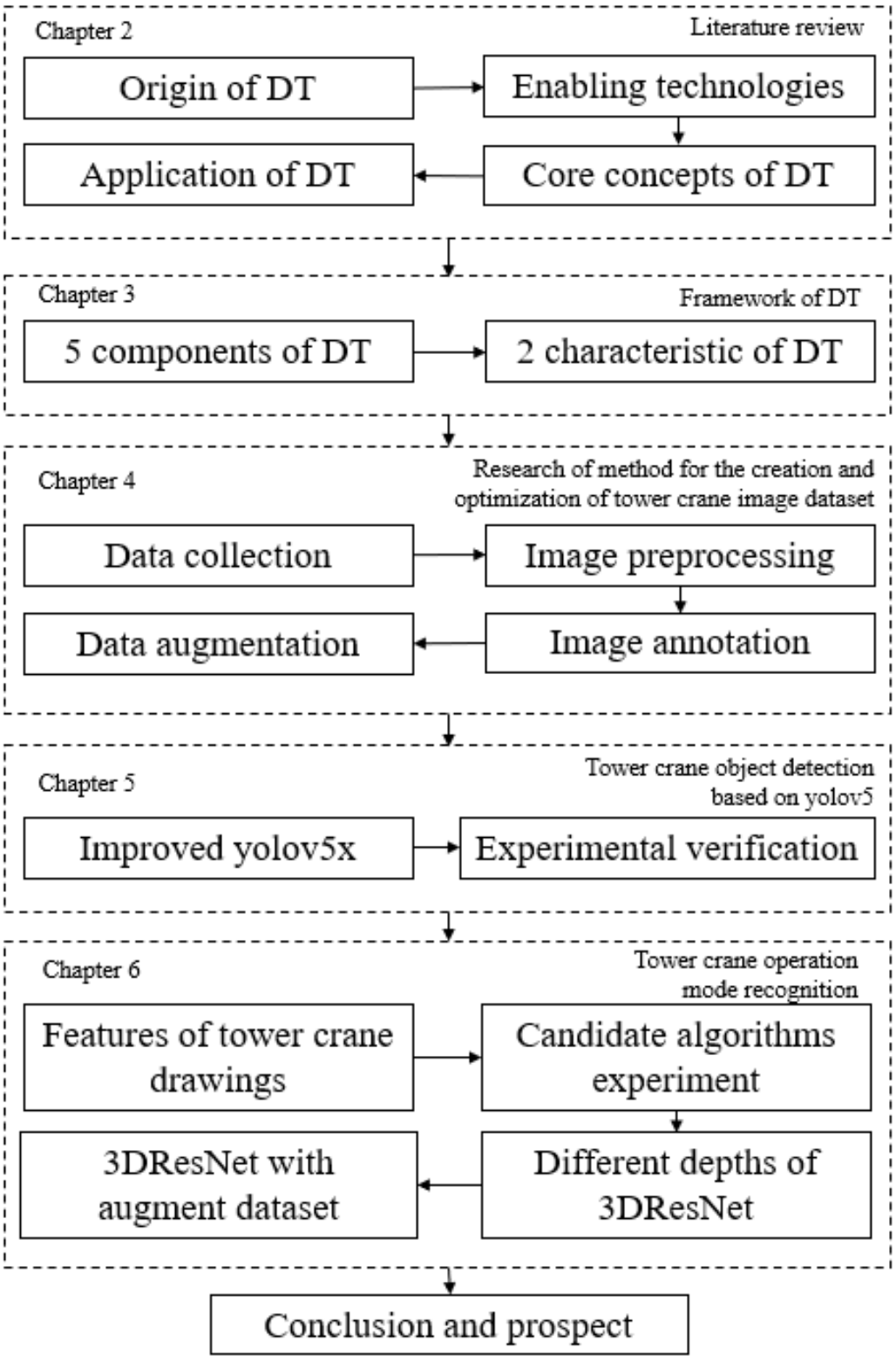


Figure 1.1: Technical route of building the digital twin of tower crane

1.5 Summary

The present study encountered many problems and inevitably has several limitations:

(1) In real-world construction projects, many different types of tower cranes are used, and construction procedures can vary considerably. Compared with mature and large-scale datasets, the number of tower crane video datasets in the present study is relatively small, and their quality is poor. The dataset might therefore be expanded to include different types of tower crane in the future. The study only modelled the QTZ80 tower crane, which is insufficient for a mature construction site digital twin; every style of tower crane should be modelled.

(2) Lack of sensor data. Many studies on digital twins uses sensors, whereas the present study uses the computer vision method to establish the physical-to-virtual connection. Our tower crane digital twin only contains video and picture data and lacks physical sensor data, so it is therefore incomplete. When the current situation improves, the results should be verified by introducing sensor data.

(3) There were insufficient data available. The selection of the shooting locations was random, so the rotation angle of the tower crane could not be calculated; only the motion state of the tower crane could be established. Subsequent researchers can fix the position of the camera and measure the distance between it and the base of the tower crane. The rotation angle and rotation speed through the data of the tower crane itself could then be measured and the three-dimensional space calculated.

2. Literature review

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 [1]. At the beginning of the 21st century, virtual models of products began to include the definition of product personality [2]. Modelling has gradually become common tools to solve some problem 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 come into being and become more and more specific.

The concept of digital twin was first proposed by Dr Grieves when he gave a lecture on product lifecycle management 2003. He said that digital twin is the digital representation of actual physical product [3]. Dr Grieves also worked with John Vickers at NASA, using digital twin technology on aviation aircraft to simulate the operational status of aircraft. According to [4], digital twin is considered to be an integrated multi-physics representation of a complex product. It has a variety of sizes and includes probabilistic simulations. Its function is to provide images that reflect the status of its twins. A digital twin is different from traditional CAD and the Internet of Things. CAD is usually related to virtual mapping and images, while the Internet of Things pays more attention to the physical world [5]. A digital twin mainly includes physical entities, virtual models, and the connection of physical and virtual parts. It is updated by modelling, simulating, and self-optimizing the physical entities [6]. More specifically, a digital twin is the modelling of physical objects using digital methods to simulate how the object will behave, in reality. This is followed by a virtual simulation of the product itself and even the entire production process. In this way, manufacturers can improve production productivity and safety. There is a closed-loop mechanism between static design and dynamic execution which means that the static design is checked through dynamic execution, and the simulation result is checked through static design for dynamic execution.

In subsequent research, digital twin technology is widely used in manufacturing. Scholars used the virtual model of the physical system to design the production line, which can be continuously controlled and optimized and can also calculate the working time of the assembly line and explore the optimal solution for the entire system [7]. With the rapid development of

Industry 4.0, smart buildings, and interconnected factories, it is expected that digital twin used for the manufacturing industry will become the system required by future manufacturing systems [8]. Padovano [9] used a digital twin to transfer the existing manufacturing environment to a manufacturing environment that satisfies the concept of future manufacturing systems. Zheng [10] applied the concept of digital twins to product functions, manufacturing processes, and product performance testing, by using a semi-physical simulation system to process information and link it to the physical world through a network module. In this literature, the digital twin can be used to represent many states of the lifecycle of the product.

2.2 Bibliometric analysis of literature of digital twin

The present study examines the emerging trends regarding the concept of the digital twin, which emerged in 2010. The search phrase was “Digital twin”. To display accurately and comprehensively trends in this research field, the search period was set to 2010--2021. As of August 2021, a total of 2,215 articles were selected. Information relating to document type, publication year, country, and keywords was used as the basis for quantitative analysis. The present study adopted the two indicators of impact factor and the h-index for a comprehensive evaluation of the influence of the papers deriving from various countries in recent years. Related documents were divided into “single country documents” and “national cooperation documents”; documents by authors from different institutions located in the same country are defined as documents from one country.

2.2.1 General trends

Figure 2.1 shows the trends of articles and proceeding papers published from 2010 to 2021. A total of 2,215 articles in 11 categories were published between 2010 and August 2021, mainly in the form of articles and proceedings papers. These accounted for 57.1% (1,265) and 36.2% (801), respectively. The remaining categories were review papers (104); early access papers (101); editorial materials (49); book chapters (40); meeting abstracts (5); new items (4); books (2); corrections (1); and data papers (1). Between 2010 and 2016, there were very few discussions about digital twins. However, since 2016, the number of documents overall has increased rapidly, and the number of related papers has continued to grow. Before 2018, the

number of conference papers was larger than the number of papers overall, but after 2019, the number increased substantially, eventually exceeding the number of conference papers.

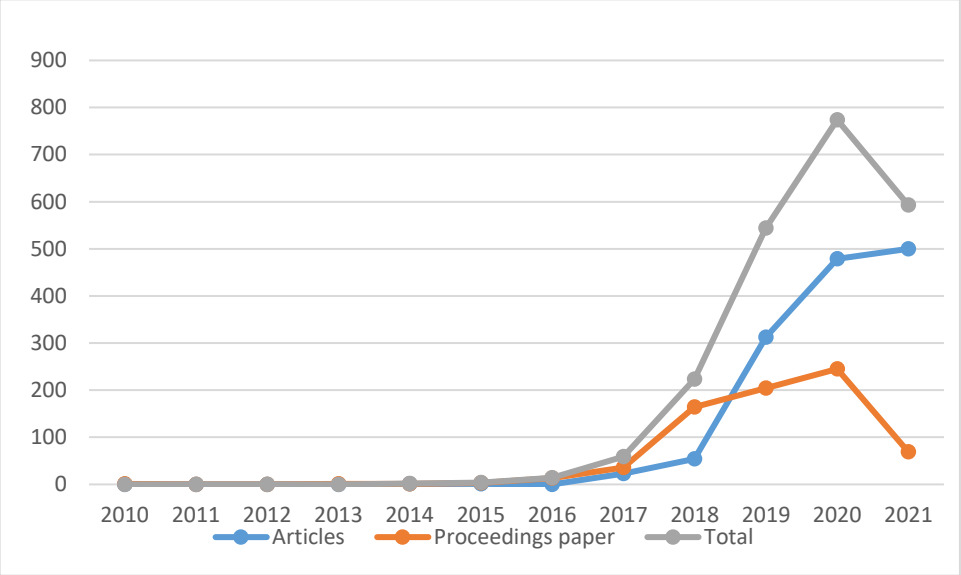


Figure 2.1: Trends of articles and proceeding papers published in 2010-2021

2.2.2 Comparison between the research contributions of countries/regions

Table 2.1 displays the top 10 countries/regions that have contributed the most in the digital twin field and summarises the number of independent and cooperative documents published by these countries/regions, the main cooperating countries, and the research institutions. The top three contributing countries accounted for 49.7%; the top ten countries/regions contributed more than 80%. Amongst them, Germany has the highest proportion (17.7 percent, 391 papers in total), followed by China (17.5 percent, 388 papers in total) and the United States (14.5 percent, 322 papers in total). There is a significant gap between the top three (Germany, China, and the United States) and the other countries/regions. The number of collaborative documents between Germany, China, and the United States is also relatively large.

Table 2.1: Top 10 contributing countries/territories in 2010-2021

Nation/region	Total number Of literature	Ranking and proportion	Number of independent literatures	Number of cooperation literature	Main cooperating nation or region	Main research institution
Germany	391	1（17.7%）	285	106	USA, France	Siemens AG
P.R. China	388	2（17.5%）	265	123	USA, Singapore	Beihang University
USA	322	3（14.5%）	190	132	P.R China, England	United States Department of Energe Doe
England	178	4（8.0%）	82	96	P.R. China, USA	University of Sheffield
Italy	148	5（6.7%）	80	68	Germany, France	Polytechnic University of Milan
France	107	6（4.8%）	36	71	Spain, Germany	Center National De La Recherche Scientifique Cnrs
Russia	103	7（4.7%）	73	30	Estonia, Finland	South Ural State University
R. Korea	92	8（4.2%）	79	13	Germany, Poland	Sungkyunkwan University Skku
Spain	92	9（4.2%）	42	50	France, England	Ik4 Tekniker
Sweden	73	10（3.3%）	41	32	P.R.China, England	Chalmers University of Technology

Figure 2.2 shows the numbers of literatures published by top 10 contribution country from 2017 to 2020. From this figure we can see that in 2019 and 2020, the number of people studying digital twins in Germany (subject to the nationality of the first author of the document) was approximately 110. The number of people studying this field was doubled from 71 to 143 in China; from 76 to 112 in the United States; and from 47 to 79 in the United Kingdom. We can see from these data that developed countries were the first to begin research on digital twins, and they are still the main contributors in this field. However, more and more people in developing countries (e.g., China) were starting to engage with the subject. On the other hand, Germany, which was the first country to study digital twins and the largest producer of studies before 2019, did not see a significant increase in the number of people in 2019–2020. In the latter year, its publication rate was surpassed both by China and the United States.

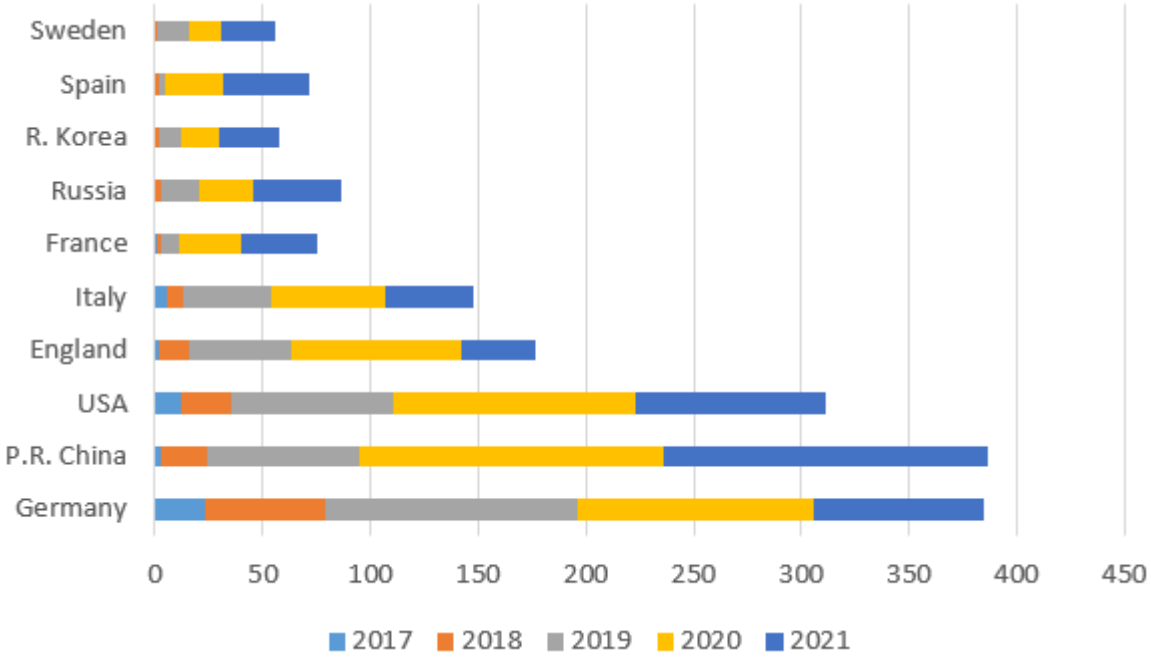


Figure 2.2: Numbers of literatures published by top 10 contribution country in 2017-2021

2.2.3 Keyword timeline

City space is a Java application used for visualization and analysis. It has been developed against the background of scientific metrics and data visualization and is widely used for citation analysis of papers [11-13]. The present study yielded 2,212 articles remaining after the original documents were filtered through the Citespace software to remove duplicates. The software's runtime was set to "2017–2021", TopN = 50, the Pathfinder method was used for pruning (and overall network pruning), and the following map was generated.

Frontier trend analysis is a way of performing document clustering by continuously citing a fixed set of basic documents. It mainly uses co-citation clustering and citations as the basis of analysis to describe the transitional situation and research nature of a certain type of research field. A timeline chart can be used to display up-to-date details about the documents and the relationship between them using two-dimensional coordinates with time as the horizontal axis. In this timeline map, the size of the node indicates the frequency of occurrence of the keyword; the year of the node indicates the time when the keyword first appeared; and the connection between the nodes indicates that different keywords appear in an article at the same time,

suggesting a different inheritance relationship between time periods. The number of documents in different years represents the results published at that time and the stage of the field.

Figure 2.3 is the keyword timeline map of digital twin, which reveals that the largest node in the literature related to digital twin is “simulation”. The largest cluster in the figure is “sustainability”. One of the keywords is “sustainability”, which started to appear in the literature around 2019. As time progresses, the keywords are “industry 4.0”, “industry foundation classes (ifc)”, “energy consumption”, and so on; the second-largest cluster is “blockchain”, which includes keywords from near and far over time such as “additive manufacturing”, “simulation”, “cloud manufacturing”, and “manufacturing system”; other early keywords include “behavior simulation”, “cloud computing”, “cyber physical system”, and “smart factory”. Research on the subject is ongoing. According to the recent clustering results in Figure 2.3, the new keywords are “product design”, “energy efficiency”, “data driven”, “bibliometric analysis”, and “supply chain management”, which indicates that, compared with theoretical research on digitalization and frameworks in 2018, the current research on digital twins is more in-depth, comprehensive, and specific.

Figure 2.3: Keyword timeline map of digital twin

2.3 Core concepts of digital twin

To realize the digital twin process, we first need to realize the physical entity and the physical environment in the virtual model through modelling and other methods. At the same time, we expect 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. In the future, digital twins need to realize real-time data transmission and autonomous analysis and decision-making. Given the above framework and the requirements, the present chapter consists of seven subsections: 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 digital twins.

2.3.1 The physical entity and the physical environment

The physical entity is the twinned object, which is physical and exists in the real world. It has been described in the literature as having specific forms, for example, “aviation plane” [14], “product” [15], “model” [16], “production system” [17], “pipeline transporter” [18], “electric vehicles” [19], and other man-made entities. With the further development of digital twins, the term has also begun to refer to non-articles with vital characteristics, such as employees [20, 21], humans [22], digital twin for farm animal management [23], agriculture [24], and manufacturing [25].

Several studies [26-28] have discussed the physical environment of the digital twin, that is, the space where the physical entity is located. The physical environment includes the external environment and physical processes. The physical process is the way the system expresses itself in the physical environment and the mechanism by which the system entities undergo state changes [29]. In product safety lifecycle management, it also refers to the factory and the external factors that affect industrial output. Under these circumstances, it is necessary to measure various factors that may affect the physical entity, then input these environmental

changes into the virtual twin environment. For example, Pylianidis [24] proposed the application of digital twins to agriculture, but the interaction between the living system and the external environment is difficult to quantify and simulate. The demand for multiple parameters encourages a combination of sensors, the Internet of Things, and digital twins. By creating a comprehensive and accurate virtual environment and searching as far as is possible for more complete and detailed external environment data, simulations can be better performed in the virtual environment, and the results obtained will be more accurate.

2.3.2 Virtual entities and the virtual environment

Virtual entities are the twin objects; they are virtual and exist in the digital world. Studies have referred to the expression of aviation aircraft, products, models, production systems, pipeline transport aircraft, employees, children, humans, and so on in this digital world. Because of the different simulation angles, there are multiple virtual entities in digital twins, and each virtual entity has a specific purpose. For example, in the product life cycle, the digital twin may represent product design, process planning, production layout, process simulation, output optimization and many other applications.

The virtual environment exists in the digital domain as the mirror image of the physical environment. The virtual environment is a virtual representation of the physical environment [29]. Virtual representation is used to determine whether physical state changes are needed and what actions need to be taken to achieve the target state. The operations that affect the necessary physical processes to achieve the target physical state are then performed. These are often physical measurement methods involving sensors, which are used to relay key metrics from the physical to the virtual to achieve digital twins. The more information that is collected in the virtual environment, the better the virtual entity can be simulated, thereby improving the integrity of the model. Under general conditions, the virtual environment includes external factors such as process control, manual operation, environmental impact, and fault detection [30].

2.3.3 The physical to virtual connection

The biggest difference between digital twin and traditional modelling is that the latter only

describes physical entities through modelling methods, but digital twin is online and instantaneously responding to changes in the state of physical entities, such as changes in the state of physical entities themselves and changes of the external physical environment, and so on, then reflect these changes to the virtual entity. In this situation, it will pass through the physical to the virtual connection. The connection from the physical entity to the virtual entity itself is a process of capturing the physical state of the object and realizing the twinning of the physical and virtual entities. By determining the variables between the physical entity and the virtual entity, the virtual entity can be updated accordingly.

At the level of technical application, the parameters of the changes in the physical entity can be updated through sensors [30-32], 5G [33], internet of things(IoT) [28, 34-36], cyber physical systems (CPS) [37], and so on. Miller [38] used the space-enhanced computer-aided design model to improve the interoperability between the model and the physical entity and expanded the 3D-CAD model through behavioral information. This enabled them to realize the connection of the physical model to the virtual model, which represents the first step in the creation of the digital twin. Schleich [15] transformed the scientific and technical basis of a geometric product specification language into a digital twin, provided it with an abstract model that covered all the necessary characteristics, and fully described the physical twin based on the physical shape of the object. Liu [17] carried out a mapping synchronization of the virtual model of the digital twin to the physical model, though they only applied it to production equipment. Jiang [39] used a discrete-time system (DES), integrated physical equipment, data, and logical models to build virtual models rapidly, and the connection between the workshop-level production line and the physical entities and the virtual twin model was realized.

MTconnect and AutomationML (a plug-and-play method) are used for data transmission between physical and virtual models. Hu [40] built a digital twin model using the MTconnect protocol to connect the physical and virtual model, a method that can be used to limit the amount of data collection and reduce communication delays. Through the MTconnect protocol, a digital twin of an intelligent machine tool embodying real-time processing data can be realized [41]. Schroeder [31] used AutomationML to exchange data between physical and virtual models. First, modelling tools were employed to build virtual models of equipment prototypes. After the models were defined, other systems were able to use them to exchange actively modular attribute information. Um [42] suggested the use of a common data model

and an appropriate communication layer to set up the digital environment automatically. Multiple assembly modules based on AutomationML are used to instantiate assembly modules in a data model composed of engineering data and a single production module.

2.3.4 The virtual to physical connection

Kritzinger [43] reviewed the historical progress of digital twin technology in the manufacturing industry. It has been characterized by a shift from a digital model with two-way manual data flow to a digital shadow in which physical object data automatically flows to digital objects. The idea of a digital twin with an automatic two-way data flow is currently attracting scholarly attention. Digital twins principally comprise physical products in physical space, virtual products in virtual space, and data and information interaction interfaces between physical space and virtual space [3]. The virtual-to-physical connection refers to an interface for digital and information interaction. Only the physical-to-virtual connection is incomplete. After the virtual model collects physical data and further processes it, the information is fed back to the physical entity through the virtual-to-physical connection for optimization [44-47] or to provide decision-making opinions [47-50] and other effects.

Zhuang's [51] defined the main components of a digital twin from a product perspective. They included product design data, product process data, product manufacturing data, product service data, and product retirement and scrap data. More attention is paid in their definition to the connection between the physical and the virtual. The object entity is described as perfectly as possible, but the connection between the virtual and the physical is not recognized. Tao [52] was the first to define a workshop digital twin from the perspective of workshop composition. The components consisted primarily of a physical workshop, a virtual workshop, a workshop service system, and workshop twin data. In Tao Fei's definition, the virtual-to-physical connection serves to transmit the results of virtual product simulation. Wang [45] used the digital twin concept to enhance the accuracy of failure prediction in the case of insufficient data, built digital twin models of different sizes of autoclaves and to simulate normal and failure environments, and use these data to enhance failure prediction.

Digital twins with physical-to-virtual connections and virtual-to-physical connections can make assumptions that can then be tested and adjusted through continuous adaptation and

improvement. The capacity to do so distinguishes the digital twin from traditional semi-physical simulation and modelling. Jones [53] believes that the value of the virtual-to-physical connection is in its ability to transmit the results of a simulation to the physical entity (the subject manager) for purposes of trial-and-error and to optimize the model. Here, product design may be given as an example. A company does not need to design a complete product to discover product defects; rather, it can modify the parameters as part of the redesign process. To a certain extent, the virtual to physical connection can realize the closed-loop function of the digital twin "design-simulation-execution-test-adjustment" and update the digital twin constantly through self-learning.

2.3.5 The digital twin process

The term digital twin refers to processes, but also technologies, methods, objects, models, and data. The digital twin process is one of change-measurement-realization, from the physical to the virtual and vice versa [53]. Changes in the physical state are expressed through parameters, and the measurement methods are used to capture the state changes that are transmitted through physical-to-virtual and virtual-to-physical connections and realized by synchronizing parameters in the physical-virtual environment.

2.3.6 The real-time characteristic of the digital twin

The digital twin is an action that synchronizes the virtual state and the physical state. When the parameters of a physical change are reflected in the virtual model, the physical virtual model will become consistent, thus realizing the "twin". Physical to virtual (external or internal changes reflected in the virtual state) and virtual to physical (predicting the physical state in the virtual environment) are the two kinds of connections needed to achieve good real-time interaction so as to promote the self-optimization of the digital twin. The real-time nature of digital twin has been discussed in several studies (e.g., [26, 53-55]. Digital twins are often combined with big data to achieve real-time interaction. Qi [56] examined the similarities and differences between digital twins and big data technology and their respective advantages and disadvantages from different perspectives (traditional and data) and analyzed the complementarity between digital twins and big data. Combining digital twins, big data, edge computing, and cloud computing will help promote the development of smart manufacturing.

2.3.7 The Autonomy characteristic of the digital twin

Historical data from sensors can be used to train digital feature models through machine learning so they can deal with faults that emerge in real-world operational processes. Combined with experience of experts, this can form the basis for the more accurate recognition of fault states in the future. The feature database can be enriched and updated with new and different types of faults; intelligent autonomous diagnoses can then be carried out. The autonomy of digital twins comprises two elements: automatic fault-tolerant control and self-optimization. Luo [57] established a multi-domain unified modelling method based on digital twins. This method makes the latter more autonomous, optimizes the operational mode, reduces the possibility of sudden failures, and improves the stability of CNC machine tool models.

Automatic fault-tolerant control is one of the core technologies of digital twin systems. The main objective of FTC design is to allow the system equipment to continue the previously set work after any malfunctions. It also reduces the extent of the malfunction, and it will not be a complete failure [58]. Usually, the FTC will use a close-loop control system to adapt to the fault or reconfigure the close-loop control system after the fault. The former one is more passive, and the latter is an active one [59]. He [60] combined a fault diagnosis scheme with a residual design model to enable stable and safe control and production in the event of a fault, tolerance control and online optimization are performed to realize the self-online optimization of the digital twin. Zhang [61] proposed a framework for data and knowledge driven digital manufacturing cell (DTMC). A fully integrated physical-virtual model senses and simulates the virtual production environment through real-time sensor data based on physical world. It can also strengthen their own execution capabilities to promote the autonomous operation of the overall system.

2.4 Application of digital twin

2.4.1 Health monitoring

Health monitoring is important throughout the entire lifecycle of equipment. Zheng [10] used different digital twin systems in different parts of the product lifecycle. For example, when testing the product performance, product manufacturing process, production equipment, and the factory operation health status detection, these will be linked through information processing and network modules to the physical world. Heng [62] pointed out that most of the existing prediction models are based on the laboratory environment, rather than the operational data collected from the physical world and environment. Therefore, considering the challenges of health monitoring, the intelligent predictive maintenance system combined with digital twin technology has attracted widespread attention.

Driven by these demands, Haag [63] applied the digital twin technology to the bending beam test bench and integrated it into the next generation manufacturing system. Ayani [64] used a simulation model (Simumatik3D) to build a digital twin model to repair the old machine. Before proceeding with machine maintenance, managers can gain confidence in the safety of the machine's future operation through virtual simulation and calculation, which improves project timing and resource integration. Talkhestani [65] proposed to check the consistency of the physical and virtual models of the digital twin in the workshop, considering the monitoring in the fields of mechanical, electrical and software. In order to meet the requirements of rapid personalized design, Wang [16] established a digital twin reference model for fault diagnosis of rotating machinery, used a parameter-based sensitivity analysis scheme to simulate imbalanced conditions, collected test data and then input the data into the digital twin, this model performed a failure analysis.

2.4.2 Smart manufacturing

From 2020 to 2021, COVID-19 had a great impact on the manufacturing industry [66]. The manufacturing industry had a large amount of labor in the past and was unable to fully produce production, resulting in great economic losses. In this situation, intelligent manufacturing has significant advantages. One of the key promoters of the intelligent manufacturing information revolution is digital twin technology [67]. In intelligent manufacturing systems, virtual entities

are high-fidelity representations of physical entities [53]. The use of digital twins in virtual manufacturing makes up for the lack of links between physical entities and virtual entities in traditional virtual manufacturing. Shao [68] specified three main scenarios for the application of digital twin technology in manufacturing. These included reducing the impact of equipment downtime (machine health monitoring), optimizing production planning and scheduling (Industry optimization) and initiating virtual testing (prediction and control). The "Machine Health Digital Twin" monitors, analyses and predicts failures in manufacturing equipment through equipment data and historic data in the production process, thereby reducing equipment downtime. In terms of production planning and scheduling optimization, the digital twin part will collect workshop data, analyses the current status of production, material inventory and customer service needs. This will drive on-time delivery through demand to optimize resource planning such as labor, equipment, materials and reduce delivery cycles and Inventory costs. The virtual test simulates actual application scenarios during the debugging process before the product is put into use, monitors the performance of the equipment and collects data, so as to achieve system optimization and continuous improvement.

Most of the current research on digital twins focuses on the manufacturing industry and its research framework [69]. Scime [70] proposed the digital twin framework in additive manufacturing. Zheng [10] built a digital twin model using body perception data. It includes the physical model of the workshop, the digital model based on the ontology and the virtual model generated by the Flexsim software. They also described the operating mechanism of the digital twin workshop in the model and the connection between the models has been built into the virtual physical model. However, most of this link is static and does not meet the requirements of real-time interaction. After that, the problem of dynamic link between models needs to be solved. Zhang [71] proposed a two-tier digital twin framework for the workshop. The first level is the digital twin model of the virtual workshop, which will contain digital twins of different service units. The second level is the digital twin of the service unit, which consists of seven digital twin files, such as, materials and logistics routes. These service units are scheduled and coordinated by the first-level virtual workshop model. Usually, the second-level service unit digital twin can operate independently and solve most scheduling problems. In some special scheduling problems, the virtual workshop digital twin will coordinate these second-tier service units digital twins and work together to complete dynamic scheduling. Qi [56, 72] described the multi-dimensional workshop digital twin model in literature, which

systematically divides this into three dimensions: workshop, automatic production line and equipment. The digital images of these three aspects have also developed more comprehensively. Tao [5] used the five-dimension digital twin workshop machine model and proposed a dynamic scheduling method enhanced by the digital twin and demonstrated the effectiveness of the method through a case analysis of the scheduling process in the processing workshop.

On the other hand, CPS is a vital component of smart manufacturing, and the digital twin is a prerequisite for the development of semi-physical simulation systems [73]. Unlike semi-physical simulation (CPS), digital twin needs to establish a two-way connection between physical and virtual models, but CPS is more biased towards network data transmission [37]. In the combination of digital twin and manufacturing semi-physical simulation system (CPS), Ghosh [44] proposed a sensor signal digital twin, which consists of five modules: input, modeling, simulation, verification and output, to adapt to CPS to help machine tools to perform tasks and conduct troubleshoot autonomously. In the digital twin framework proposed by Leng [74], the flow-type smart manufacturing system is modeled. The digital twin system is used for remote semi-physical simulation debugging in a distributed environment and is applied on the smartphone assembly line. On-site debugging, this method has the advantages of low debugging cost and short debugging time. Rocha and Barata [75] used a digital twin optimizer to predict the evolution of the Cyber-Physical Production System and reconfigure the system when there is a need for improvement. Fan [76] realized the digital information management of the project design, optimization and production stage of the automobile camshaft parts production line through the human and cyber physical manufacturing system.

What's more, the influence of human beings is indispensable in the digital twin of intelligent manufacturing [73]. Wei [22] proposed the Human digital twin. At present, the research is still facing problems, such as, human complexity, difficulty with modeling, large-scale data fusion analysis, diversity of data sources, data variability and heterogeneity. Graessler [20] taked the skills and needs of employees at work into account, the digital twin model was applied to employees, through automatic calculation and determination of work processes such as labor distribution, which enabled employees to be integrated into an automated distributed production control system. Nikolakis [21] collected a limited number of data to identify human movement. Sensors are installed on employees' arms and wrists. The data collected by these

sensors correspond to the key coordinates of the human body in the 3D space. By using this data, it is possible to identify and confirm the work patterns of some employees who are working. These worked as a starting point and began to gradually incorporate human influence into the digital twin model of workshops and factories.

2.4.3 Industry optimization

The characteristics of digital twins include real-time reflection, interaction, convergence and self-evolution [30]. Digital twin technology can promote the design and optimization of intelligent manufacturing systems [77]. Uhleman [78] indicated that the existing manufacturing industry should upgrade the multi-mode data collection technology and local independent optimization environment to meet the requirements of small and medium-sized enterprises with regards to flexibility, user-friendliness, scalability and service-oriented digital applications. Zhu [79] combined the digital twin technology in the thin-walled part manufacturing which require high levels of precision such as aerospace and has high levels of replacement frequency, high production cost, and long production cycle time. After combining with the digital twin technology, the cutting machine direction and path optimization, algorithms can be trained according to the updated digital twin model, but this method requires a certain amount of offline operation data calculation. Fang [80] applied digital twins in workshop scheduling, optimized workshop scheduling through scheduling resource parameter update methods and dynamic interactive scheduling strategies.

Digital twins can be used to improve the quality and level of products. Dong [81] used digital twins and integrated backtracking to redesign products. Lu [82] used customized software (Catia) to build the constant speed joint of the car, and the digital twin lifecycle management system from design, manufacturing to operation has changed the original production mode and improved the traceability of products. Liu [17] applied the digital twin-driven method to the manufacturing system in the automated flow shop. When the workshop is dynamically performing tasks, it can also collect information for decision-making, and use a two-layer interactive coordination mechanism to upgrade and develop the system to achieve optimal performances, which improves the applicability of the design.

Digital twins can be used for industrial optimization to reduce costs and increase benefits. Maschler [83] collected data during the design and testing phase of the physical systems and combined artificial intelligence, machine learning and digital twins to accelerate machine debugging, reduce costs and increase benefits. Fedorko [18] applied digital twin technology to the pipeline conveyor to record and simulate the data of the experimental process, thereby reducing the time for certain specific measurements. Botkina [84] used the component information of the international standard ISO 13399 (for cutting tool data representation and exchange) to develop a digital twin system for cutting tools, and used the previously developed IoT-based line information system architecture to optimize machining operations. At present, the development of digital twins is facing the situation of large framework and high cost. Kutzke [46] reduced its cost entry barrier through the digital twin of this subsystem. In the multi-objective optimization problem, the digital twin is allocated to the sub-system of the UAV, in order to maintain the performance of the subsystem and improve the overall reliability.

2.4.4 Autonomous prediction and control

Model-based predictive control technology measures and analyzes the data of physical entities and uses it to predict and reflect changes in physical entities, which is similar to digital twin technology. With the development of complex system control, intelligent control methods are needed to assist managers in data collection and decision-making analysis. Under this demand, Dotoli [85] integrated fuzzy logic, artificial neural networks and evolutionary algorithms and applied them to the control of industrial machinery, focusing on two model-based control technologies - model prediction control and technology based on computational intelligence.

Digital twin technology needs to have predictive capabilities [86], which realizes predictive control of physical entities through virtual-to-physical connections. Uhlemann [87] suggested that in order to reduce the product lead time, a systematic method must be used to collect data. Three forms of digital twins are used: digital shadows for real-time connection, digital twins for data processing, and digital twins with optimized parameter sets. Optimizing the parameters through the constantly updated database and putting forward suggestions for improvement of production control on this basis. Mukherjee [88] used 3D printing to produce complex components. Through digital twin technology, machine learning and big data, the composition of the model can be controlled and collected to predict whether the rules and

quality of future production components can meet the requirements. This method can reduce the number of false tests, optimizes the quality of products, reduce defects, and shortens the time between design and production. In terms of aircraft manufacturing and design, Tuegel [14] used digital twin to predict the life expectancy of aircraft and its components, and converted the parameters that potentially affect the life of the aircraft (the characteristics of the component materials, the manufacturing quality or the assembly method of the aircraft) into a probability of different results, to calculate the probability distribution of the remaining life. The digital twin also used, amongst others, actual flight parameters, historical records to simulate the virtual flight mission again.

At the same time, the digital twin needs to have decision-making capabilities. The decision module of the digital twin system needs to find the implicit knowledge in the twin data to make controlled decisions. For instance, Zhou [50] proposed a knowledge-driven digital twin manufacturing cell, which enables digital twins to have the functions of analysis, decision-making and dynamic adjustment. In terms of the quality of metal processing products, Liu [47] constructed a digital twin model to express the quality factors of the product and summarized all the factors leading to poor product quality, realized the transformation from data to knowledge and promoted the development of autonomous decision-making. Rosen [6] proposed an important method to improve the manufacturing industry from an automated system to an autonomous system. They used sensors to collect external real-time data and used automatic reasoning to make plans and determine specific manufacturing options. Next, actuators were used to perform a series of actions outlined in the plan. Digital twin technology promotes autonomous systems by collecting all important prior knowledge about products and production processes, realizing the process from sensors to actuators to an autonomous execution.

2.4.5 Application of digital twin in Engineering construction

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 [89] and IOT technologies has not been widely applied to constructions yet [90, 91]. Fuller [36] reviewed digital twins from smart cities, manufacturing, and healthcare, and proposed that the growth of artificial intelligence, IOT, and IIoT promoted the development of digital

twins.

In order to describe the digital twin of a single building or even a city, full element building modeling and simulation technology are required. The traditional method uses building information modeling (BIM). The development of the BIM industry has driven the digital development of the construction industry. Boje [90] used 4Dbim, which includes the time dimension in the design phase. In order to reflect the advantages of practitioners during the bidding phase, these BIM models are now often mandatory. Digital twin is different from traditional BIM. BIM does not pay attention to the relationship between the model and physical entities. BIM can be used in the design of buildings, 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 (ECoM4DT) [92], which is based on traditional modeling and simulation and can systematically guide the modeling process of the digital twin. For the completed building, Shanbari [93] proposed lidar modeling, Zhang [94] proposed point cloud modeling to describe the digital twin of the building, and Kaewunruen [95] used Revit and Navisworks software to build the BIM model of the railway transportation system. Combined with digital twin technology, to manage the entire lifecycle of the railway capacity system, reduce costs and increase sustainability. Lu [96] used point clouds and labelled point clusters to model concrete bridges as part of the rapid modeling of digital twins. Angjeliu [97] used the point cloud to model the vault of the Milan Cathedral as the first step of connecting the physical to the virtual model in the digital twin to prepare for the subsequent structural monitoring, operation and maintenance of the building. Digital twin technology has also participated in the urban planning and decision-making of Zurich [75]. Deng [34] realized smart cities and digital twin cities through mapping, BIM, IOT based on 5G technology, blockchain, edge computing, etc.

In the application of digital twin in the construction phase, Tran [98] used it in the prefabricated facades system of prefabricated and modular buildings. Through comparing the 3D semantic model reconstructed by point cloud and the digital model itself to measure the geometric data of the physical appearance. This method established a geometric quality assessment framework, but requires offline calculations, which cannot meet the real-time nature proposed in the core concept of digital twins in this paper. Pan [28] integrated BIM, IOT, and data mining

to transfer daily data logs to digital twin and developed a digital twin with connected physical virtual models to analyze and optimize the construction process. On the issue of material distribution and transportation, Buckova [99] proposed to combine the transportation of building materials with digital twins to optimize material handling, production logistics paths, movement modes to improve transportation efficiency and reduce material loss during transportation. During the operation process of construction, Greif [100] used a lightweight digital twin in the decision-making system of material scheduling and replenishment. Decision support is made through the analysis of historical data and replenishment decisions are made intelligently with digital twin data. Boje [90] proposed to apply the 5DBIM of the cost estimation method to the construction stage. Through artificial intelligence and digital twins, the scheduling of construction material resources can be optimized through active modeling, the information of workers on the construction site can be collected and abnormal behavior of workers can be detected in time.

Ozturk [101] summarised the development and shortcomings of digital twin in the architecture, engineering, construction, operations, and facility management industries. The industry is faced with the problems of low integration efficiency and lack of data, and most of the current research is focused on the theoretical level, through the construction of a digital twin framework to monitor the lifecycle. The author proposes to transform data into knowledge through "integrated cognitive technology" and "automated knowledge management" to promote the development of construction operation and maintenance of construction projects. Digital twin technology can also be applied to the operation and maintenance of linear infrastructure [102] and historical preservation buildings. Angjeliu [97] built a digital twin model of the Milan Cathedral to study how to maintain and enhance structural stability. For problems, such as, building structure safety and wall aging, Karve [103] used digital twin to study crack damage and maintenance, which can detect crack changes in time and perform maintenance. Rocha [75] detected structural damage by combining digital twins, physical models, and machine learning. Lydon [104] promoted the development of building energy performance through sensor data interaction and digital twin of analog thermal system in the design of thermal system of house roof structure. In terms of bridge and tunnel applications, Yu [48] established a digital twin decision analysis framework for the O&M management of tunnels, designed a COBie to OWL converter, and realized the effective integration of space-time data of digital double tunnels. Jiang [105] combined probabilistic multi-scale models and

digital twin to predict the fatigue life of steel bridges and established the fatigue crack initiation mechanism by considering the uncertainties of the material on-premise structure and random factors. Asset inspection is also very important in construction operation and maintenance projects. Establishing a sound asset lifecycle management is of great significance for ensuring the safety and integrity of fixed assets and improving economic benefits [55, 106]. Digital twin has the ability to manage and predict building facilities, as well as data analysis and decision-making capabilities. Lu [107] used digital twin technology to monitor and diagnose the operation of assets in buildings, improve detection efficiency and take timely preventive measures before causing losses.

2.5 Enabling technologies

Currently, there is no software specially designed for digital twin. In this part, enabling technologies will involve modelling from physical to virtual models, and data interaction between physical and virtual models.

In terms of modelling, the civil engineering industry pays more attention to this part. Point cloud modelling can be used for objects that will not change greatly over a certain period of time in the future, such as bridges and buildings. It is more efficient to model quickly through point clouds, but for those objects whose posture and shape change over time cannot be quickly modelled by point cloud modelling or lidar scanning, BIM models are generally used, and much BIM software are models created based on Revit. For example, Naviworks is a software based on Revit. Combined with the digital twin technology, the entire life cycle management can be carried out from the design to the construction stage, reducing costs and increasing benefits. It shows its design effect, and can carry out collision experiments and 4D simulation construction, but it has three shortcomings. One is that the rendering process requires a lot of computer configuration and takes a lot of time. The second is that the application effect of the software in large-scale projects is not satisfactory. After the model is adjusted, all components and schedule plans must be re-linked. The third is that some surfaces in Navisworks have parameter limitations. For curved surface models, Bentley BIM software should be used, but the use of Bentley BIM software will encounter interface compatibility problems. There are other simulation softwares, like the Unity3D game engine. Unity3D supports the NVIDIAPhysX physics engine, which can simulate rigid body or soft body, joint

physics, vehicle physics, path simulation, human flow simulation, etc. In a specific digital twin application, it can be completed by combining the functions of the Unity3D software.

In the present manufacturing industry, the display mode of 3D real-world recurrence is mostly static display, and the display power of the actual 3D dynamic characteristics is weak, and it is difficult to capture real-time changes and production processes of the workshop during the operation process. Jones [53] proposed that the virtual entity is a high-fidelity representation of the physical entity, and Zheng [10] used FLEXSIM to construct the virtual model of the workshop. FLEXSIM is a general-purpose discrete simulation software. The software uses a highly developed object to model and the structure is hierarchical, and has an efficient simulation engine which can be simulated and visualized at the same time. However, if there is no entity required for modelling in FLEMSIM, it is difficult to carry out secondary development. At the same time, in the digital twin framework of many manufacturing industries, the changes of the virtual entity during the production process are ignored. The virtual entity may be static and look like a photo. Practitioners are more concerned about the interaction between data and how to deal with manufacturing execution issues such as analyzing data.

The data interaction between physical and virtual models is usually completed through the Cyber physical system or the collaborative cooperation of sensors, IoT and bigdata. Big data technology collects product information from sensors, such as product operating status, real-time data about component health status, and historical data related to digital twin integration (maintenance records, energy consumption record data). After the big data technology analyses these data, the product's data transmission continues to predict the health of the product and the possibility of failure. Specifically, Figure 2.4 is the framework of data interaction between physical and virtual models which can be divided into data collection, data transmission, data storage, data processing, data visualization, and data security.

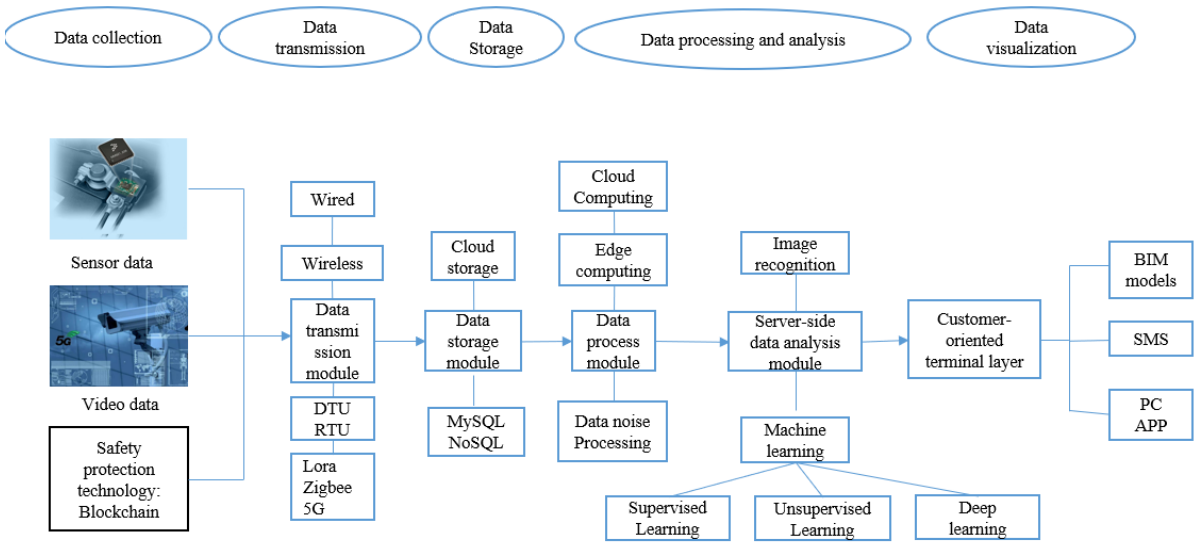


Figure 2.4: Framework of data interaction between physical and virtual models

Data collection: Collect data according to need through different methods. Common data collection methods are: video data, sensor data, radio frequency identification RFID [108, 109], QR data [110, 111], drones [112], handheld mobile devices [113].

Data transmission: mainly divided into wired and wireless data transmission, wired transmission has little interference and high security. However, the construction cost is high, and the wired layout is relatively restricted by the external environment. Wired transmission often includes cable transmission and optical fiber transmission. The overall cost of wireless transmission is low and stable, and the network is flexible and expandable. However, the use of microblog transmission is susceptible to external electromagnetic interference and is relatively affected by the weather. Data transmission can be realized through DTU (Data Transfer Unit) and RTU (remote transfer unit) to realize data collection and transmission. RTU is applied in the monitoring site and plays the role of connecting, that is, collecting sensor data and uploading it to the monitoring centre through the 4G network. Wireless transmission often includes GPRS, 3G, 4G, 5G, NB-IoT, CDMA, LoRa, Bluetooth, zigbee. GPRS and CDMA are 2G technologies, which have basically been eliminated, but they are relatively stable in long-distance wireless transmission. 5G technology [114, 115] is proposed in many digital twin frameworks, but it faces the problem of an insufficient number of 5G base stations and high cost. The MQTT protocol is a publish/subscribe message protocol designed for remote devices with poor hardware performance and poor network conditions. Due to their light

weight, simplicity, openness and ease of implementation, many digital twin applications now use the MQTT protocol [116-119].

Data storage: After data is collected and transmitted, a place is needed to store it for subsequent data processing and analysis. For instance, cloud storage [120], traditional Oracle [121], MySQL [122, 123], and the latest technology NoSQL [124] (Hbase, Cassandra) are often used. As a relational database, MySQL has the advantage of small size, high speed, low total cost of ownership, and open source. The disadvantage of MySQL is the lack of a security system and that it does not have stored procedure language. The advantages of NoSQL are its low cost, fast reading and writing, and that it is easy to expand, and the disadvantages are that the existing products are not mature enough, do not support industry standards such as SQL, and have no official support.

Data processing and analysis: In the process of digital twinning, many diversified and heterogeneous data are often collected. It is difficult to remove noise from the original data and obtain usable data. The main data analysis methods are supervised learning, unsupervised learning, and deep learning. Supervised learning uses a set of samples of known categories to adjust the parameters of the classifier to achieve the required performance. Common algorithms include the K-nearest neighbor algorithm [125] (lazy learning), decision tree [126] (through inference and analysis, input a series of data, and gradually reduce it to get a decision plan) and naive Bayes [127, 128] (based on probability theory, requires fewer parameters, and certain amount of missing data is also acceptable). Unsupervised learning is used in situations that lack prior knowledge and are difficult to manually label. Common situations include K-means clustering algorithm [129], spectral clustering, and principal component analysis [130, 131]. Deep learning provides advanced analysis techniques for processing and analyzing large amounts of data. It can also be divided into supervised and unsupervised learning. For example, image recognition and generation of confrontation network(GAN) [132, 133] are methods of supervised learning, image generation is an unsupervised learning method. Some data (such as video and image data) is directly uploaded to the cloud server, which requires a relatively large amount of traffic and has a high cost. The edge processing method is often used to preprocess the data and then upload it, thereby improving efficiency and saving costs.

Data visualization: Can be combined with BIM models, such as Navisworks. Or show results

information to customers through APP SMS or PC client terminal responses. Visualizing data through rich varieties of charts is conducive to customers' decision-making.

Data Security: Blockchain is a shared database: the data or information is stored in it, due to its high security, non-copy and forgery, traceability and other characteristics. Since 2019, it has received widespread attention, and research on the combination of blockchain and digital twins has also emerged. Li [134] proposed the establishment of a blockchain-based digital twin sharing platform to realize software copyright protection, simplify the integration of heterogeneous manufacturing resources in decentralized and distributed environments, and conduct exemplary case studies with real 3D printing scenarios as motivation to verify and evaluate the proposed platform. In real-time modelling of the Internet of Things in the digital twin, the blockchain authenticates the data exchange and establishes a digital twin blockchain framework that can trace all data exchanges. This solves the problem of limited information storage on the blockchain, and enables the data of the digital twin to be traceable [135].

2.6 Summary

At present, digital twin technology is used to solve the problems of machine health detection and to optimize industrial workshops. Through simulation-feedback execution, it has promoted the development of smart manufacturing and the digitalisation of the construction industry. Using cyber-physical simulation and the Internet of Things, the problem of real-time data transmission between physical objects and virtual models in digital twins has been solved, and many CPS-based digital twin application frameworks have been established. The combination of blockchains with digital twins solves the problem of data resource integration, improves data security, and is also a resolution for the shortcomings of insufficient data security in CPS. The development of artificial intelligence, 5G, sensors, the Internet of Things, blockchain, software, and hardware can all promote the development of digital twins to a higher level. Existing technologies can be improved in terms of algorithm optimization and the improvement of digital twin models, and further achieve the effect of improving production efficiency, resource utilisation, and optimizing the industrial chain, making the manufacturing and civil construction industries more economical and environmentally friendly.

Current digital twin technology still has the following problems, and further research is needed to provide a solution:

(1) Digital twins are faced with the problem of excessively high modelling accuracy requirements. Most studies have proposed a large framework, but only a small part of the research has been selected as a case study. Moreover, most of the current modelling methods are only for specific scenarios, so it is necessary to develop a unified, more applicable, and faster digital twin modelling method.

(2) The influence of external and human factors in the physical environment on physical entities. In the physical entities and physical environments, the influence of the external environment on the entity is considered, but how to reflect these external changes in the model requires further research. At the same time, it is also necessary to consider the impact of human factors impact the simulation. The environment and personnel are both modelled into artificial systems in the form of agents. The analytical space of environmental agents is more open, while the human agents are similar to humans. This takes the mapping of the actual system to the artificial system as its internal cognitive process and responds to the social environment

by constantly changing the internal cognitive information system.

(3) Multivariate and heterogeneous data processing. In the data exchange process of the digital twin, sensors will generate a large amount of diverse and heterogeneous data, which has the characteristics of multi-modality, high repeatability, and exists in massive quantities. Uploading it directly to the cloud will cause high traffic charges. Real-time historical data coverage has other issues, so it is necessary to process and analyse these data locally using edge computing efficiently and accurately.

In the future, digital twin is a new direction of construction industry research and a prerequisite for the construction industry to promote the progress of digitalization. The digital twin of a single building or the whole life cycle of a group of buildings requires further research. At the same time, in manufacturing, the impact of more external factors on workshops and factories will be included, and the efficiency of optimization can be improved through simulation, monitoring, diagnosing, predicting, and also controlling the state of the virtual model.

3. Framework of digital twin of tower crane

To develop a digital twin, we first need to define the physical entity and the physical environment in the virtual model through modelling and other methods. At the same time, we expect 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, the section will focus on the development of digital twin of a construction crane, by studying the 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.

Figure 3.1 demonstrates the example of digital twin process of a construction crane, by which physical entity and environment transfer relevant data to virtual environment through physical to virtual connection in real time and autonomously, while virtual entity analyses the data and then transfer execution command to physical entity through virtual to physical connection.

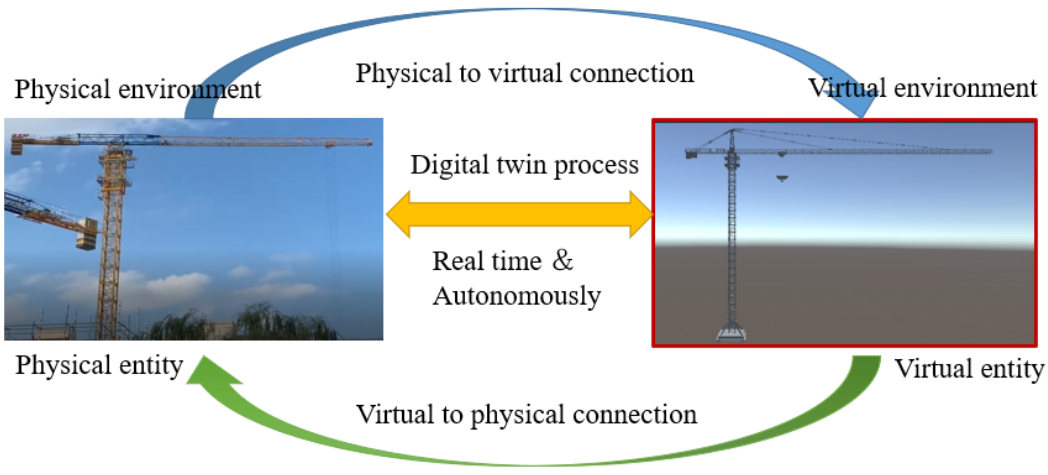


Figure 3.1: Core concepts of Digital twin

3.1 Physical entity and physical environment

In this research, the physical entity refers to the tower crane. The QTZ80 tower crane, shown in Figure 3.2, 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 3.2: 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. Creo (Pro/E) software is used to draw the components of the tower crane according to the CAD drawings of the QTZ80 tower crane, and then assemble these tower crane components in the 3D unity software. As is shown in Figure 3.3, 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.,)

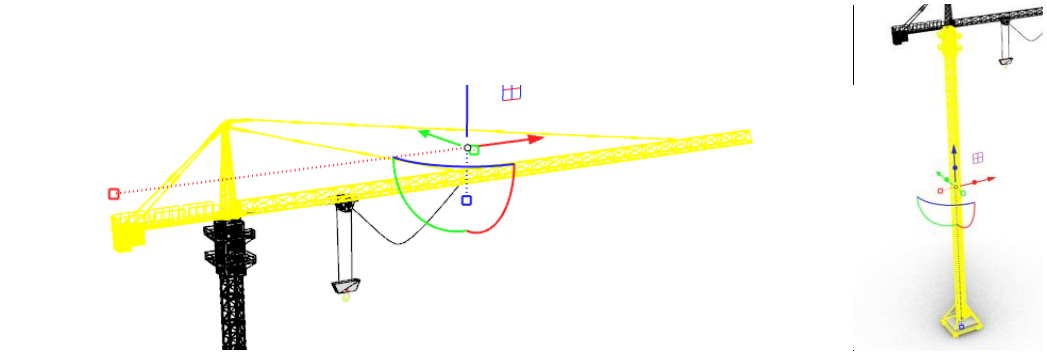


Figure 3.3: Tower crane boom and main body

3.2.1 The virtual model of QTZ80 tower crane

The QTZ80 tower crane is a new type of lifting transportation equipment designed to meet the construction and equipment installation of high -rise buildings. Its arm length is usually 60 meters, a rated maximum lifting is 6 tons, and the rated lift -to -weight torque is 80 tons*meters. It has the characteristics of extensive adaptability, convenient installation and disassembly, high work steady efficiency, and complete safety protection equipment.

Overall layout: The tower body consists of the overall structure of the standard section. The lower part of the tower is connected to the foundation through the direction of the basic section. The standard section of the tower is welded by the square steel pipe and seamless steel pipe. The section size is 1.78 meters *1.78 meters, and 2.8 meters high. The upper and lower panels are all box -shaped structures.

In the figure below, Figure 3.4 is the standard CAD drawing of the tower crane. This section will draw the standard section model of the tower crane based on this drawing. Divided into beam installation (Figure 3.5), column installation (Figure 3.6) and overall installation (Figure 3.7).

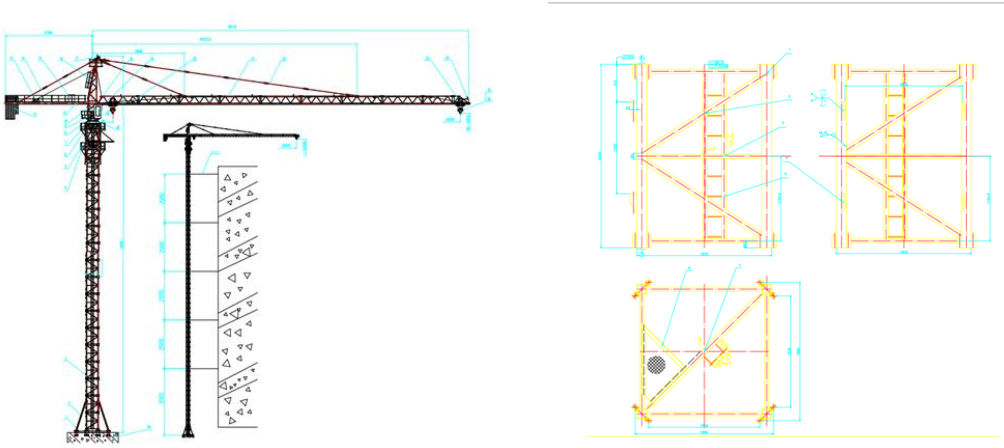


Figure 3.4: CAD drawing of QTZ80 tower crane and its section CAD drawing

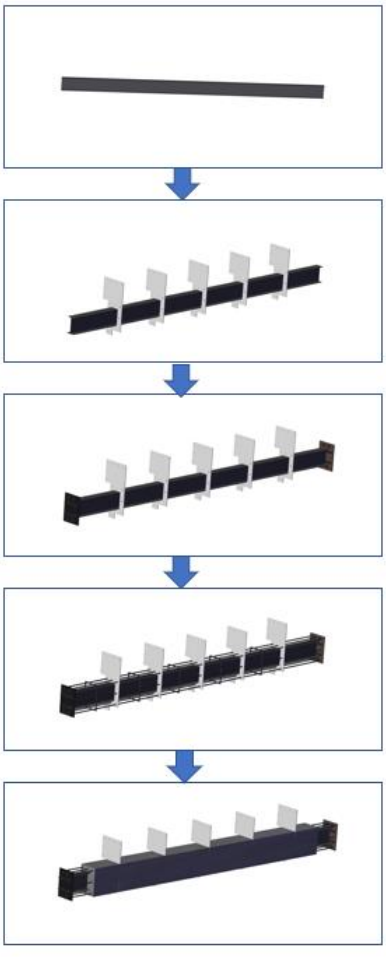


Figure 3.5: Beam installation

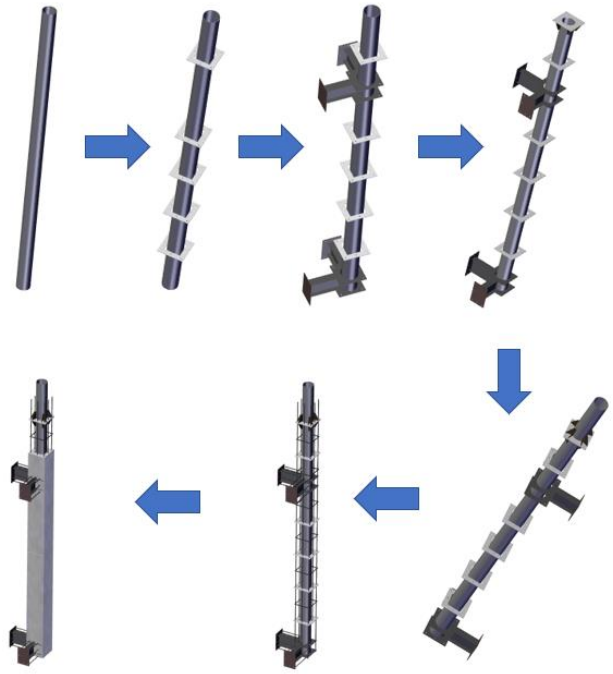


Figure 3.6: Column installation

Figure 3.7: Overall installation

After completing the above Beam Installation, Column Installation and Overall Installation, we can get the basic framework of the tower crane standard section. By connecting these frameworks, the foundation body of the tower crane can be formed, as is shown in Figure 3.8.

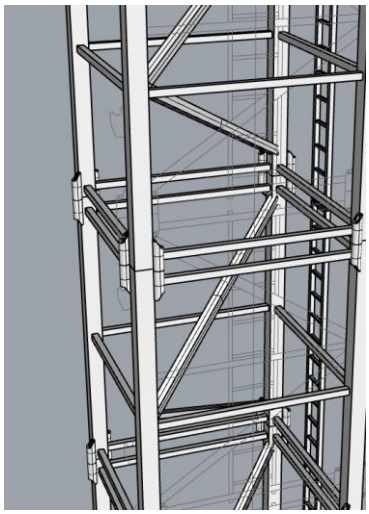


Figure 3.8: QTZ80 tower crane body

In conclusion, this section used Creo software to draw the components of the QTZ80 tower crane and assemble these components in the unity3D software. Figure 3.9 shows the original tower crane CAD drawing and the finally virtual model of the tower crane in 3D software.

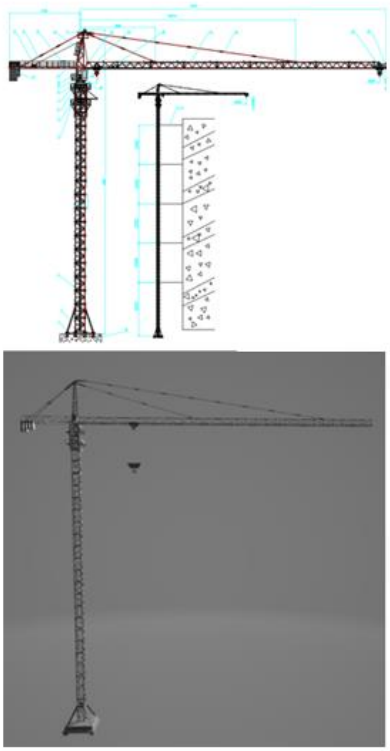


Figure 3.9: Model of QTZ80 tower crane

3.3 The physical to virtual and virtual to physical 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, 5G, IOT, CPS. In this research, tower crane object detection, operational mode recognition and modelling are used to capture the changes of tower crane and update the virtual crane accordingly.

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 or to provide decision-making opinions.

Through the physical to virtual and virtual to physical connection, 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.

3.4 The real-time and autonomous 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.

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 and optimization of tower crane image dataset

4.1 Introduction

Data is the most critical element in machine learning. To build a powerful computer vision deep learning model, high-quality datasets need to be applied during the training phase. As training algorithms are becoming more and more mature and efficient, using high-quality, large-scale image datasets maximizes the efficiency of deep learning processes such as image recognition, with the ability to train models with higher quality and greater accuracy. In the computer vision field, many high-quality datasets are available, such as ImageNet, CIFAR-10, Cityscapes, and other databases. ImageNet, for example, is one of the most popular datasets for computer vision projects, providing an accessible database of images organized according to the WordNet hierarchy. There are over 100,000 synsets in WordNet within which ImageNet provides an average of 1,000 images for each synset. It therefore provides tens of millions of cleanly ordered images for most concepts within the WordNet hierarchy. However, currently, there is no publicly available annotated image dataset showing construction site tower cranes. Collecting relevant video and image data and performing a series of optimization processing is therefore the primary task of this study.

This chapter proposes a study with the overall objective of creating and optimizing tower crane image datasets. A tower crane video image recognition dataset will be created based on this method. The main contributions to research are as follows: (1) A tower crane video image recognition dataset is established, providing sufficient and diverse training and testing data for subsequent tower crane image recognition based on convolutional neural networks. (2) The use of deep learning to determine the operation mode of tower cranes and establish a tower crane motion mode judgment dataset is examined, providing data reference for selection, design and optimization of subsequent deep learning algorithms.

The specific structure of this chapter is as follows: Section 3.2 shows the overall process of creating and optimizing a tower crane video and image dataset. Section 3.3 introduces 10 common high-quality open access datasets and sets out a scheme for the collection of tower crane image datasets. Section 3.4 introduces the methods currently used for image pre-processing, including grayscale processing and binarization. Using these two processing

methods, irrelevant information in the images can be removed, and data is simplified to the greatest possible extent, so that actual, useful information can be restored and emphasized, ultimately enhancing the detectability of key information in images. Section 3.5 introduces the common labelling methods and common labelling tools for image annotation and describes the labelling rules for tower crane images applied in this study. Section 3.6 introduces the tower crane segmentation algorithm and mode recognition dataset. Tower crane segmentation, as a key step connecting tower crane identification and tower crane motion status identification, is a significant part of this study. Section 3.7 performs data augmentation based on existing datasets to expand the amount of data, thereby reducing the possibility of overfitting. The mosaic enhancement method is used in the image recognition deep learning algorithm, while in the operation state recognition deep learning algorithm, tower crane images are rotated to augment the dataset. Section 3.8 introduces the tower crane dataset created for this study, which uses these methods for data acquisition, data pre-processing, data annotation, and data enhancement. Section 3.9 provides a summary of the content of this chapter.

4.2 Process framework

Under the premise that computing power and algorithms have been basically determined and are relatively mature, data plays a decisive role in the process of realizing general deep learning tasks. It can be argued that the lower limit of the impact of data on the final effect of deep learning tasks is very low, and that the upper limit is very high. This is to say that, if there is a problem with data quality, a good model can be wasted. However, if the quality of the data is very high, a mediocre algorithm model can be applied with extraordinary effect. The focus of this section is therefore establishing a large-scale tower crane image dataset and optimizing data processing. As shown in the Figure 4.1, this research includes tower crane image recognition, tower crane segmentation, and tower crane motion status recognition. A large-scale, high-quality dataset is the basis of these deep learning functions and therefore for improving the accuracy of deep learning recognition.

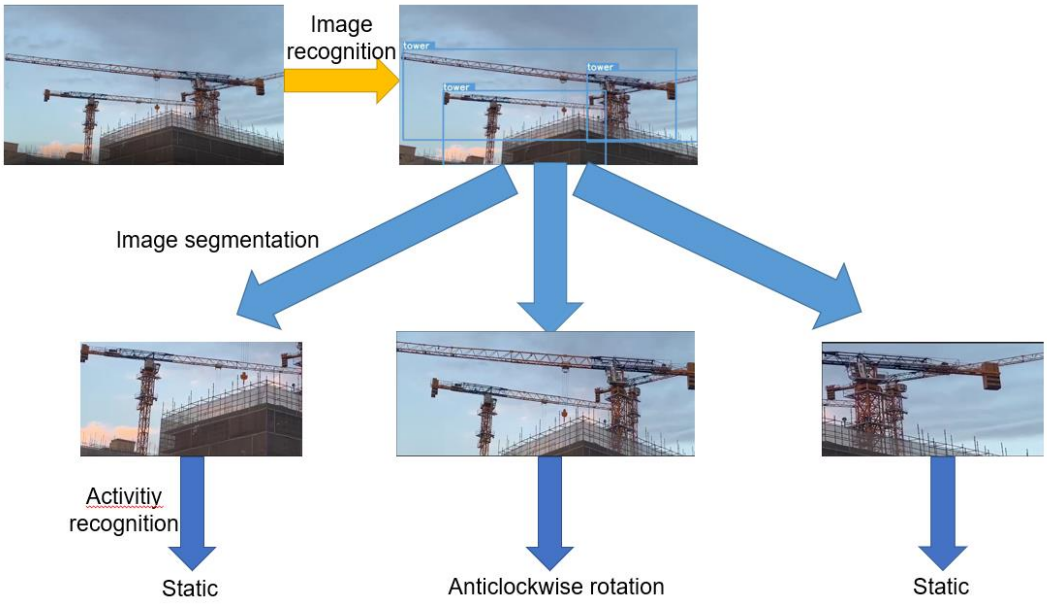


Figure 4.1: Framework of operational mode recognition in this research

This section divides the creation and optimization of tower crane image datasets into video data collection (datasets with small quantity and medium quality data), data pre-processing (datasets with small quantity and high-quality data), and image labelling (making the data with labels). Tower crane segmentation (connecting deep learning for image recognition and deep learning for motion state recognition), data enhancement (large quantity and high-quality datasets that can be used for deep learning). As a result, an initial data set with small volume and medium quality is turned into a training data set with large volume and high quality. The subsequent sections of this section will explain this in detail.

4.3 Data collection

Datasets are the core of deep learning. High-quality, large-scale datasets can effectively improve the accuracy of training models. At present, academics, as well as some high-tech enterprises, have developed deep learning datasets, of which there are now many. The preparation of data is a necessary task before training the model and is obviously very time-consuming. The following Table 4.1 shows 10 popular high-quality open access datasets, including datasets that integrate object classes, faces, human behaviour, etc., and introduce them according to name, object type, and annotation features. Rich datasets also play a positive role in the development of algorithms. However, so far, there is no public or generally available image dataset with category labels for construction site tower cranes. Collecting relevant video

and image data and performing a series of optimization processes to generate such a dataset is therefore the primary task of this research.

Table 4.1: 10 popular high-quality open access datasets

Name	***Type***	***Features***
CIFAR-10	Object recognition	60,000 32×32 color images in 10 categories
Cityscapes	Semantic City Scene	5,000 frames of high-quality pixel-level annotation, and 20,000 weakly annotated frames
Fashion MNIST	Clothing dataset	10 categories, 60000 training dataset, 28×28 grayscale image
ImageNet	Common Objects	over 100,000 synsets
IMDB-Wiki	Human face dataset	523051 human face images
Kinetics-700	Human pose	700 categories, each contains at least 600 video clips
MS Coco	Common Objects	91 categories, 2.5 million labeled instances in 328k images
MPII Human Pose	Human Pose	25K images, more than 40k labels, 410 human activities
Open Images	Common Objects	16 million bounding boxes for 600 object classes on 1.9 million images
The 20BN V2	Human Pose	220847 videos of human activities

Source: *http://news.51cto.com/art/202001/609142.htm*

The video data used in this study were mainly from google videos and on-site tower crane operation videos. Some low-quality images and videos have been filtered out. Each video only contains a single state of the tower crane: stationary, rotating clockwise, or rotating counter clockwise. Adobe Premier Pro CC was used to divide each video into individual frames. In order to ensure the diversity of data sources, the collected video had to meet two requirements as far as possible:

(1) The tower crane videos are all different; they are from different construction projects, the backgrounds of the videos are different, the environment and weather are different, and they have different video styles.

(2) the videos contain as many different types of tower cranes as possible, with different attitudes and different geographical locations, which gives the tower cranes in the samples a wide range of expressions.

4.4 Image preprocessing

4.4.1 Grayscale and thresholding pre-treatment

Image pre-processing is suitable for the eradication of superfluous image-related data. It also reduces the complexity of information, reinstates and highlights pertinent data, and renders critical data easier to locate. Thus, image pre-processing can ensure that feature extraction, image segmentation, and image recognition are more accurate and dependable. Common image pre-processing methods include Grayscale processing, thresholding processing, histogram processing, and image filtering, alongside advanced cutting-edge detection techniques. The current research employs grayscale processing and binarization processing for the image pre-processing of tower crane operating status recognition. These techniques both facilitate data simplification, whereby the processed images can be used to enhance the training accuracy of the machine learning algorithms.

It is possible to adopt image grayscale processing as a pre-processing stage in overall image processing management, in advance of the ensuing upper-level operations that include the segmentation, recognition, and analysis of images. Grayscale indicates the colour depths of the points that exist between black and white, thereby calibrating the colour value attached to each pixel in the grayscale image. Thus, grayscale typically ranges between 0 and 255, with 0 signifying black and 255 denoting white. The grayscale value indicates the shade of the colour, whereas the gray histogram represents the number of pixels that have a grayscale value that matches each grayscale value in a digital image. Image processing tends to process the three components of RGB is isolation. Moreover, the morphological characteristics of the image are not reflected in RGB. Rather, RGB only modifies colours in accordance with optical precepts. Grayscale processing comprises the conversion of colour images into grayscale images. RGB colour image consist of three elements, namely: red, green, and blue. The images are displayed as groups of these three. Grayscale transforms the r, g, and b colour elements so that they are equal, providing the three RGB quantities are equivalent. For example, RGB (100, 100, 100) means that the grayscale value is 100. Hence, increased grayscale values correlate with brighter pixels. A small grayscale values is manifested in darker pixels. Grayscale images can be achieved via multiple techniques:

1. The component technique

The grayscale value of the three grayscale images comprises the image brightness of the three

colour image elements employed as the grayscale,with one grayscale image being chosen in accordance with the requirements of the application.

2. The maximum value technique

The maximum value of the three-component luminance in the colour image has a maximum value that is employed as the scale of the grayscale image.

3. The average technique

The colour image has a three-component luminance that can be averaged in order to acquire a grayscale values.

4. The weighted average technique

The three elements are weighted and averaged with different weights in accordance with their comparative significance and further indicators. This is because human vision exhibits greater sensitivity to green and less to blue. The grayscale image produced by this method is more suitable because it is secured by balancing the averages of the three R, G, Ba elements.

This study uses the weighted average method to grayscale the RGB image:

$$\mathrm{Gray}(i,j) = 0.299 \times \mathrm{R}(i,j) + 0.587 \times \mathrm{G}(i,j) + 0.114 \times \mathrm{B}(i,j) \qquad \text{Eq. (4-1)}$$

Among them, Gray (i, j) represents the gray value of the pixel, R (i, j), G (i, j), B (i, j) represent the pixel in the red, green and blue channel respectively.

Once the image grayscale processing has been performed, the tower crane images require binarization. Specifically, image binarization is a critical aspect of image pre-processing. Following the binarization of an image, additional denization and feature processing can be conducted. An overly diminutive threshold denotes the inclusion of noise in the target, wherein deep learning judgment is impacted, and training ability is restricted. Image binarization allows the grayscale values of a pixel to be fixed as either 0 or 255. Hence, the resultant image is monochrome because each pixel of a binary image is characterized by one of two possible values, namely: pure white or pure black. The simplicity of binary image data is simple has caused numerous vision algorithms to be dependent upon binary images. Binary imaging not only represents a more valuable approach to the analysis of shapes and contour, but also has the ability to mask areas of the original image that are neither interesting nor valuable. Global Thresholding is the most popular of the numerous binarization methods. It can be represented as follows:

Use a fixed threshold uniformly for all pixels in the input image:

$$g(x,y) = \begin{cases} 255, & \text{if } f(x,y) \geq T \\ 0, & \text{else} \end{cases} \qquad \text{Eq. (4-2)}$$

Among them, g (x, y) is the grayscale values of the point after binarization, f (x, y) is the grayscale values of a pixel in the image, T is the gray threshold, which belongs to the hyperparameter value, which can be obtained by Otsu's method.

In the Figure 4.2 below, after removing some noise, the images from left to right show the results of grayscale processing and binarization. The step of image preprocessing recovers useful real information in the image, enhancing the detectability of key information. It is beneficial to the 3Dresnet algorithm in extracting information and improving training accuracy.

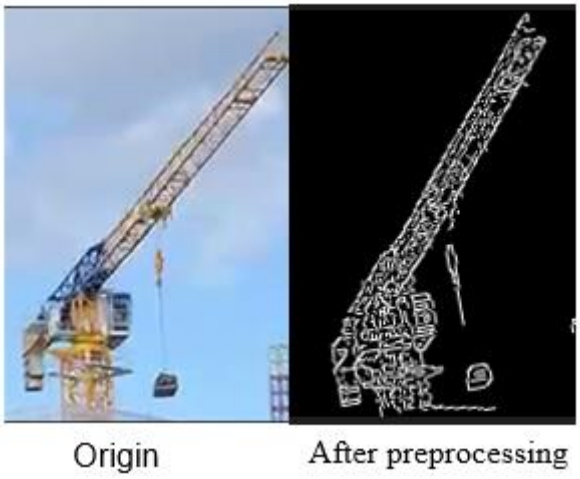


Figure 4.2: Sample of image preprocessing

4.4.2 Optical flow

Optical flow denotes the instantaneous speed of the pixel movement on the imaging plane. Thus, extremely small-time intervals, as can be found between two concurrent video frames, equates to target point displacement. Changes in the time domain of the image sequence pixels and the relationship between bordering frames can be employed to ascertain the agreement between prior and present frames, thus facilitating the calculation of object movement between adjoining frames.

The optical flow technique is predicated in two principal suppositions, namely: (1) the

inconvenience of constant brightness, (2) changes in time do not cause drastic changes in the pixel positions. In this dataset, the duration of the videos taken by the tower crane was between 7-15 seconds. During this period, the brightness will basically not change, which conforms to the first assumption. Since the rotation of the tower crane is a slow process, the displacement between frames is also quite small, in line with the second assumption.

In the 20 frames of pictures divided by the tower crane, one picture was selected from every five frames as an interval, including the beginning. The end of each tower crane contained five pictures. The optical flow method was used to process these five pictures, the result of which is provided in Figure 4.3. Below, Figure 4.4 reflects the small optical change of each frame. Each optical flow result was then stored in the new crane folder.

Figure 4.3: The status of tower crane in 5 frames

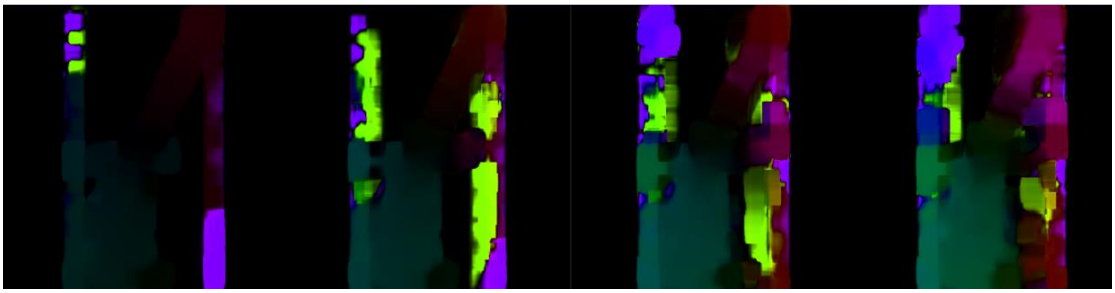


Figure 4.4: Optical flow of changes in each frame

4.5 Image annotation

Data annotation is an important link in the realization of artificial intelligence. Recognizing

tower crane images and operating status through deep learning models and algorithms requires the use of large volumes of image data for algorithm training. Artificial intelligence requires more human intervention than is generally realized. The successful operation of artificial intelligence requires a lot of manual work, and image labelling is one such task which cannot be done automatically. Clean and comprehensible data must be entered into the system for it to function as required. When it comes to images, computers need to see what the human eye sees. To prepare high-accuracy training data, images have to be annotated to obtain the correct results. The training involved in deep learning is not simply inputting pre-processed images into the model, but also telling the neural network what objects particular areas of the images represent. Then, through training and learning using large amounts of image data, the logical association between the image features and the object ontology is established. The data will then be effective for training deep learning neural networks.

4.5.1 Common annotation method

In target detection, different annotation methods need to be used according to different requirements. Common annotation methods can be divided into the following categories: point annotation (face recognition, gesture recognition), line annotation, bounding box annotation (2D, 3D), pixel annotation, and semantic annotation.

Point annotation (Figure 4.5): Point annotation is also called key point annotation. Key points are manually marked at specified positions. Point annotation is generally used for such applications as human skeleton recognition, face feature point recognition, etc. In the Figure 4.5, landmark annotations are performed on the eyebrows, eyes, nose, lips, and cheeks of the face. The point annotation needs to be as detailed as possible, comprehensively marking out the details of the target to permit accurate identification and authentication.

Line annotation (Figure 4.6): Line annotation is used in the field of autonomous driving to identify lanes. Straight lines and curves of different colours are used to mark lane boundaries. For example, red, green, blue, and yellow are used to mark fast lanes, overtaking lanes, emergency lanes, etc. The driving system uses these indicators to identify and precisely locate lane boundaries to avoid situations such as lane congestion, inadvertent lane changes, etc.

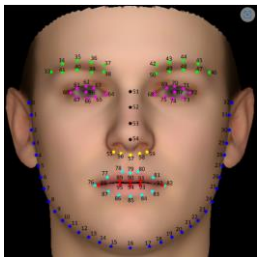


Figure 4.5: Point annotation

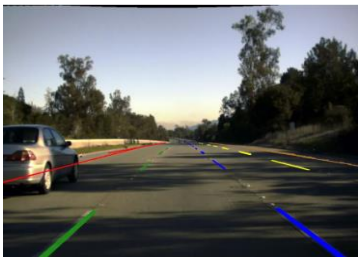


Figure 4.6: Line annotation

Bounding box labelling (Figure 4.7). The bounding box labelling method is used for target object detection, identifying the specific position of an object with certain characteristics in the image. It is divided into 2D labelling and 3D labelling according to different application scenarios and requirements. The bounding box used in 2D annotation is a rectangle around the target object, sized by its upper-left and lower-right corners. In some static images, in order to accurately label the target, polygonal labels are used for irregular objects. Based on 2D annotation, 3D annotation also reflects the three-dimensional 'thickness' of an object, usually representing the object as a cuboid although, in some special cases, the 3D annotation will be an irregular polygon. As shown in the figure below, 3D annotation completes the drawing of polygons by marking multiple control points.

Figure 4.7: Bounding box labelling (2D & 3D)

Semantic annotation (Figure 4.8). For example, the Citescapes dataset mentioned in the previous Section 4.3 is a semantic understanding of urban street scenes, which is widely used in traffic control for autonomous vehicles, including the handling of vehicle breakdowns, pedestrians, obstacles, sidewalks, motor vehicles, buildings, etc.

Pixel annotation is a type of semantic annotation (Figure 4.9). The traditional bounding box annotation has a number of problems—such as requiring a large number of datasets, noise

around the object in the annotation box, and difficult-to-detect objects that are occluded. These problems can be solved by accurate pixel labelling. The most commonly used tools for pixel annotation rely heavily on slow point-by-point object selection tools, where the annotator must pass through the edges of the object. Pixel labelling is therefore highly labour-intensive and is accordingly expensive.

Figure 4.8: Semantic annotation

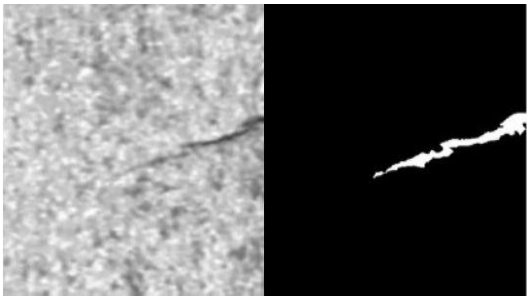


Figure 4.9: Pixel annotation

4.5.2 Common annotation tools

Data annotation is a key link in the operation of most artificial intelligence deep learning algorithms. To realize artificial intelligence, the ability to understand and detect objects must be taught to computers, just like human beings, so that AI can learn to recognize and understand in a similar way to humans. Table 4.2 introduces six common image annotation tools and describes their features, annotation frameworks and output formats. This research uses the well-known annotation tool, Labellmg, to manually annotate the images of tower cranes.

Table 4.2: Examples of annotation tools and its features, frame and output format

Name	Features	Frame	Format
LabelImg	Well-known annotation software	Rectangle	XML
Labelme	Support object detection, image semantic segmentation data annotation	Rectangle, polygon, line, point, circle	VOC, COCO
RectLabel	Image annotation tool	Rectangle, polygon, polyline, point	YOLO, KITTI, COCOJSON, CSV
CVAT	Support image classification, object detection frame, image semantic segmentation	Rectangle, polygon, line, point, tag	COCO, PASCAL, VOC
VOTT	WEB-based visual data annotation tool, support image and video data annotation	Rectangle, polygon, polyline	Tensorflow, VoTT, CSV, CNTK
LabelBox	Suitable for labeling of large projects, support image, video and text annotation	Rectangle, polygon, line, point, nested classification	JSON

4.5.3 Rules of tower crane annotation

The tower crane images in the dataset have obvious 2-dimensional features. Therefore, this study uses the 2D rectangular frame annotation method for image annotation. When annotating tower crane images, the following rules must be followed. (1) The size and position of the labelling box should be appropriate. The rectangular labelling box should just surround the target, with the gap size within a reasonable range, and contain the outline information of the labelling target. (2) In order to improve the versatility of the model, occluded targets and some small targets should also be marked. For example, some tower cranes are partially blocked by buildings and trees, but it is obvious that they are tower cranes, so they should be marked appropriately. Figure 4.10 shows the example of tower crane annotation. In these images, the left image represents the correct way of annotation which meets the two requirements mentioned above. From the image in the middle, we can see that the labelling image is apparently larger than the tower crane itself, and in the right image, the labelling box is smaller.

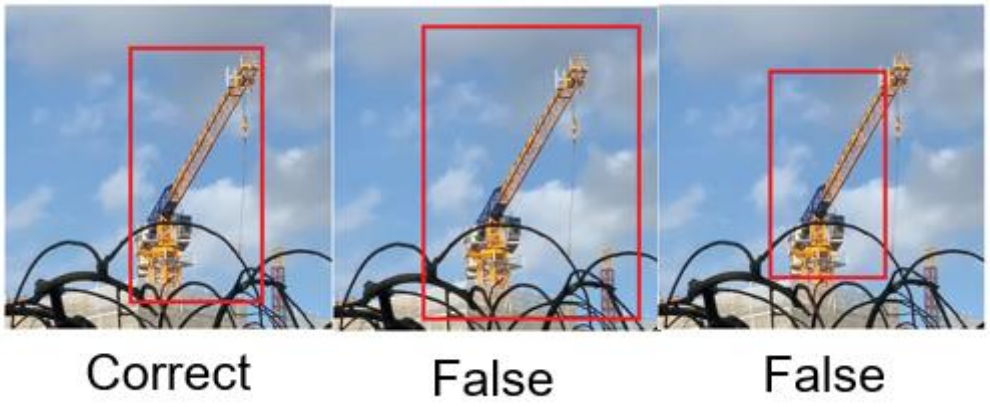


Figure 4.10: Example of tower crane annotation (Correct, False, False)

4.6 Tower crane segment and state recognition dataset

Tower crane image recognition is the first step in tower crane operation mode recognition. Once tower crane image recognition 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 for use in this study and a set of segmentation tower crane algorithms 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 4.11. 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.

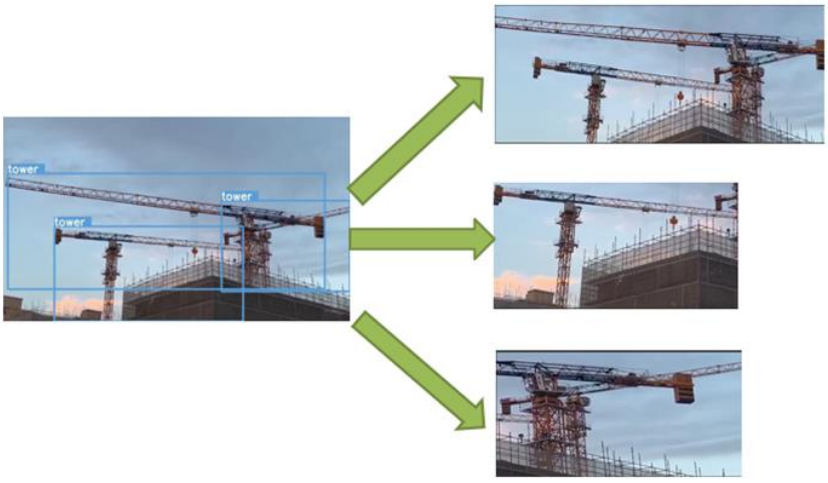


Figure 4.11: Tower crane segment

However, there are some cases where two tower cranes are in the same image frame, as shown in Figure 4.12. 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 4.12: Example of two tower cranes overlapped in one image

Accordingly, for each group of tower crane datasets, some cases of overlap or duplication are manually filtered out. However, in this study, some samples with two or more tower cranes are retained, as this increases the robustness of the model. In 5-20s videos, there are often hundreds of image frames, with 20 of them being filtered and taken out. For the operating state of the tower crane described by these 20 frames, the motion state of the tower crane is indicated, (0 is static, 1 is clockwise rotation, 2 is counter-clockwise rotation). For example, tower 2 in test3 is rotating clockwise, and is thus marked as test3/tower2 1.

Figure 4.13 is the core part of the tower crane segmentation algorithm. The algorithm is divided into 3 steps: (1) Normalization process. The normalization algorithm is used to process the data and limit it to the required range for this research. In this study, the role of normalization is reflected in image processing. The area that was originally difficult to divide on a map can be easily given a relative position after normalization. (2) The second step is the modelling process. Here, the model trained by the yolov5 algorithm can quickly find the position of the annotation frame in the video image. (3) The third step is to generate the images, separating the tower crane element in the image and saving it in the corresponding folder for subsequent image data processing. Figure 4.14 shows the python codes of tower crane segment algorithm.

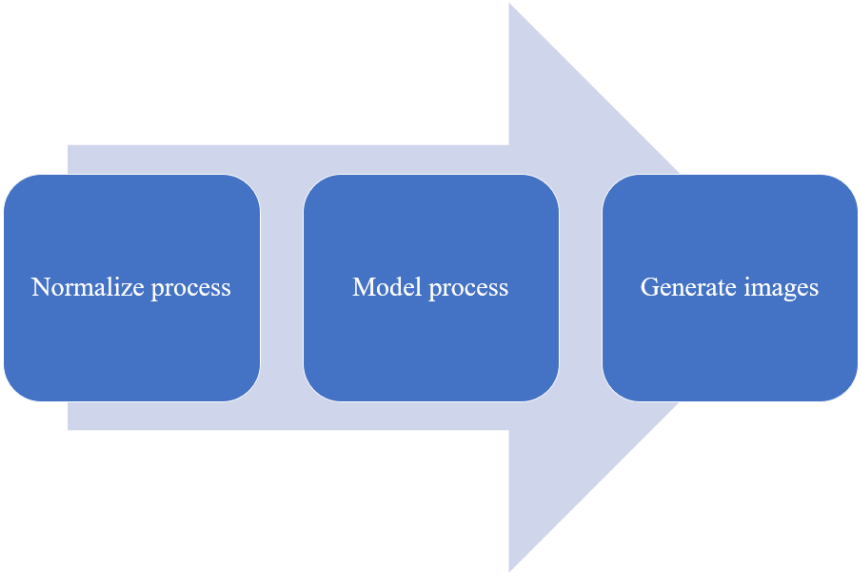


Figure 4.13: Flow diagram of the tower crane segment algorithm

```
def detect(self, frame):
    im0 = frame
    img = letterbox(frame, new_shape=416)[0]

    img = img[:, :, ::-1].transpose(2, 0, 1)
    img = np.ascontiguousarray(img, dtype=np.float32)
    img /= 255.0

    img = torch.from_numpy(img).to(device)
    if img.ndimension() == 3:
        img = img.unsqueeze(0)
    pred = model(img)[0]
    pred = non_max_suppression(pred, opt.conf_thres, opt.nms_thres)

    boxes = []
    confidences = []
    classIDs = []
    imgs = []

    for i, det in enumerate(pred):

        if det is not None and len(det):
            det[:, :4] = scale_coords(img.shape[2:], det[:, :4], im0.shape).round()

            for *xyxy, score, cls in det:
                label = '%s ' % (names[int(cls)])
                #plot_one_box(xyxy, im0, label=label, color=colors[int(cls)])
                imgs.append(get_tower(xyxy,im0))
                # ------------------------------------------------------------------------------------

                boxes.append([int(xyxy[0]), int(xyxy[1]), int(xyxy[2] - xyxy[0]), int(xyxy[3] - xyxy[1])])
                confidences.append(float(score))
                classIDs.append(int(cls))

        return imgs
```


Figure 4.14: Python codes of tower crane segment algorithm

After the tower crane is segmented, there are often hundreds of image frames in the 5-20s videos. In this study, 20 frames were extracted by manual filtering and the images of these 20

frames stored in the test folder in an irregular manner, for example: tower0_1, tower0_20, tower 0_39. For subsequent processing, the suffixes of these pictures need to be uniformly modified. The figure below is the algorithm for numbering the images 0-19. Finally, the 20 images in the test folder will be stored with suffix numbers 0-19. (e.g., tower0_0, tower0_1, tower0_2, … tower0_19). Figure 4.15 is the tower crane number algorithm.

```
dir = './temp'

pattern = re.compile(r'\d+')


for item in os.listdir(dir):
    files = []
    seq = []
    for file in os.listdir(dir+ '/' +item):
        files.append(file)
        head = file.split('_')[0]
        result = pattern.findall(file)[1]
        seq.append(int(result))
        seq.sort()

    for i in range(0,len(seq)):
        old = dir + '/' + item + '/' + head + '_' + str(seq[i]) + '.png'
        new = dir + '/' + item + '/' + head + '_' + str(i) + '.png'
        try:
            os.rename(old, new)
        except:
            print(old)
```

Figure 4.15: Tower crane numbering algorithm

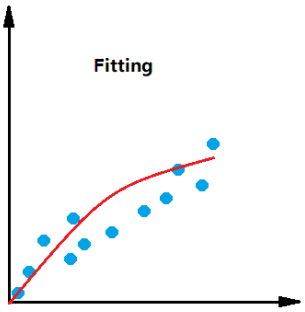
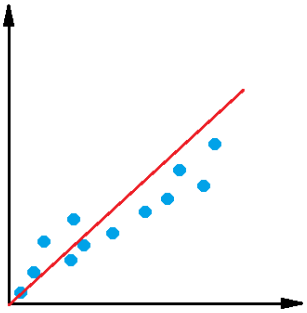
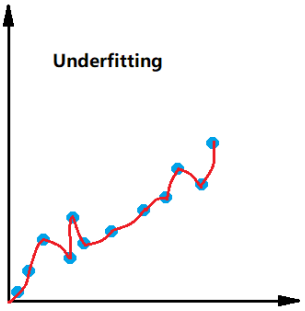
4.7 Data augment

The results of deep learning are often expressed in the form of fitting curves, and there are three usual cases—fitting, overfitting, and underfitting—according to different situations. The figure below shows the characteristics of fitting, underfitting, and overfitting curves. Underfitting indicates limited model complexity and the inability of the model to secure a sufficiently low training set error in respect of the rules underlying the data. However, it is typically possible to resolve these issues by rendering the network more complex through the addition of new properties. Overfitting is indicative of an excessive gap between the training error and the test error whereby the model recalls the properties characteristics that cannot be

applied to the test set. Moreover, the model is unable to comprehend the fundamental data properties and has limited generalizability. The situations that generally lead to overfitting are: (1) The training set samples are single and insufficient. For example, in the motion pattern recognition of tower cranes, the amount of data for the clockwise motion of tower cranes is much larger than the amount of data while in counter-clockwise rotation or static. Therefore, for this study, pattern recognition training was concentrated. The tower crane rotates clockwise, or counter-clockwise, or is stationary and the ratio of data for each condition is close to 1:1:1. (2) Noise interference is too large during the training process. For this study the interference data in the training data was eliminated as far as reasonably possible by manual filtering. (3) The model is too complicated. Acquiring and using more data (dataset augmentation) is the fundamental approach to solving the problem of overfitting of the trained model. Regularization is also a commonly used method to prevent model overfitting, by adding a regularization term to the target equation, and using a regularization coefficient to balance the target equation and the regular term.

However, unlike traditional computer vision projects, which have rich, complete and large-scale databases that can be used for training and testing of image recognition and operation pattern recognition, there is no ideal, available video image data set for tower crane operation at construction sites. Therefore, considering the difficulty of video collection, as well as the labour and time costs of manually annotating tower crane images, the generalization of computer vision neural networks can be improved to avoid overfitting caused by insufficient training datasets.

Table 4.3: Definition and curve of fitting, underfitting and overfitting

Curve	Training data	Test data	Figure
Fitting	GOOD	GOOD	
Underfitting	POOR	POOR	
Overfitting	*GOOD*	*POOR*	

Generally speaking, the purpose of data augmentation is to create more similar data from existing, limited data without substantially increasing the data volume, thereby improving the generalizability of the model. Data augmentation can be divided into supervised data augmentation and unsupervised data augmentation. Supervised data augmentation includes single-sample data augmentation and multi-sample data augmentation. This study mainly involves single samples of tower cranes and single-sample data enhancement includes geometric transformation (flipping, rotation, cropping, scaling, etc.); colour transformation (noise, blur, colour transformation, erasure). Figure 4.16 shows some single-sample data enhancement methods. There are generally two approaches to unsupervised data augmentation. (1) GAN (generative adversarial network), which learns the distribution of data throughout the model and randomly generates pictures that are consistent with the distribution of the training data set. (2) AutoAugment, which can learn a suitable data enhancement method for

the current task automatically through the model.

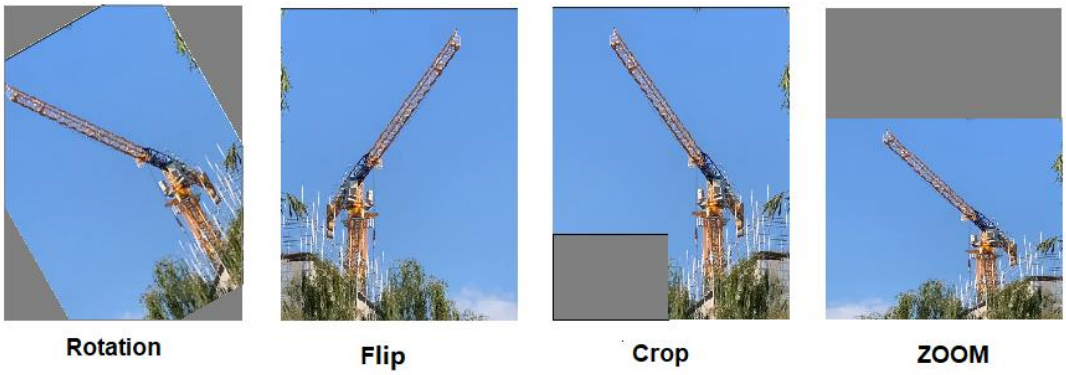


Figure 4.16: Single-sample data enhancement

The Yolo series algorithm proposes a mosaic data augment method which refers to and improves the Cutmix data augment. This method randomly crops four different images and then stitches them into one image as training data. Thus, a rich picture background is obtained, and batch size is improved to a certain extent. Mosaic data augmentation is divided into 3 steps.

(1) Randomly select 4 images from the original dataset (Figure 4.17)

Figure 4.17: Select images from the dataset

(2) Figure 4.18 shows the second step: performing random single-sample data enhancement on each of the four pictures (flip, rotate, crop, zoom, colour transformation, etc.). Then place the images according to the original image size, with the first image placed in the upper left

corner, the second image in the lower left, the third in the lower right, and the fourth in the upper right.

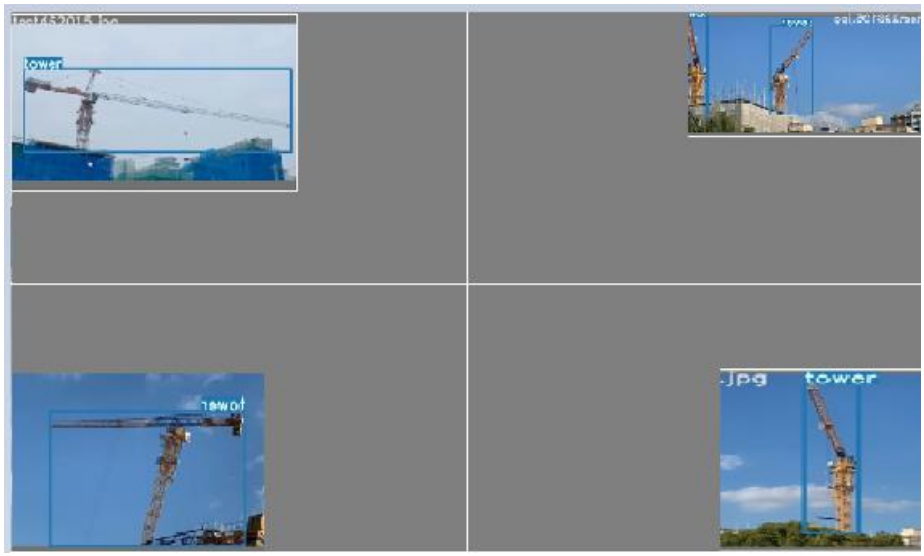


Figure 4.18: perform random single-sample data enhancement

(3) Figure 4.19 shows the third step in which the fixed area of the four images is intercepted by the matrix algorithm, and the four images are spliced into a new image.

Figure 4.19: Combination of pictures and boxes

(4) After transforming the image, the final output is obtained, and the corresponding code (Figure 4.20) is as follows:

```
img4, labels4 = random_affine(img4, labels4,
                              degrees=self.hyp['degrees'],
                              translate=self.hyp['translate'],
                              scale=self.hyp['scale'],
                              shear=self.hyp['shear'],
                              border=-s // 2)  # border to remove

return img4, labels4
```

Figure 4.20: Image transformation code

4.8 Creation of tower crane dataset

4.8.1 Data collection and preprocessing

This study developed a dataset that includes a video set, an image set, and an annotation set. The video is mainly composed of google video and the tower crane operation videos shot on site, comprising a total of 583 videos with a length of about 7-15 seconds each. The videos only contain a single state of the tower crane: static, or clockwise rotation, or counter-clockwise rotation. Adobe Premier Pro CC software was used to divide each video into frames and output them in a JPG format, producing a total of 45588 images. Due to the small changes in the tower crane during rotation operations and while stationary, there were some low-quality videos and noisy images on the network which were removed by manual screening. Finally, 8545 images were selected for labelling using the target detection labelling tool, Labellmg, to label the images, which were then stored in XML format. Ultimately, more than 20,000 tower crane images were labelled.

In the dataset of tower crane operation pattern recognition, this study segmented the tower cranes in the images using the tower crane segmentation algorithm and saved them in different folders. Most images where the crane trajectories overlap was manually filtered out, although some with two or more tower cranes were retained to increase the robustness of the model. Finally, 1373 sets of data were obtained, including 27460 tower crane images which were divided into 1167 sets of training datasets, 119 sets of validation datasets, and 87 sets of test datasets.

4.8.2 Image annotation

After completing image acquisition and preprocessing, Labellmg was used to label the tower crane images for subsequent deep learning of image recognition. The images were labelled according to the tower crane labelling rules set out in Section 3.5. A labelled image is shown in Figure 4.21. In the images, the tower crane is closely enclosed by a rectangular frame which covers all parts of the tower crane according to the situation. Some image noise is also included, such as leaves, buildings, etc.

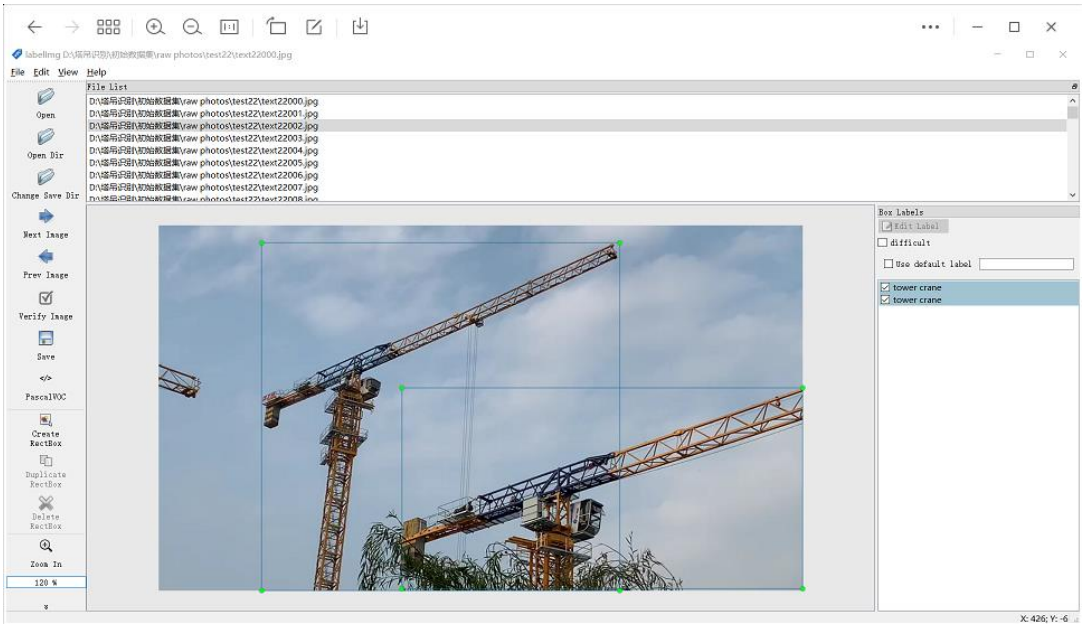


Figure 4.21: Tower crane annotation using Labellmg

Section 4.5.2 describes how the output format of the results annotated by Labellmg is an xml file. However, the yolov5 algorithm used in image recognition in this study requires a txt file. It is therefore necessary to convert xml files to txt format in batches using python. The Figure 4.22 below shows the python code generated for conversion from xml format to txt format.

```
if i == 4:
    for item in list(children_node)[i]:
        location.append(item.text.strip())

    num_1 = int(location[0]) / W
    num_2 = int(location[1]) / H
    num_3 = int(location[2]) / W
    num_4 = int(location[3]) / H

    center_x = num_1 + ((num_3 - num_1) / 2.0)
    center_y = num_2 + ((num_4 - num_2) / 2.0)

    print(int(location[0]) / W)
    output.write('0' + " "
                 + str(center_x) + " "          #Indicates the X coordinate of the labeled crane
                 + str(center_y) + " "          #Indicates the y-coordinate of the labeled crane
                 + str(num_3 - num_1) + " "     #Indicates the target frame width of the labeled crane
                 + str(num_4 - num_2) + "\n")   #Indicates the target frame height of the labeled crane
```

Figure 4.22: Xml to text file algorithm

Here, we take the numbers in one of the txt files.
0 0.06953125 0.2744565217391305 0.13593750000000002 0.5434782608695653

There are five numbers in total. The first digit: "0" represents the label, that is, the tower crane. In the Labellmg annotation tool, different objects in the same image can be individually classified and labelled, and their names can be selected after completing the labelling box. In this study, tower crane is the first label, so "0" is used to represent it. The second digit: "0.06953125" indicates the x coordinate of the labelled crane. The third digit "0.2744565217391305" indicates the y coordinate of the labelled crane, the fourth digit "0.13593750000000002" indicates the target frame width of the labelled crane, and the fifth digit "0.5434782608695653" indicates the target frame height of the labelled crane. Through the processed txt files, the data can be directly extracted by the yolov5 algorithm for deep learning.

4.8.3 Data augment

After completing image data acquisition, preprocessing, and labelling, data enhancement and augmentation were carried out to expand the tower crane image dataset. Mosaic enhancement was used in the image recognition deep learning algorithm. Data augmentation was performed on the tower crane image dataset using Mosaic data augmentation and traditional single-sample data augmentation methods (flip, rotate, crop, scale, colour transform, etc.). After data

enhancement was completed, the original data and the enhanced data were scrambled and mixed into a new dataset, increasing the volume of the data, with the new dataset providing enough data to support parameter optimization in the model, thereby improving the generalizability of the deep learning models. In 3DResnet, a data augmentation method using rotated images was applied. In the model training of tower crane operation attitude recognition, in order to increase the training dataset, the dataset was made as diverse as possible so that the trained model has stronger generalizability. Improving the relevant data in the dataset through data augmentation can prevent the network from learning irrelevant features, encourage it to improve data-related performance, and significantly improve the overall performance. In this study, the tower crane was rotated 10° and 20° clockwise and counter-clockwise. As shown in the Figure 4.23 below, the tower crane was rotated by 10 degrees and -10 degrees, and the original image of the tower crane by 20 degrees and -20 degrees. These operations increased the number of datasets from 1373 to 6865, a factor of five, which will improve the overall performance of the subsequent model.

Figure 4.23: Dataset augment

4.9 Summary

Image data is the core of deep learning. For deep learning models for supervised learning and semi-supervised learning, the outcome of training depends not only on the depth and quality of the algorithm itself, but also on large-scale, high-quality datasets. To address the current

lack of tower crane image datasets in the engineering field, this study established a dataset for tower crane image recognition and tower crane operation status recognition using image acquisition, image preprocessing, image annotation, tower crane segmentation and data enhancement. In the tower crane image recognition dataset, there are 583 tower crane operation videos and a total of 45588 tower crane images. In the data set for tower crane operating status recognition, there are 1373 groups of data and a total of 27460 images. This solves the problem of the current small quantity and poor quality of tower crane data sets in for use in image recognition during the construction process. Deep learning algorithms can now be developed based on this dataset.

5. Tower crane object detection

5.1 Introduction

At present, tower crane safety detection is not intelligent, and its generalization ability is weak. When encountering new environments or new rules, it does not have the ability to self-learn and solve problems. It needs to rely on human experience to solve the problem, and there is a lack of a visual and intelligent solution based on computer vision. Therefore, it is necessary to adopt a more intelligent and adaptable tower crane safety monitoring scheme to improve the efficiency and quality of site inspection.

Firstly, this chapter briefly analyzes the tower crane image datasets, research objectives, and background of this research. Based on these internal and external reasons, the yolov5 algorithm is selected as the basic neural network in this research. The algorithm network architecture and detection principle of yolov5 are in-depth researched, and a targeted network structure optimization scheme is proposed. Section 5.2 introduces the state of the art of target detection algorithms, introduces the progress of image recognition algorithms, and summarizes the research status of its application in civil engineering. Section 5.3 introduces the yolov5 series of algorithms, and introduces some general concepts related to target detection (bounding box, anchor box, intersection ratio, loss function) to promote the understanding of the network structure and detection principle of the Yolov5 series of algorithms, and then introduces the network structure of Yolov5 algorithm, GIoU loss function and ROI pooling etc. Section 5.4 introduces an improved yolov5 algorithm with the improvement method of modified loss function and edge extraction. Section 5.5 is the experimental demonstration part, which describes this research from the experimental setting, experimental environment, evaluation indicators and experimental results. Section 5.6 summarizes this study.

5.2 State of art of target detection algorithms based on deep learning

There has been a requirement for target detection algorithms to develop in line with the progression of deep learning technology. Deep learning neural networks, part of the broader family of machine learning methods, have eclipsed the traditional feature-based techniques, where properties are derived using various algorithms taken from the information present in the image itself. The position and category of objects in images and videos can be established

by object vision, which is a crucial element of computer vision. The many parameters that cannot be encompassed by handcrafted design can be determined automatically by deep learning to identify feature representation from big data. Consequently, deep learning is used as the basis for target detection algorithms that has latterly become the established form of usage. This chapter will present the development of deep learning and computer vision, together with the current standing in research and development of deep learning-based target detection algorithms, with the advancement in civil engineering of deep learning.

5.2.1 The development of computer vision and deep learning

5.2.1.1 Computer vision

Comprehension of the content of pictures is the function of computer vision, which is a major derivate of artificial intelligence (AI). Formerly, a computer only recorded the size, format and storage size of an image. The development of computer vision meant that an image was no longer just a file but could now be understood through AI, thereby establishing itself as having an important real-world role. Human key point and object detection, scene text recognition, instance and semantic segmentation, image and video classification and object tracking are the eight focal tasks of computer vision. The function of the human brain has been used as the principal guideline for the development of the latest level of computer vision skills. Figure 5.1 demonstrates the working principles of human brain vision and computer vision.

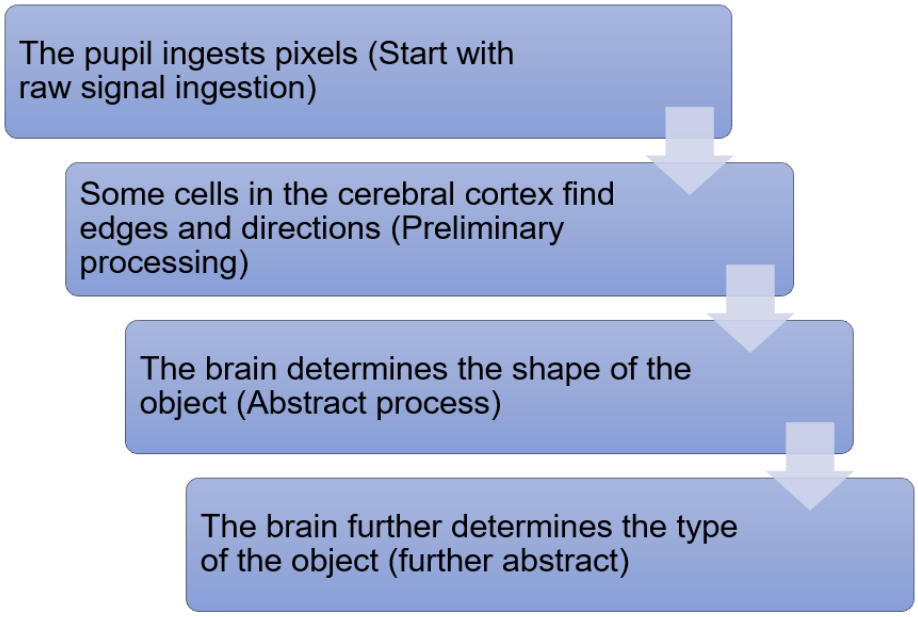


Figure 5.1: The working principle of human brain vision and computer vision

5.2.1.2 Deep learning

Known as the father of neural networks, Geoffrey Hinton invented the backpropagation (BP) algorithm in 1986 for use in multilayer perception (MLP). From this conception, AI research has constantly kept deep learning as a prime subject for consideration and research. The nonlinear classification and learning problems were solved by his use of the sigmoid nonlinear function for mapping. Thereafter, there was an increased frenzy for adopting the method of neural networks. Deep learning has replaced traditional algorithms due to its excellent performance. In many cases, industrial detection encounters complex pictures. Although noise can be removed to improve picture quality through manual effort, many traditional algorithms are prone to failure. Traditional methods of feature extraction tend to rely on manually designed extractors; this requires professional knowledge and complex parameter-tuning processes. At the same time, each method is aimed at specific applications, so they cannot be generalized easily, and they lack robustness. However, deep learning provides a solution to these issues. It does not need to pay too much attention to noise because it is based mainly on data-driven feature extraction. By learning a large number of training set samples, the final network can generate good results. It can also reduce noise.

The common network model structures of deep learning are a fully connected FC network structure; a convolutional neural network; and a recurrent neural network. The fully connected layer FC network structure is the most basic deep neural network. It is used to classify the extracted features at the early stages; each node of its fully connected layer will be connected to the previous node, so the parameters of the network are wide. It therefore requires a great deal of storage and computing space. A convolutional neural network processes data with similar network structures through a convolutional layer (the pooling layer and the FC layer). By using the convolution kernel as an intermediary, the number of parameters is reduced, thereby reducing the amount of storage and computing space that is needed. A recurrent neural network processes sequence data, mining time series information and semantic information within them.

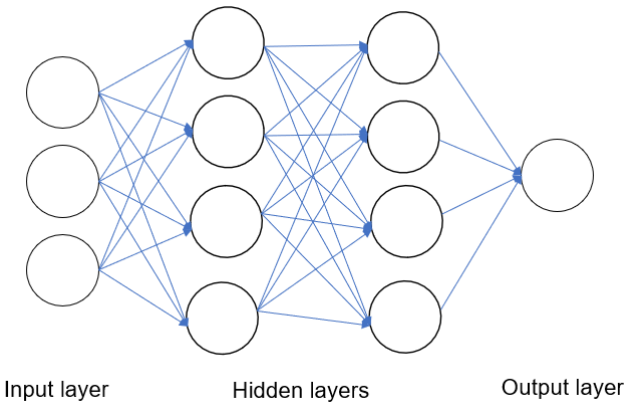


Figure 5.2: Network structure of deep learning

For instance, a tower crane image includes information such as attitude, color, texture, and light. The distributed feature representation of deep learning can use multi-layer nonlinear mapping to distinguish different and use different neurons to represent different factors in the final hidden layer. Thus, the nonlinear relationship between the factors becomes linear, thereby improving the extraction and recognition of different features. Through its analysis of the structure of the deep learning network model, the number of parameters required by deep learning is much smaller than the traditional manual labelling algorithm. The next section will introduce the research status of target detection algorithms; traditional target detection methods; and deep learning target one-stage and two-stage detection methods.

5.2.2 State of art of algorithms

Object detection is an important branch of computer vision and it has various application like intelligent video surveillance [136, 137], autonomous vehicle [138], manufacturing inspection [139] and other fields. With the improvement of computer GPU computing power, the development and iteration of deep learning convolutional neural algorithms and artificial intelligence technologies, and some mature image training sets, such as the development of COCO ImageNet [140], the development of target detection models has been promoted.

The target detection in the traditional era was constructed by manual feature, and sequentially selects regions on the detection image by sliding windows. These methods take a long time, the algorithm effect is poor, and they do not have the generalization ability. Limited by the computer power at that time, acceleration technology is generally used to improve the

detection efficiency. The following Figure 5.3 presents common traditional era object detection methods, and the table below introduces the tradition target detect algorithms and its publication year and a brief introduction.

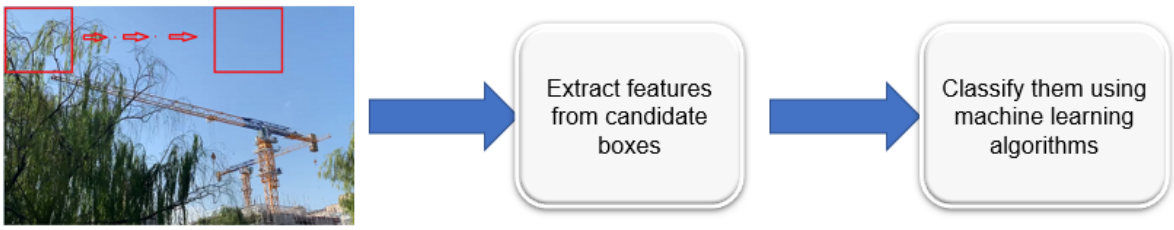


Figure 5.3: Traditional target detection method (slide window)

Table 5.1: Traditional target detection algorithms

Algorithm	Year	Brief introduction
SIFT [141]	1999	A local feature description algorithm by creating an image Gaussian pyramid, accurately position the key points to determine the location, and then construct their descriptors.
Viola Jones Detectors [142]	2002	Combining Integral Image, Feature Selection and Detection Cascades to improve detection speed
HOG Detector [143]	2005	Using overlapping local contrast normalization to improve accuracy, mainly used to detect pedestrian.
DPM [144]	2008	As an extension of the HOG detector, contains root-filter and some part-filters.

In the era of deep learning algorithms after 2014, deep convolutional neural networks and GPU computing power have been improved step by step. Object detection is gradually used in multi-class detection [145, 146], human face recognition [147], pedestrian and vehicle detection [148, 149], edge detection and other fields. The target detection algorithm mainly completes two tasks: target classification and target regression and the essential difference lies in whether the output label has a distance metric. Both classification and regression are supervised learning, making predictions on input data. The output of target classification is discrete and is the category of the object. For example: pedestrians, vehicles, etc. The output

of target regression is continuous and is the value of the object.

According to the different steps required for the task to be completed, the target detection algorithm is divided into two-stage detection (the process from coarse to fine) and one-stage detection(complete in one step) [150]. The following table compares one stage detection and two stage detection. Respectively from: essence of method, recognition accuracy, training speed, representative algorithms.

Table 5.2: Comparison of one stage detection and two stage detection

	One stage detection	Two stage detection
Essence of method	Regression	Classification
Recognition accuracy	Lower accuracy	Higher accuracy
Training speed	Quick	Slow
Representative algorithms	Yolo series, SSD, DSSD, RetinaNet, M2Det	RCNN, SPPNet, Fast RCNN, Faster RCNN, FPN

Two stage detection transform the object detection problem into a classification problem. Two-stage detection will first generate a candidate region, and then classify and locate it through the CNN classification algorithm. Among them, the development of two-stage detection is shown in Figure 5.4, mainly from, RCNN [151], which searches 2000 candidate regions from top to bottom through selective search) – SPPNet [152], it use CNN operation to extract features on the entire image, improving the speed)- Fast RCNN (add a ROI POOLing layer, using softmax for classification, using multi-task loss function)- Faster RCNN [153] (no longer using selective search, proposed the concept of RPN network and anchor box, RPN network and detection network share convolutional features, can evolve End-to-end training)- Feature Pyramid Networks [154](All region proposals share all parameters, improving algorithm efficiency), the following figure introduces the development of the main two-stage target detection.

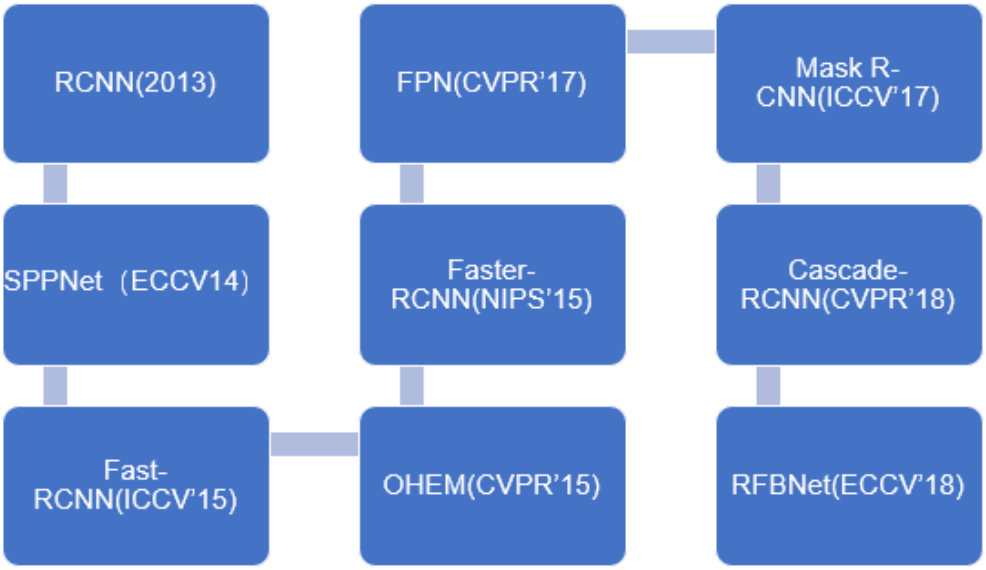


Figure 5.4: The development of two stage detection

One stage detection converts the target detection into a regression problem, and the unified network can be used to directly predict bounding boxes and classification categories. Figure 5.5 shows the development of two stage detection. Take the most widely known yolo algorithm as an example. The Yolo series of algorithms [155] was the first to propose the idea of One stage detection. Yolo (you only look once) is widely used because of its speed. The traditional Fast R-CNN algorithm uses the selective research method to select 2000 region proposals for each image, but yolo only needs to predict more than 100 region proposals for each image. One stage detection development from yolo series to single shot multi-box detector(SSD) which is a one-stage multi-category single-shot detector [156], to deconvolutional single shot detection (DSSD) which adds prediction and deconvolution modules [157], using RESNET101 as the backbone network- RetinaNet [158](accuracy reaches the level of two stage detection detectors)- M2Det [159](construct feature pyramid via MLFPN). The figure below shows the development of one stage detection and the published conference.

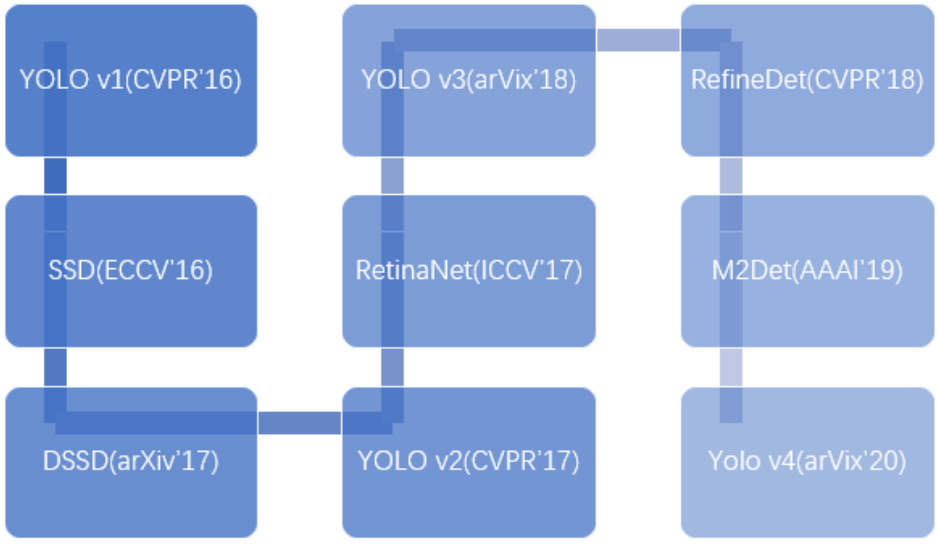


Figure 5.5: The development of one stage detection

5.2.3 Development in engineering

As can be seen from the state of art in the previous section on deep learning, neural networks, target detection, and computer vision, these deep learning-based target detection methods have made breakthroughs in many fields. However, the civil engineering industry is one of the industries with the slowest progress in digitalization and smart intelligence. Deep learning has inspired innovations in many fields, including civil engineering. In recent years, scholars have carried out in-depth exploration and expansion in the field of multidisciplinary integration of architectural engineering, computer vision, and deep learning. At present, the application of computer vision deep learning in these field of civil engineering: artificial intelligence and design, intelligent construction, intelligent management of construction sites, post-maintenance and early warning detection, intelligent materials, etc. The development of automated and intelligent construction sites has become the consensus of the construction party and the first party. Therefore, this research proposes to use deep learning and computer vision to carry out a series of target monitoring research based on tower crane video images through the powerful capabilities of image recognition, image segmentation, and pattern recognition in the field of computer vision, so as to realize online and offline monitoring of tower cranes. The operation situation has laid a solid foundation for the subsequent tower crane construction and the digital twin construction of the entire construction site.

5.3 Yolov5 series algorithm

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 to 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 great differences in recognition accuracy and detection speed and should be selected according to the research direction and goal.

High-quality, large-scale datasets can effectively improve the accuracy of training models. The initial data collected in this study are small compared to some mature datasets. Yolov5 passes each batch of training data through a data loader and simultaneously augments the training data with scaling, color space adjustment, and mosaic enhancement. In the development of the tower crane digital twin, tower crane recognition is the first step. The purpose of tower crane image recognition is to identify the tower crane in each frame of a video. This not only needs to meet the requirements of high detection accuracy, but also requires qualified detection speed. The reasoning time of a single image of the Yolov5 series can reach 7ms (140FPS), which is the current state-of-the-art in the field of object detection. One of the features of Yolov5 is that the weight file is very small, it can be mounted on mobile devices with lower configuration, it can quickly converge on multiple data sets, and the model is highly customizable. Therefore, considering the above factors comprehensively, this study is based on yolov5 to design, optimize, train and test deep neural networks.

5.3.1 Related content

In traditional image classification tasks, we generally use convolutional neural networks to extract image features, and then use these features to predict the classification probability and establish a classification loss function according to the sample labels. In the object detection problem, learn from the experience of image classification, split the target detection task, and generate a series of regions which may contain objects on the input image. These regions can be regarded as an image separately, and the image classification model is used to classify them to see which category or background it belongs to. Before studying the network structure and detection principle of the Yolov5 series of algorithms, this section first introduces some general concepts related to target detection to promote the understanding of the subsequent

content. The introduction mainly includes bounding box, anchor box, Intersection-over-union, loss function.

5.3.1.1 Bounding box

When using deep learning method to detect targets, it is necessary to predict the two core tasks of object category (target classification) and position (target detection) at the same time, so some concepts related to position need to be introduced. We usually use the bounding box to represent the position information of the object, which is a closed rectangle that can just contain the object. There are two types of bounding boxes for target detection. In the detection task, the true bounding box of the object given by the label of the training dataset is called the ground truth box, which refers to the real position of the object in the image. The other is the possible location of the target object predicted by the trained deep learning model. The prediction box is the bounding box predicted by the model. The image pixel coordinate system differs from that of the standard Cartesian coordinate system. The three-dimensional space object is represented by a projection on the image plane, discretizing the pixel, with its coordinate origin in the top left-hand corner of the charge-coupled device (CCD) image plane. The *u*-axis is parallel to the CCD plane horizontally to the right and the *v*-axis is perpendicular to the *u*-axis downward, with *u* and *v* representing the coordinates.

Usually, the common bounding boxes in the yolov5 algorithm have the following three expressions:

(1) The pascal_voc method, that is, the coordinates of the frame are encoded as [x_1, y_1, x_2, y_2], where x_1, y_1 represent the coordinates of the upper left corner of the frame, and x_2, y_2 represent the coordinates of the lower right corner of the frame.

(2) The coco method, that is, the frame coordinate encoding is [x_1, y_1, w, h], where x_1, y_1 represent the coordinates of the upper left corner of the frame, and w and h represent the width and height of the frame.

(3) The yolo method, that is, the coordinate encoding of the frame is [x_c, y_c, w, h], and these 4 data are normalized by data. Where x_c and y_c represents the centre position

of the border, and w and h are the width and height of the border.

In the detection task, we hope that the model trained by deep learning can output the bounding box predicted by the model according to the input image, as well as the type of objects contained in the bounding box and the probability that the objects in the predicted box belong to this category. Usually we have this format: [L, P, H], where L represents the category of the predicted object, and P represents the probability that the predicted box predicts the object to belong to this category. Among them, H represents the position information of the prediction frame, and one of the above three position expressions is selected for representation according to different methods. Among them, a picture may generate multiple prediction boxes, and then we will use the anchor box part to understand how the deep learning algorithm accomplishes this task.

5.3.1.2 Anchor boxes

Following the collection from the input regions of a large number of images, the target detection algorithm then establishes whether the target for the research needs is contained within them and then, to more precisely forecast the target ground-truth bounding box, adjusts the region edges. At this stage, this section utilizes a method to generate multiple imaginary bounding boxes to overcome the problem that multi-scale target detection cannot be carried out as a window can only detect one target. Different sizes and aspect ratios are generated on each pixel. This concept, first advanced in the faster Region-Based Convolution Neural Network (R-CNN) paper, terms these bounding boxes as anchor boxes. In the YOLO series algorithms, which are based on regression, the use of anchor boxes allows the shape and size of the most commonly occurring boxes to be calculated by the k-method from the ground truth boxes in the training set, thereby allowing the predetermination of statistical priors, probability calculated before some evidence is taken into account. The model can converge more quickly if experience is included, allowing multi-scale learning. Figure 5.6 shows A1, A2 and A3 as anchor boxes (marked in red), with the ground-truth box showing the ground truth (GT) marked in blue.

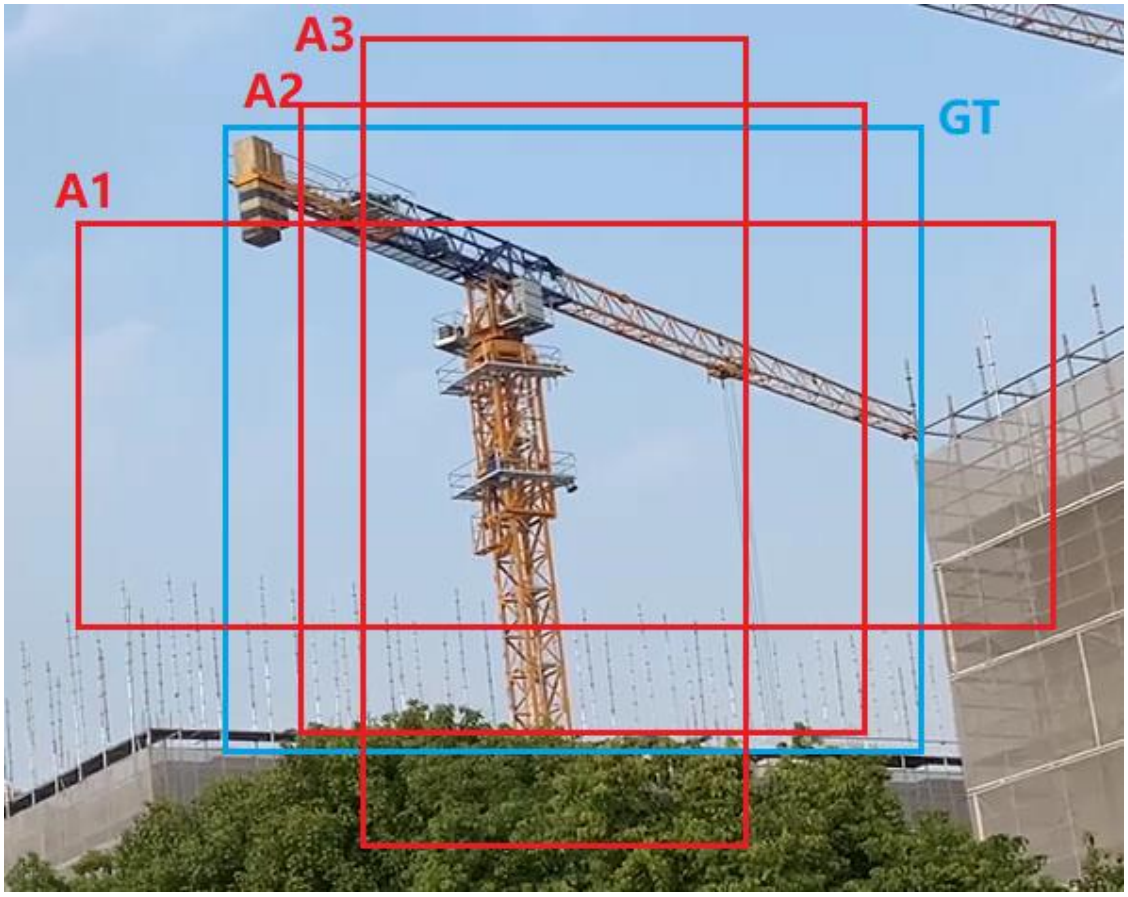


Figure 5.6: Ground truth frame and anchor boxes in tower crane image

Several different-sized anchor boxes will be created in the image when using the deep learning algorithm to detect the object (Figure 5.6). There is a need to further predict the object type if the target object appears in the anchor box. As the bounding box of the object is unlikely to conform to the first version of the anchor box, there is a need to fine-tune to create a prediction box based on the anchor box to describe the position of the object with accuracy. The model needs to progressively learn and adjust the parameters during the training process, to establish whether the anchor box representing the candidate area contains objects, forecast the object category and by what margin the boundary box requires to be adjusted relative to the anchor box. The training phase flow chart and the prediction phase are shown in Figure 5.7 and Figure 5.8.

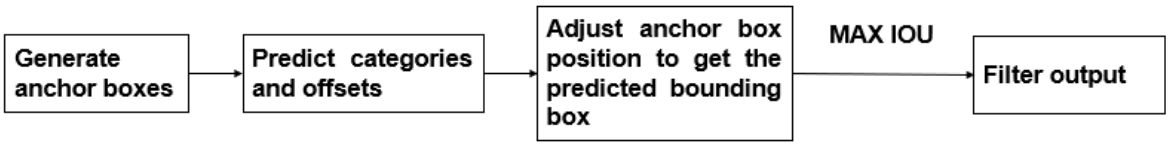


Figure 5.7: Training phase flow chart

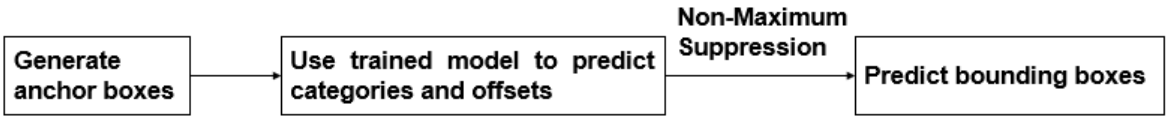


Figure 5.8: Prediction phase flow chart

5.3.1.3 Intersection over union

Through the introduction of anchor boxes in Section 5.3.1.2, the role of anchor boxes is to find areas in the image where there may be detected objects. In the anchor frame-based target detection algorithm, when the anchor frame contains an object, the trained model needs to be used to predict the category of the object and fine-tune the coordinates of the anchor frame. After multiple adjustment processes, a final prediction frame is obtained. In order to determine whether the anchor frame contains objects, the concept of Intersection-over-union is proposed. When the Intersection-over-union between the anchor frame and the real frame is small, we think that the anchor frame does not contain the detected object, when the Intersection-over-union of the box is relatively large. The object is considered to be contained in the anchor box.

Intersection over Union (IoU) is an important metric in target detection, which describes the overlap ratio of the pixel area between the real and candidate frames. The degree of correlation between the ground truth and candidate prediction boxes can be established by using the Intersection over Union concept. The IoU of the two boxes, considered as the sum of the two pixels, equates to the overlapping part divided by the two boxes' union area. Figure 5.9 displays the union of the combined area of the two boxes in green, and the intersection of the overlapping area of the two boxes in blue.

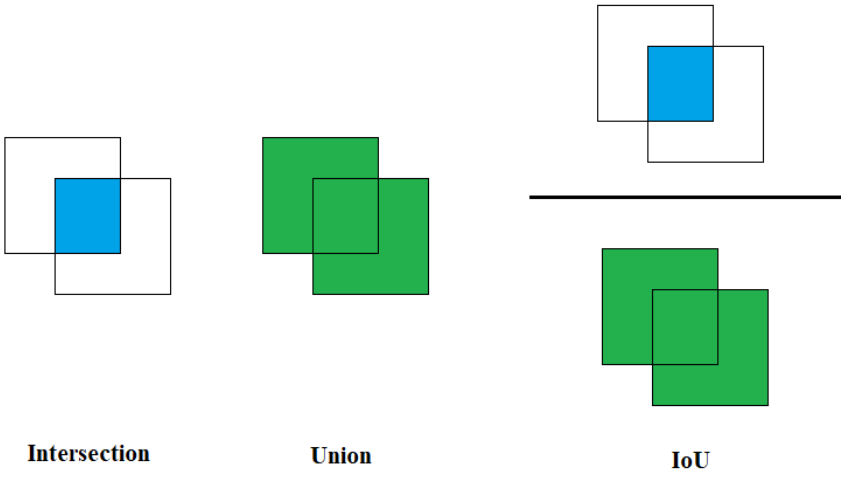


Figure 5.9: Diagram of IoU

The result of IoU is generally expressed in decimal form, and its value range is 0-1. Generally speaking, when IoU is larger than or equal to 0.5, which means that the result of deep learning is acceptable and can be detected correctly. If IoU is equal to 1, that is, the predicted box overlaps the actual bounding box perfectly, because the intersection is the union. If IoU is equal to 0, that is, the surface base where the predicted box and the actual bounding box have no intersection, and the intersection is an empty set. It means that the prediction frame is completely separated from the actual label frame, and the detection result is expected to be bad. It is generally agreed that 0.5 is the threshold, which is used to determine whether the predicted bounding box is correct. The higher the IoU, the more accurate the predicted bounding box. The figure below shows the relative positions of the predicted and ground-truth boxes for 9 different intersections from 0.00-0.95. Figure 5.10 shows the relative position relationship between the prediction frame and real frame under different IoU.

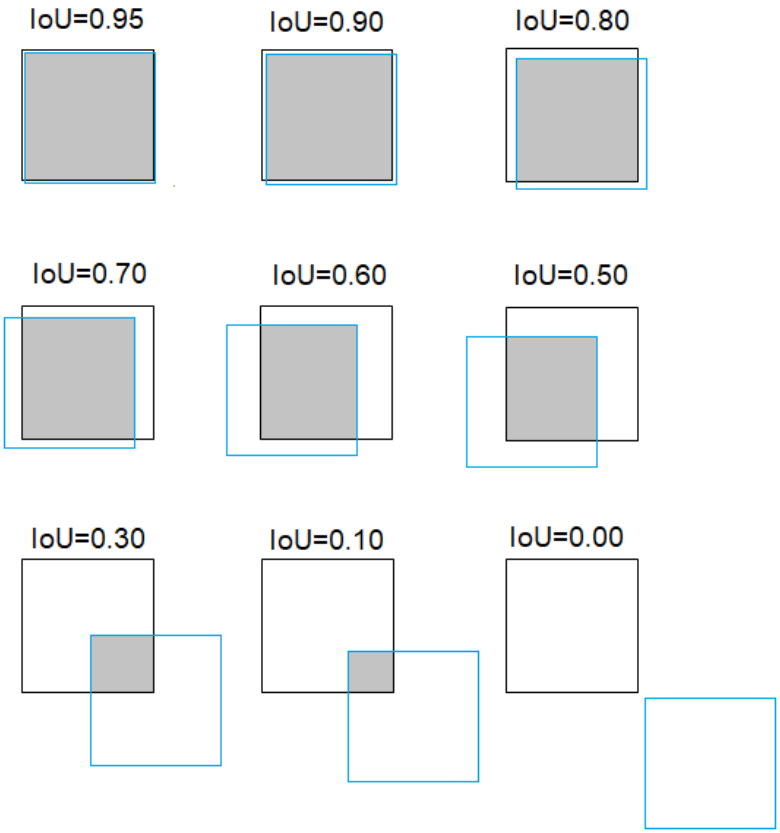


Figure 5.10: The relative position relationship between the prediction frame and real frame under different IoU

5.3.1.4 Loss function

To measure the difference between the real value Y and the model predicted value f(x), the loss function is used as an operational function. Normally, L(Y,f(x)) is used to denote the non-negative real-valued function. To minimalize (optimize) the loss function, the deep learning model training process is employed, whereby, for model parameter fitting, f(x) and y are as close as possible. To establish the training method of the minimum value of the function, normally the gradient descent algorithm is used. The smaller the difference between the real value and the model predicted value, the smaller the loss function value, which will mean real data will be better fitted with the model, ensuring greater resolution of the model. In the normal course of events, different loss functions are needed when training models for different deep learning algorithms. A number of factors determine which loss function is employed. This can depend upon the selection of machine learning algorithms, if there are outliers, the ease of finding the function derivative and the running gradient time efficiency descent, together with

the projected result confidence level. Essentially, two major applications are applicable for the loss functions most often used in target detection. These are classification loss for discrete variables and regression loss for continuous variables. Table 5.3 is the examples of regression loss and classification loss and their functions. Entropy and cross-entropy, Kullback-Leibler (K-L) divergence and Dice loss originating from the Sørensen–Dice coefficient, are included in classification losses. Regression losses include L1 loss, known as the Least Absolute Deviations, L2 or Least Square Errors loss, IoU loss, the generalized (GIoU) loss and the Smooth L1 loss.

(1) In the process of experimental research, the parameters of the model are continuously adjusted according to the loss function value.

(2) Combine the advantages of different loss functions, and reasonably combine each loss function to better measure the similarity between samples.

(3) Construct feature space based on probability distribution measure or length distance based on the most prominent main features of the data.

(4) Select a reasonable feature normalization method, so that the transformed feature vector can be as consistent as possible with the previous content.

Table 5.3: Regression loss and classification loss

	Classification	Formula
Regression Loss	L1 Loss	$L(Y\|f(x)) = \frac{1}{n}\sum_{i=1}^{N}\|Y_i - f(x_i)\|$
	L2 Loss	$L(Y\|f(x)) = \frac{1}{n}\sum_{i=1}^{N}(Y_i - f(x_i))^2$
	Smooth L1 Loss	$L(Y\|f(x)) = \begin{cases} \frac{1}{2}(Y - f(x))^2, \|Y - f(x)\| < 1 \\ \|Y - f(x)\| - \frac{1}{2}, \|Y - f(x)\| \geq 1 \end{cases}$
	GIoU Loss	$IoU = \frac{A \cap B}{A \cup B}$ $GIoU = IoU - \frac{\|C \backslash (A \cup B)\|}{\|C\|}$
Classification Loss	K-L Divergence	$L(Y\|f(x)) = \sum_{i=1}^{Y} Y_i \times \log(\frac{Y_i}{f(x_i)})$
	Cross Entropy	$L(Y\|f(x)) = -\sum_{i=1}^{N} Y_i log f(x_i)$
	Softmax Loss	$L(Y\|f(x)) = -\frac{1}{n}\sum_{i=1}^{n} log \frac{e^{f_{Y_i}}}{\sum_{j=1}^{c} e^{f_j}}$
	Focal loss	$FE = \begin{cases} -\alpha(1-\rho)^{\tau}\log(\rho), y = 1 \\ -(1-\alpha)\rho^{\tau}\log(1-\rho), y = 0 \end{cases}$

5.3.2 Framework

In the traditional Faster R-CNN, the RPN and the Fast R-CNN share the convolutional layer, but in the model training, the RPN network and the Fast R-CNN network need to be repeatedly trained, and candidate frame extraction and classification are required. In contrast, the YOLO series of algorithms do not have the process of obtaining a region proposal, and yolo only needs to process it once. For the target detection problem, Faster R-CNN divides the detection results into classification problems and regression problems to solve, and yolo unifies them into a regression problem. The network structure of yolov5 will be introduced below. There

are four network models of yolov5, which are divided into Yolov5s, Yolov5m, Yolov5l, and Yolov5x. Among them, the Yolov5s network is the network with the smallest depth and the smallest feature map width in the yolov5 series. The other three algorithms are based on Yolov5s, which continuously deepens and widens the width of the feature map.

Figure 5.11 shows the framework of Yolov5 algorithm, Yolov5s network structure consists of four parts: input, backbone, neck, and prediction.

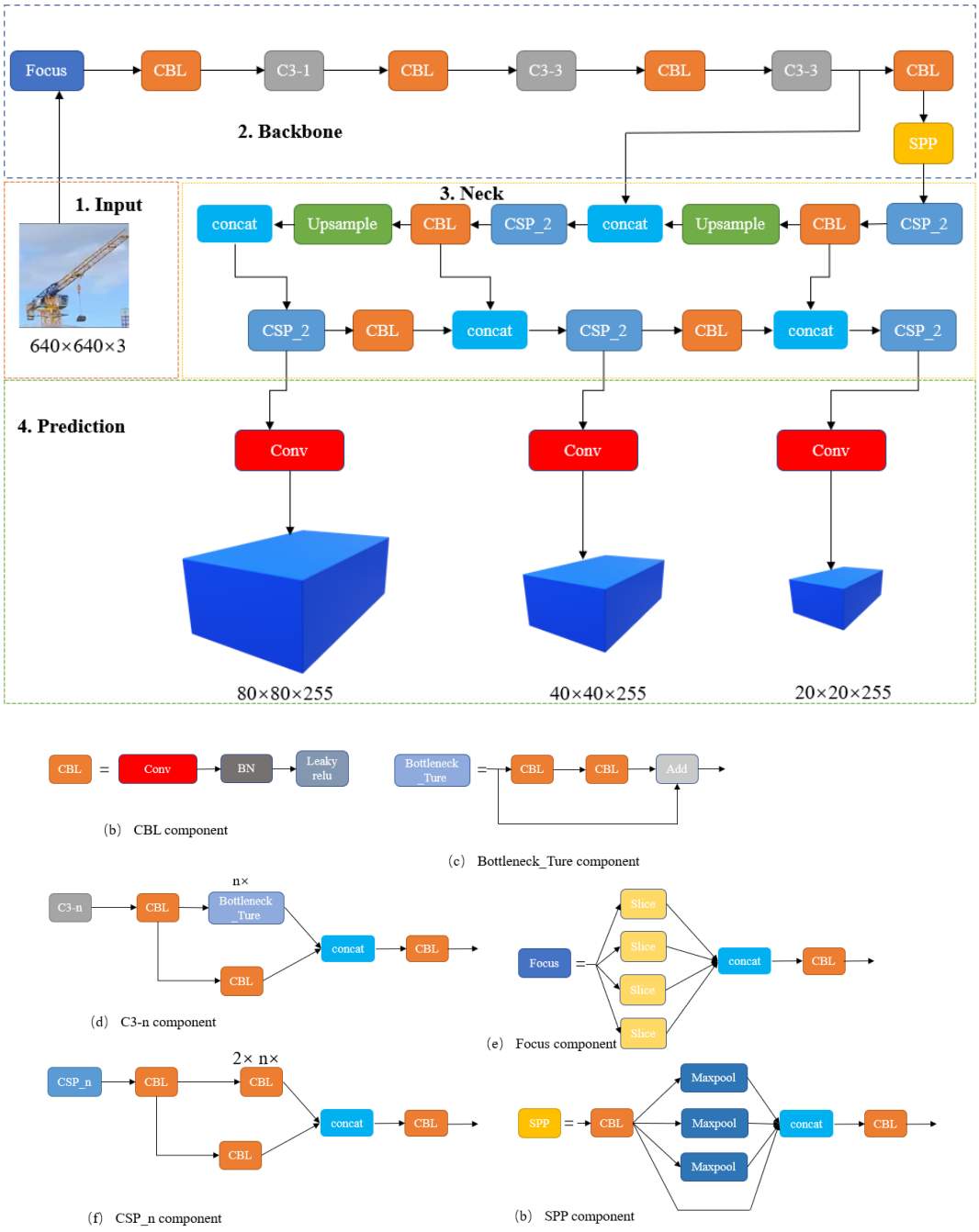


Figure 5.11: Network structure of Yolo series algorithm

In the input part, mosaic data enhancement, adaptive anchor box calculation, and adaptive image scaling are used for data enhancement.

Mosaic enhancement is mainly divided into four steps:
(1) Randomly select 4 images from the initial dataset
(2) Apply traditional data enhancement methods (scaling, cropping, rotating, etc.) to these four images, and place them in the four corners of the original image
(3) Use the matrix algorithm to stitch the four pictures into a new picture
(4) Transform the picture and output it finally.

Adaptive anchor box enhancement: The initial anchor box function is added to the code in Yolov5, which can adaptively calculate the best anchor box values for different training datasets during training process.

Adaptive image scaling: By calculating the scaling ratio, the scaled size and the black border filling value, the images of different input sizes are unified to the standard size, thereby improving the training and inference speed

Backbone structure is combined by CSP 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 obtain twice the sampling feature map without losing any information. After the slicing operation, the spliced image is changed from the previous RGB three-channel mode to 12 channels, as shown in Figure 5.12. Taking the 640×640 input as an example, the original 640×640×3 image input focus structure, after slicing operation, becomes a 320×320×12 feature. Since yolov5s has 32 convolution kernels, after the convolution operation, it finally becomes a feature map of 320×320×32.

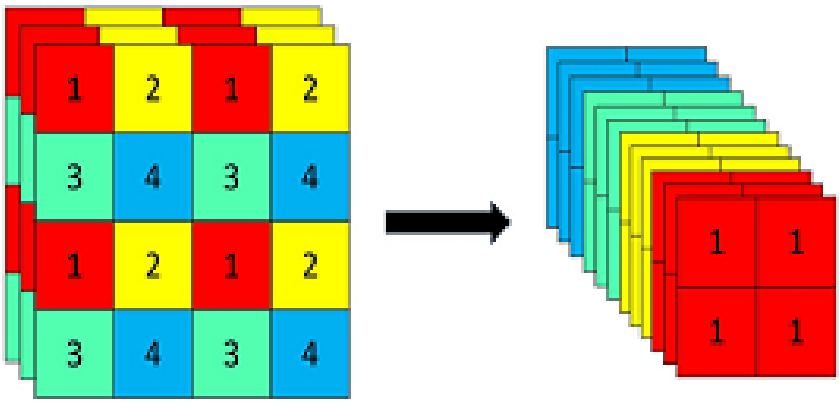


Figure 5.12: Slice operation in focus component

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 5.13 demonstrates the FPN+PAN structure. Among them, the FPN layer conveys strong semantic features of tower crane from top to bottom through up-sampling, and the PAN layer conveys strong localization features of tower crane from bottom to top through subsampling.

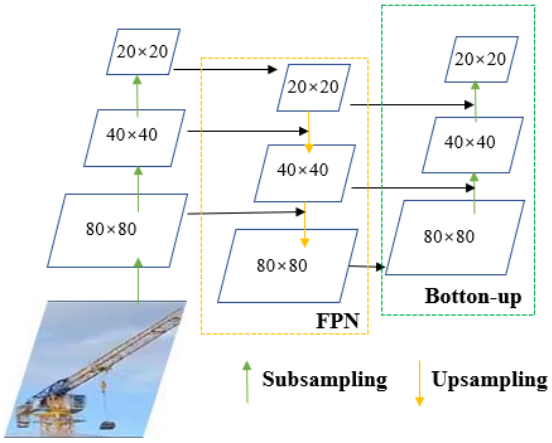


Figure 5.13: FPN+PAN component

On the output side, yolov5 uses the CIoU_Loss loss function, which will be introduced in detail in Section 5.3.3. In the postprocess of target detection, when dealing with many tower crane target frames in a picture, the method of non-maximum suppression is usually required. In the future work of this research, the loss function will be optimized.

Although the Yolov5 algorithm is good enough, there are still some components can be improved. In the input component, some latest data augmentation methods can be used instead of mosaic augmentation to get a larger dataset. On the other hand, the backbone network can be modified to improve the efficiency and accuracy of image processing. In the neck layer, the original FPN+PAN can be improved, the information loss in the feature map generation process can be reduced through the improved feature pyramid model, and the representation ability of pyramid subsampling and upsampling can be improved. In the precision layer, the output prediction result of yolov5 which use CIOU_Loss as the loss function will ignore some detection target occluded by other objects, so loss functions like distance-intersection-over-union (DIoU),non-maximum suppression (NMS), and loss function (DIoU_nms) can be used to improve it.

5.3.3 CIoU loss function

The difference between the predicted value f(x) of the model and the real value Y is measured by the loss function acting as an operation function. Normally, using L(Y,f(x)) to express, it is a non-negative real-valued function. In engineering, the most frequently used loss functions are either of the two major application situations. For discrete variables, the classification loss is used and for continuous variables, the regression loss is used.

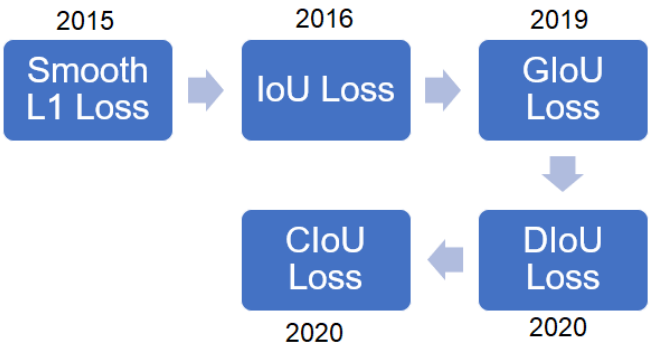


Figure 5.14: The development of loss function

The development of the loss function in object detection is evident in the above Figure 5.14. This is from IoU loss, where only the overlapping area of the target frame and the detection frame are taken into consideration. This was used until 2019, to solve the problem when the bounding boxes do not overlap, the proposed GIoU loss. The following year, a proposal was

put forward suggesting Complete-IoU (CIoU) loss and Distance-IoU (DIoU) loss, respectively. This results in better performance as the efficiency of the convergence is accelerated. The distance between the real box and the anchor box is the penalty term of the loss function in DIoU. The effect of post-processing can be enhanced by combining non-maximum suppression (nms) with the DIoU loss (DIoU-NMS). To embellish the accuracy and speed of the prediction box regression in the YOLOv5 algorithm, the CIoU loss regression method is used. The gradient calculation based on DIoU loss does not include the CIoU loss formula that adds a balance parameter, being the *v* part, which is the similarity calculation of the tow box aspect ratio metrics. The diagonal distance of the smallest circumscribed rectangle C is referred to as Distance C. The distance between the centre point of the target frame and that of the prediction frame is termed as Distance 2.

$$CIoU_{LOSS} = 1 - CIoU = 1 - (IoU - \frac{{Distance_2}^2}{{Distance_C}^2} - \frac{v^2}{(1-IoU)+v}) \quad \text{eq (5-1)}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w^p}{h^p})^2 \quad \text{eq (5-2)}$$

This

5.3.4 Non-maximum suppression

The non-maximum suppression algorithm is used to suppress the remaining values in the field when determining the local maximum. This is critical in applications in the field of computer vision that includes object recognition, video object tracking texture analysis and edge detection, as well as important use in algorithms such as dimension-based partitioning and merging (DPM) clustering, You Only Look Once (YOLO), single-shot detector (SSD) and Fast Region-based Convolutional Network Method (FAST R-CNN).

Target detection can project multiple numbers of candidate prediction frames in close proximity to the target object following the deep learning algorithm learning, where the prediction frames overlap. This is unlike image classification, which is likely to have only a single output. The model also presents a regional proposal higher than the actual number, thereby improving the recall value. Consequently, to find the best prediction bounding box and eradicate redundant bounding boxes, there is a need to use non-maximum suppression.

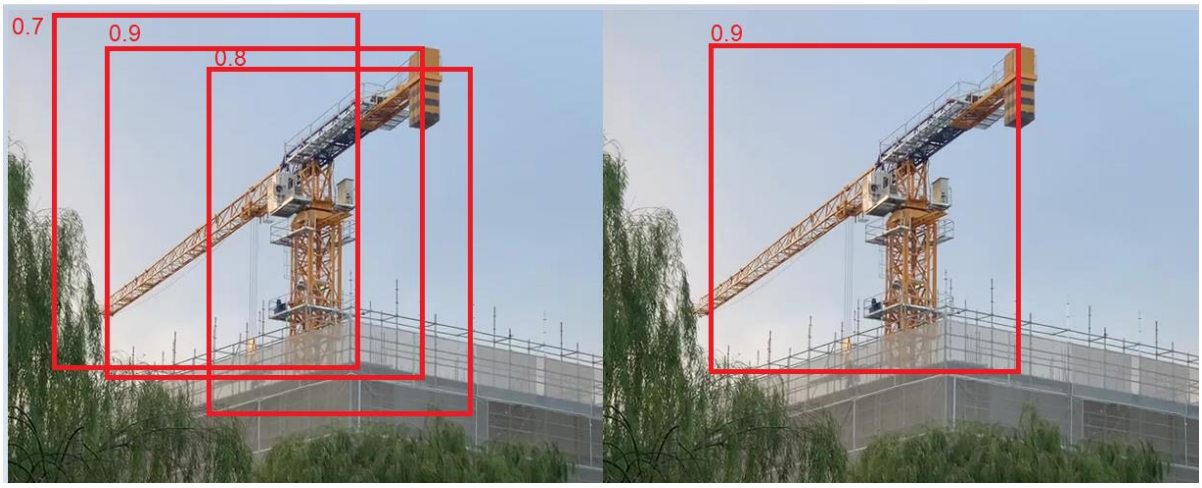


Figure 5.15: Results of candidate prediction frames

The picture on the left of Figure 5.15 is the result of candidate prediction boxes for tower crane image recognition, and each prediction box has a confidence score. Therefore, we need to set a threshold for the predicted bounding box and its corresponding topping degree to delete some bounding boxes with large overlap and repeat the above operation for the remaining detection boxes to guide the processing of all candidate boxes in the image. The picture on the right is the result of using non-maximum suppression, which also matches the expected result of our target detection. The following Figure 5.16 describes the process of non-maximum suppression:

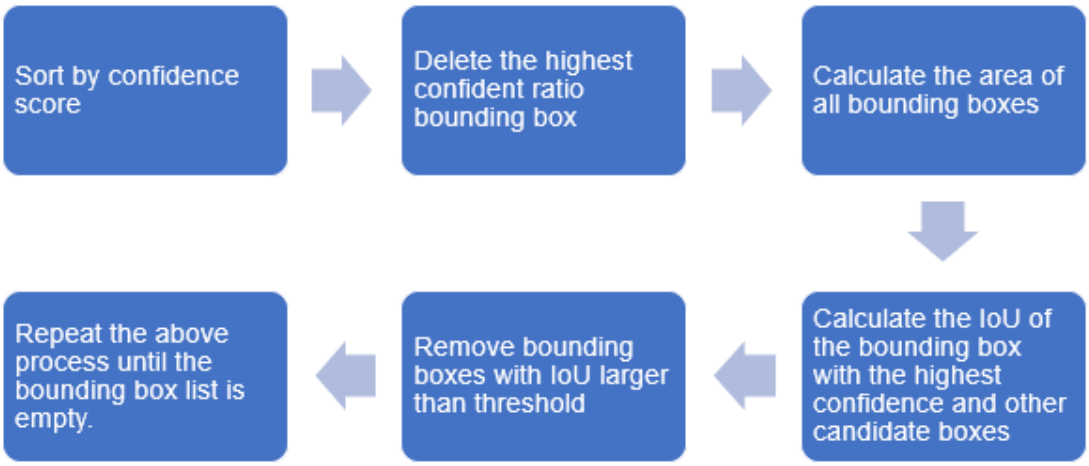


Figure 5.16: Flowchart of non-maximum suppression

5.4 Improved Yolov5

5.4.1 Distance-intersection-over-union (DIoU)_non-maximum suppression (NMS) loss function

Candidate anchor boxes are likely to appear close to the real box in the target detection prediction stage. The same target may be surrounded or overlapped by them. In assessing the post-processing step of target detection, in this step, the non-maximum suppression (NMS) algorithm is normally used to eradicate the possibility of redundant detection frames. The best bounding box is retained according to preset conditions by using NMS to merge like bounding boxes close to the recognized object.

Only the IoU factor needs to be considered in the standard 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 images, 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 is used to decide whether or not to delete a frame by measuring the distance ratio of the two prediction frames.

$$s_i = \begin{cases} s_i, IoU - R_{DIoU}(M, B_i) < \varepsilon, \\ 0, IoU - R_{DIoU}(M, B_i) \geq \varepsilon, \end{cases} \quad \text{Eq (5-3)}$$

Here, the classification confidence is s, the NMS threshold is ε and the box with the highest confidence is M. $B_i$is the other box frames

$$R_{DIoU} = \frac{\rho^2(b, b^{gt})}{c^2} \quad \text{Eq (5-4)}$$

b, b^{gt} represent the centre points of the predicted box and the real box, respectively, and ρ represents the Euclidean distance between the two centre points. c represents the diagonal distance of the smallest closure region that can contain both the predicted and ground-truth boxes.

5.4.2 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 grayscale 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. Normally, the first or second derivation can detect edges. The maximum value of the corresponding edge position can be detected by the first derivative taking the maximum value. The zero-crossing point as the position of the corresponding edge can be taken by the second derivative.

Edge detection can be conducted using the Sobel operator, often called the Sobel filter, a discrete differentiation operator that creates images emphasizing edges. Used to calculate the approximate values of the brightness and darkness of an image, this operator combines Gaussian smoothing, a 2-D convolution operator, and differential derivation in its process. The edge is recorded where a particular point in the area exceeds a certain number, according to the brightness and darkness next to the image edge. The image considers that adjacent point distances have different influences on the current pixel. Image sharpening and highlighting of the edge contour can be achieved by determining that the shorter distance has a greater influence on the current pixel.

The phenomenon where the grayscale weighted difference between the left and right and upper and lower adjacent points of the pixel, also achieves its extreme value at the edge, can be detected by the Sobel operator. It provides greater accuracy on edge direction information and has a noise smoothing effect. Noise can be eliminated as the operator utilizes differential derivation (differentiation) with Gaussian smoothing. The Sobel operator tends to be used as an edge detection method when not very high accuracy is the requirement.

The following formula demonstrates the algorithm template, where the horizontal direction is represented by *dx* and the vertical direction by *dy*.

$$d_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad d_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

This operator includes two sets of 3x3 matrices (horizontal and vertical direction) and perform planar convolution with the image to obtain the horizontal and vertical brightness difference approximations respectively, A represents the origin image, the formulas are shown below, $f(x, y)$ represents the of the grayscale value of (x, y) point of the image.

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \qquad \text{Eq (5-5)}$$

$$\begin{aligned} G_x = {} & (-1) \times f(x-1, y-1) + 0 \times f(x, y-1) + 1 \times f(x+1, y-1) \\ & + (-2) \times f(x-1, y) + 0 \times f(x, y) + 2 \times f(x+1, y) \\ & + (-1) \times f(x-1, y+1) + 0 \times f(x, y+1) + 1 \times f(x+1, y+1) \\ & = [f(x+1, y-1) + 2 \times f(x+1, y) + f(x+1, y+1)] \\ & - [f(x-1, y-1) + 2 \times f(x-1, y) + f(x-1, y+1)] \end{aligned}$$

$$\begin{aligned} G_y = {} & 1 \times f(x-1, y-1) + 2 \times f(x, y-1) + 1 \times f(x+1, y-1) + 0 \times f(x-1, y) \\ & + 0 \times f(x, y) + 0 \times f(x+1, y) + (-1) \times f(x-1, y+1) \\ & + (-2) \times f(x, y+1) + (-1) \times f(x+1, y+1) \\ & = [f(x-1, y-1) + 2 \times f(x, y-1) + f(x+1, y-1)] \\ & - [f(x-1, y+1) + 2 \times f(x, y+1) + f(x+1, y+1)] \end{aligned}$$

The arithmetic square root of the horizontal and vertical grayscale values of each pixel of the image is used to calculate the number of the grayscale level of the point. However, usually, in order to improve the efficiency of the training process, an approximation without the square root is used, if the gradient G is larger than a certain threshold, the point (x,y) is considered as an edge point.

$$G = \sqrt{(G_X)^2 + (G_y)^2} \qquad \text{Eq (5-6)}$$

Figure 5.17 below shows the python code of sobel operator of edge extraction to processing the original images and Figure 5.18 shows the tower crane image processed by sobel operator.

```python
from skimage import filters,io
import numpy as np
import matplotlib.pyplot as plt
gx = np.array([[-1,0,1],[-2,0,2],[-1,0,1]])          #gx 3x3 matrix
gy = np.array([[1,2,1],[0,0,0],[-1,-2,-1]])          #gy 3x3 matrix
def sobel(img):
    height = img.shape[0]
    width = img.shape[1]
    tmp_img = img.copy()
    for i in range(1,height-1):
        for j in range(1, width-1):
            tmpx = np.sum(np.sum(gx * img[i-1:i+2,j-1:j+2])) #calculate gx grayvalue
            tmpy = np.sum(np.sum(gy * img[i-1:i+2,j-1:j+2])) #calculate gy grayvalue
            tmp_img[i,j] = np.sqrt(tmpx**2 + tmpy **2) #calculate arithmetic square root
    return tmp_img
```

Figure 5.17: Python code of sobel opeator

Figure 5.18: Tower crane images using sobel operator

Figure 5.19 below shows the framework of the improved yolov5, compared with the original yolov5 framework, enhanced yolov5 added edge extraction in the input part and used DIoU_nms to predict the detected object in the prediction part.

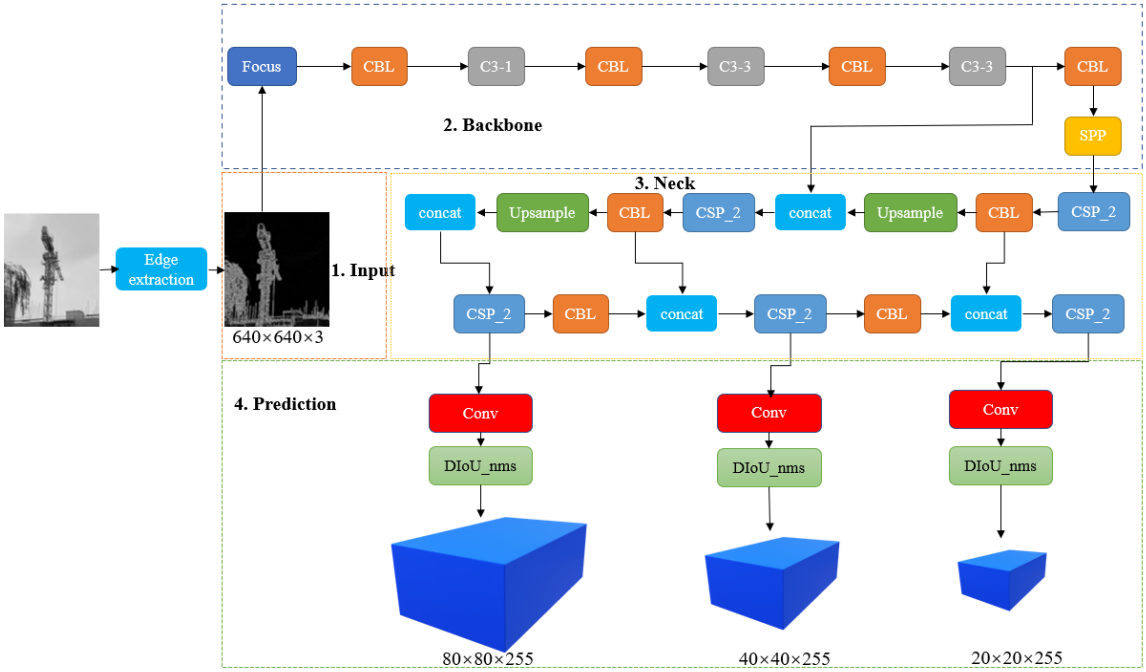


Figure 5.19: Framework of the improve yolov5

5.5 Experimental verification results

This section will use the tower crane image dataset established in chapter 4 and use the algorithm optimization strategy introduced in Section 5.4 to experiment and analyze the improved Yolov5 to verify the target detection ability of the deep learning network model.

5.5.1 Experimental setup

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, 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.

5.5.2 Experimental environment

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.

5.5.3 Evaluation indicators: F1 function, precision, recall, map

Numerous evaluation indicators have been devised by scholars for quantitative analysis of target detection algorithm performance. They all indicate the performance of the algorithm to its level to a certain extent. Frame per second and floating-point operation volume FLOPS are detection speed indicators. General precision indicators are precision, recall, accuracy and mean average precision (mAP). The target detection is 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.

The model's ability through recall to detect positive samples among the actual positive samples, and by precision discriminate against negative samples are in the predicted positive samples set. Recall is the number of labelled images detected from the labelling perspective, and precision is the number of accurate predictions made from the prediction result. However, the

confusion matrix needs to be loaded initially before the precision and recall images are detected.

Table 5.4: Definition of confusion matrix

Confusion matrix	Prediction result is positive	Prediction result is negative
Prediction is true	TP	TN
Prediction is false	FP	FN

Table 5.5: Definition of TP, TN, FP, FN concept

TP	Positive predicts to be positive
TN	Positive predicts to be negative
FP	Negative predicts to be positive
FN	Negative predicts to be negative

(usually the category we concern is positive and others are negative)

As an example, the tower crane detection of the tower crane image recognition can be taken. 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*. Figure 5.20 is the example of the concepts of TP, TN, FP, FN.

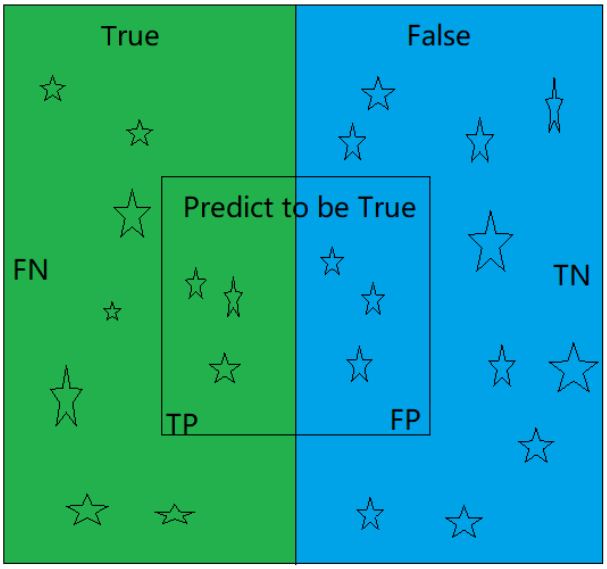


Figure 5.20: Example of the concepts of TP, TN, FP, FN

The formula below are the definition of precision, recall and accuracy using TP, FP, TN, FN.

$$Precision = \frac{TP}{TP+FP} \quad \text{Eq (5-7)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{Eq (5-8)}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{Eq (5-9)}$$

The weighted average of recall and precision is shown as the *F1* score, considering both *FN* and *FP*. It is normal for the *F1* score to be used when the distribution of data set categories is not balanced. However, the cost caused by *FP FN* is almost identical and when there is a balance in the distribution of data set categories, then accuracy is needed.

$$F_1 = \left(\frac{2}{recall^{-1}+precision^{-1}}\right) = 2 \cdot \frac{precision \cdot recall}{precision+recall} \quad \text{Eq (5-10)}$$

The precision-recall (PR) curve and average precision (AP), the area under the PR curve, need to be introduced, before introducing map, and the PR curve is a curve with recall as the abscissa and precision as the ordinate.

The expectation is that the greater the recall and precision of the model, as shown in Figure 5.21, where the curve is closer to the upper right-hand corner of the figure, the greater will be the effect of the model.

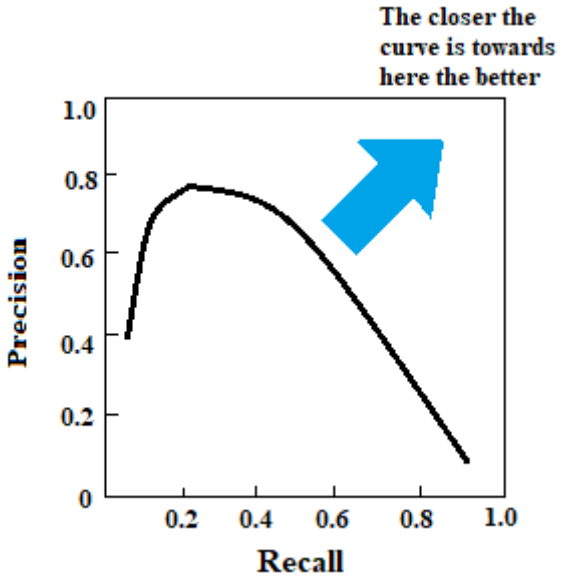


Figure 5.21: Comparison in PR curve

In the picture, the area below the PR curve is termed the average precision (AP). A confidence level for all detected objects will be presented by target recognition detection algorithms such as YOLOv5. The sample is considered positive when the preset confidence level is exceeded by the confidence level, otherwise, it is considered negative. Consequently, using different thresholds, different recall and sets of precision values will be returned. The average accuracy of recognition of the tower crane target will be deemed to be the average accuracy in the area contained between the *X* and *Y*-axes in the precision-recall curve. In the Figure 5.22 below, by utilizing the calculation formula of the *AP* rectangle rule and the schematic diagram, the calculation method of *AP* can be determined.

$$AP = \sum_i^n (R_i - R_{i-1}) P_i \qquad \text{Eq (5-11)}$$

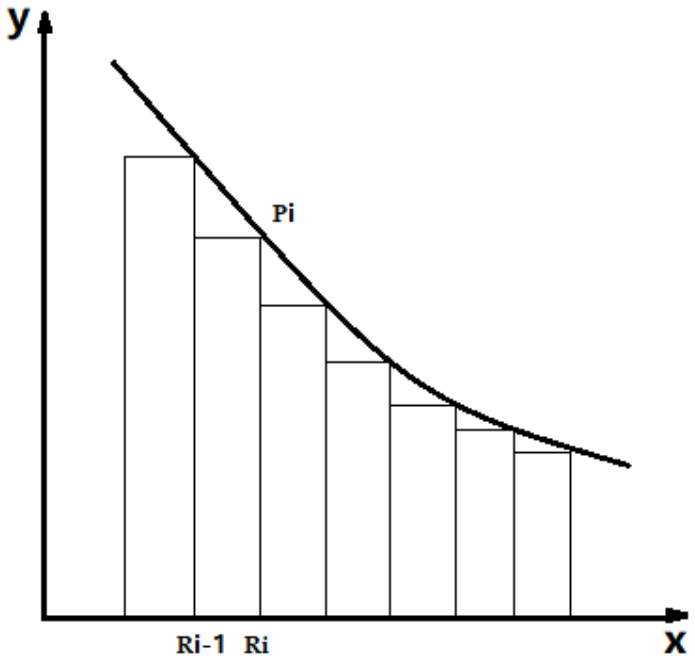


Figure 5.22: Schematic diagram of AP rectangle rule calculation (P is the precision value corresponding to the point, and R is the recall value corresponding to the changed point)

Mean average precision (mAP) is the average value of each AP category. This method is used to measure the tower crane determination ability of the target detection model as the evaluation index.

It is proposed that mAP@0.5 (IoU=0.5) and mAP@0.5:0.95 are used as the detection

indicators of YOLOv5. In the IoU concept, mAP@0.5 (IoU=0.5), TP is IoU > 0.5 detection boxes, *FP* is the number of detection boxes with IoU <= 0.5, or the number of redundant detection boxes that detect the same ground truth.

5.5.4 Low accuracy and high accuracy labelling

According to Section 3.8.2, image annotation requires that the tower crane is surrounded by a rectangular frame, which covers all parts of the tower crane according to the situation. This study compared low accuracy labelling (the rectangular frame is obviously larger than the tower crane) and high-accuracy labelling (the rectangular frame fits the tower crane). As can be seen from the Figure 5.23, if the low accuracy labelling method is used, the image recognition accuracy of the tower crane is relatively high, and the recognition accuracy has reached nearly 98% in 100 epochs. In the video verification, the frames of the tower crane are also like the annotation: the size of the prediction box is larger than the tower crane itself. However, if high-accuracy labelling is used, at the 300th epoch, the final precision only reaches 85%, but during the video verification, the size of the prediction box just fits the tower crane. Considering the follow-up, we need to identify the motion state of the tower crane, and we need to separate the tower crane from the pictures and videos. In order to ensure the accuracy of the tower crane motion pattern recognition, high-accuracy labelling will be used.

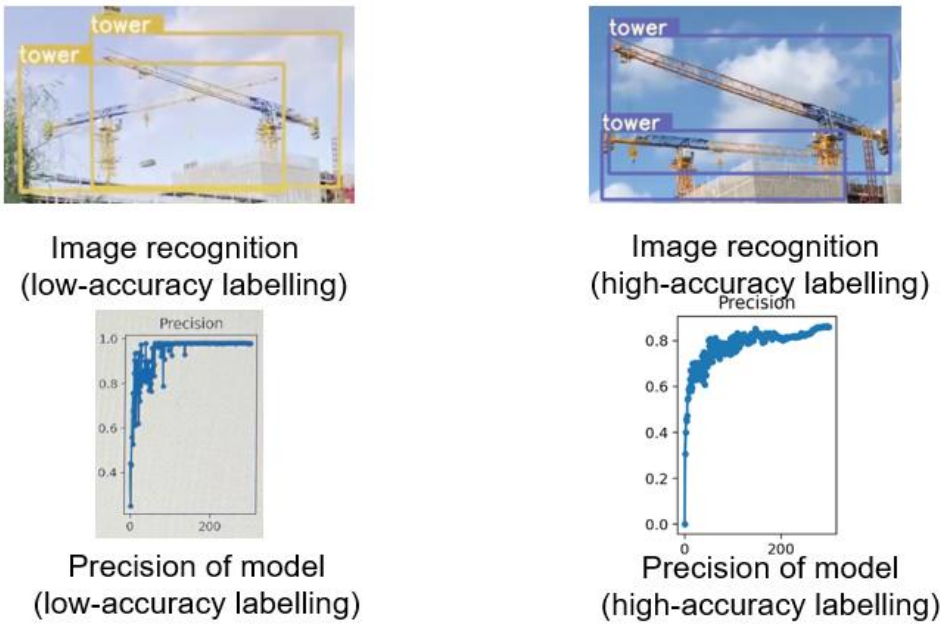


Figure 5.23: Low-accuracy labelling and high-accuracy labelling

5.5.6 Comparison of other algorithms

In this study, four Yolov5 series algorithms with different depths: Yolov5s, Yolov5m, Yolov5l, and Yolov5x are trained and tested. 3,265 new tower crane images will be used as samples to detect the advantages and disadvantages of Yolov5 algorithms with different depths. The CPU used in this study is an Intel Core (TM) i9-11900F@2.50GHz. The memory is 64GB; under the condition of an NVIDIA GeForce GTX 3090 24GB graphics card. The prediction precision, recall and F-score values of these four depth algorithms are compared.

Table 5.6: Comparison of Resnet algorithm with different depth

Method	Precision	Recall	F-score
Yolov5s	89.64%	0.9897	0.941
Yolov5m	91.68%	0.9927	0.953
Yolov5l	93.01%	0.9897	0.959
Yolov5x	93.85%	0.9912	0.964

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 F1 score above 94% (94.1%, 95.3%, 95.9%, 96.4% respectively).

Yolov5x is the deepest in the Yolov5 series algorithms. Take the input picture of 608*608 as an example. In the convolution operation of the focus structure, the number of convolution kernels is 32. After the focus structure, the size of the feature map becomes 304*304*32. The convolution operation in the Focus structure of yolov5m uses 80 convolution kernels, so the feature map after the Focus structure becomes 304*304*80. 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 5.24 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 5.25 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.

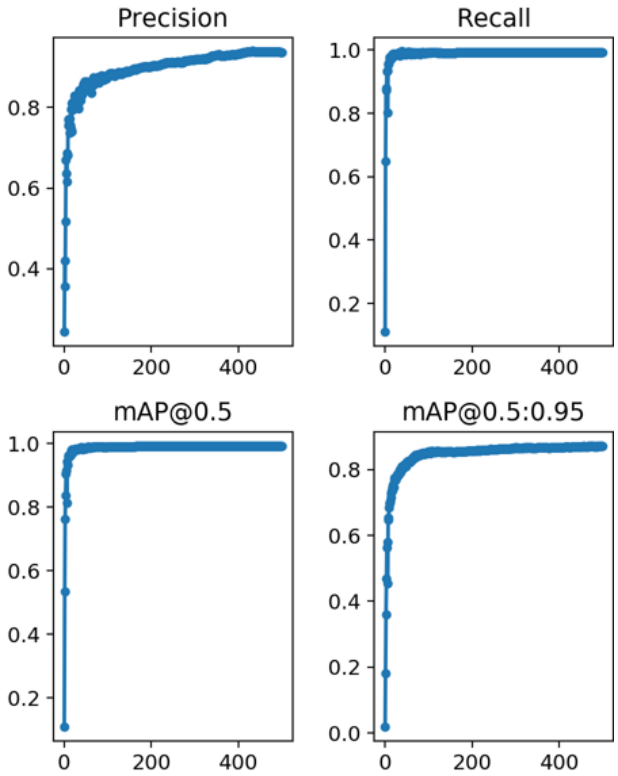


Figure 5.24: Common evaluation index curve

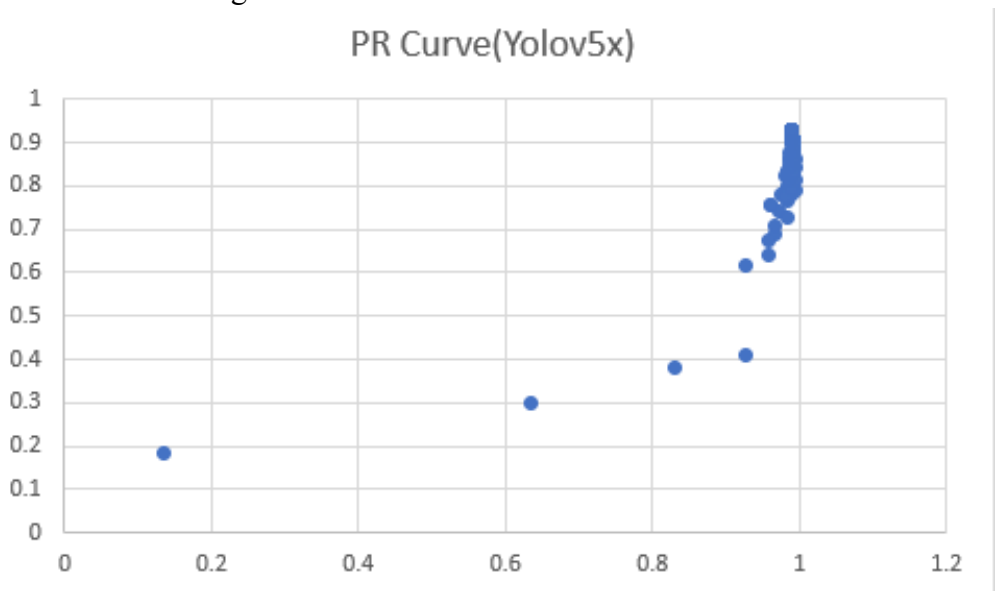


Figure 5.25: PR curve（yolov5x）

Considering the requirements of this research, this 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 will use the built-in Yolov5s algorithm for work deployment. In offline detection, we do not have many requirements for detection speed. Here, we can use the Yolov5x target detection model to detect tower cranes with better accuracy.

5.5.7 Improved Yolov5 algorithm using DIoU_nms

Although the Yolov5 algorithm is good enough, there are still some components can be improved. Non-maximum suppression(nms) operation is usually used to filter out the target frame in the post-processing process of target detection, and the weighted nms is used in yolov5, which has poor detection ability for occluded targets. Under these circumstances, DIoU_nms will be used to detect the occluded objects in this experiment. From Table 5.7, we can see that using this improving method can obviously improve the precision of Yolov5 series algorithms. Table 5.8 shows the precision, recall and F-score of the previous traditional Yolov5x and the improved Yolov5x with DIoU_nms, these three evaluation indicators are increasing, and the improved F-score demonstrated the model is better than the previous model.

Table 5.7: Precision of Yolov5 series algorithms using DIoU_nms

Method	Precision	Precision using DIoU_nms
Yolov5s	89.64%	89.98%
Yolov5m	91.68%	91.97%
Yolov5l	93.01%	93.34%
Yolov5x	93.85%	94.23%

Table 5.8: Comparison of Yolov5x algorithm using DIoU_nms

Method	Precision	Recall	F-score
Yolov5x	93.85%	0.9912	0.964
Yolov5x with DIoU_nms	94.23%	0.9925	0.967

The tower crane detected in a conventional YOLOv5 model is displayed in the left picture (Figure 5.26). Following the replacement of the original NMS algorithm by using the DIoU_nms, the tower crane can be detected by the YOLOv5 model (Figure 5.26). It can be seen that by using the conventional algorithm, the two cranes are close and will not be detected.

However, the improved DIoU_nms Yolov5 logarithm allows a better resolution to the problem.

Figure 5.26: Tower crane detection after using DIoU_nms

5.5.8 Ablation experiment

In order to better analyze and verify the effectiveness of the Yolov5 improvement strategy used in this section, 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:

Table 5.9: Detection results of Yolov5 under different improvement strategies

	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

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.

5.6 Summary

In the tower crane operation detection part of the digital twin technology, the creation of the tower crane image dataset is the basic work, and the quality of the tower crane object detection directly affects the accuracy of the tower crane operational mode recognition. Combined with the characteristics of the tower crane images, this chapter expounds the tower crane detection method and principle based on Yolov5 and proposes a series of optimization strategies for the network structure of the deep learning model according to the problems and limitations of the loss function and the comparison algorithm on the yolov5 algorithm. And the algorithm has been optimized and improved.

This chapter conducts a comparative study on four yolov5 algorithms with different depths. Through the comparison of the detection speed of a single image and the detection accuracy and F1 value, combined with the actual situation of the construction site, the yolov5x algorithm is selected. The traditional yolov5 object detection method often ignores the overlapping tower cranes, and the tower cranes are only partially occluded. In this study, the updated DIoU_nms method is used as the loss function part of Yolov5, and then the sobel operator in edge detection is used to further improve the detection accuracy.

In conclusion, this chapter verifies the detection performance of the improved yolov5 by means of experiments, and the experimental results demonstrate the effectiveness of the

improved strategy. At the same time, this chapter also conducts ablation experiments of the neural network model to deeply study and analyses the impact of different improvement strategies on the performance of the original model.

6. Tower crane operation mode recognition

6.1 Introduction

The first section discusses the development of operation mode recognition. Up until now, no suitable algorithm for tower crane operation pattern recognition has existed. We therefore needed to choose a neural network with high accuracy. Section 6.2 introduces some candidate algorithms (i.e., long-short term memory [LSTM], convolutional neural network [CNN], and residual neural network [RNN]) and their operating principles. A combination of LSTM and CNN algorithms was proposed. The residual network framework 2DResNet18 is discussed and 3DResNet is used to support the network structure of ResNet series algorithms. Section 6.3 introduces two methods to improve model accuracy based on dataset augmentation and edge extraction. Section 6.4 presents the experimental setup and environment, evaluation indicators, and the results. Section 6.5 summarizes the study.

6.2 Candidate algorithms used in this research

6.2.1 Long-short term memory (LSTM) and Convolutional neural network (CNN)

Generally, it is supposed that there is no link between time (t) and time (t+1) when considering the picture content of archetypal neural networks. To determine some similar continuous photo data, recursion neural networks (RNNs) were designed using time dimension information analysis. An RNN algorithm is long short-term memory (LSTM), used initially in general recurrent neural networks to resolve model gradient dispersion and long discrepancies. Able to retain its usefulness over the long term, LSTM is proficient in presenting time period information.

There is a greater information transmission band of cell state in LSTM, relative to the accepted RNN algorithm, as the 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.

In deep learning, a commonly used algorithm is convolutional neural network learning (CNN). AlexNet, GoogleNet, LeNet-5, ResNet and VGG-16 are commonly used CNN models. CNN is composed of convolutional and neural network. Convolutional means that each small pixel area on the picture is processed; the picture information is continuously; and the graph can be recognized by the neural network. Neural network is comprised of a series of neural layers and there are many neurons in each neural layer. To recognize things, the neurons of the neural network are essential. Convolutional neural network will order the collected information of a small area of pixels and scan the edge information generated by using a similar batch filter.

The edge information allows the structure of the higher-level information structure by summarizing the neural network, resulting from adopting the specific training steps. Thereafter, small batches of pixel blocks are gathered by the ongoing function of the filter. A 'picture' with greater height, smaller width and length (containing edge information) is created by the output value. Multiple convolutions such as compressing the width and length and increasing the height can be performed by taking the same steps.

By combining the CNN and LSTM networks, a spatial-temporal network method was developed. Any image greater than 2k needs to be downsampled as the LSTM input must have fixed image pixels. The following has to be taken into consideration in this research:

(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 recognize the operation mode of tower crane.
(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.

6.2.2 Residual network (ResNet)

6.2.2.1 2DResNet

The Residual Network (ResNet) method has been put forward by scholars to overcome the problem that when the network depth is deepened, the gradient disappears. This network is also the first network with a 100-layer depth. Termed the observed value, H(x) is the mapping to be solved, where x is the estimated value otherwise known as the identify function. The purpose is to create the residual mapping function F(x) to solve the network by converting ResNet. Thus, F(x)=H(x)-x, whereby the problem becomes H(x)=F(x)+x, Figure 6.1 below shows the framework of residual network.

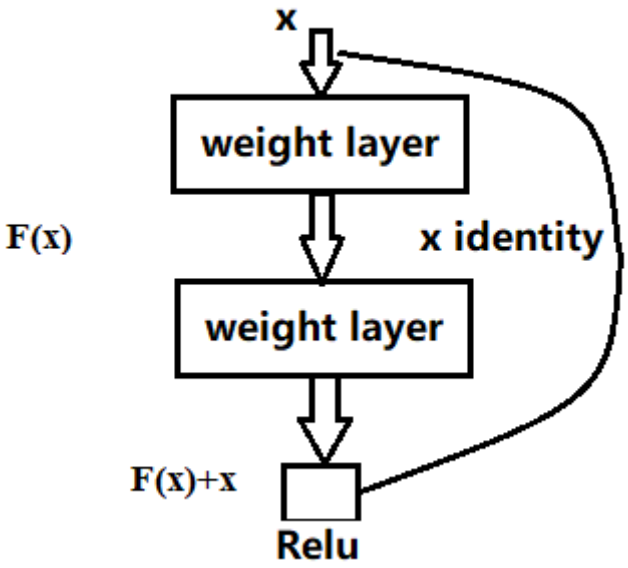


Figure 6.1: Framework of residual

x_m and x_{m+1} are the input and output of the residual cell， $g(x_m) = x_m$ is identical mapping， r is relu activation function， F is residual function。

$$y_m = g(x_m) + F(x_l, W_l) \quad \text{Eq (6-1)}$$

$$x_{m+1} = r(y_m) \quad \text{Eq (6-2)}$$

$$x_n = x_m + \sum_{i=1}^{n-1} F(x_i, w_i) \quad \text{Eq (6-3)}$$

The feature formula gleaned from the shallow layer (m) to the deep layer (n) is shown in the above formula, when updating the weight value of the F(x) part can produce new features.

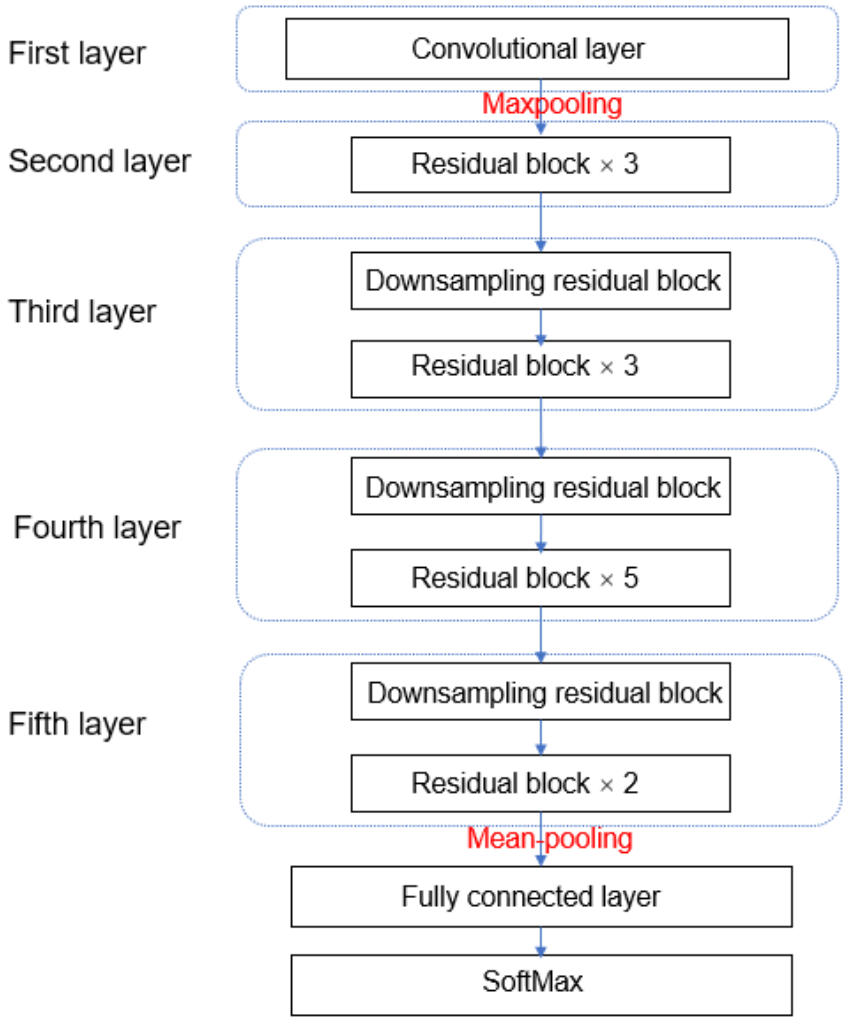


Figure 6.2: Network structure of 2dresnet34

The 2Dresent structure of the network can be seen in the above Figure 6.2. One normal convolutional and a max-pooling layer form the first construction layer. Thereafter, six residual modules make up the second construction layer. Subsequent construction layers of seven, eleven and five residual modules comprise the following third, fourth and fifth construction layers, each starting with a down-sampling residual module (each residual block contains two weight layers; for example, in the fourth layer, there are one downsampling residual block and 5 residual blocks. In total, there are eleven residual modules in the fourth layer).

6.2.2.2 Improvement between 3DResNet and 2DResNet

Contemporary video datasets such as ImageNet are in their millions. Earlier video resources

were restricted and there is inclined to be overfitting of these model networks. The training results using 2DResNet may be overfitting as there are a limited number and type 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 (3DresNet). Each 3DConv module can be connected by using the residual structure of 3DresNet = ResNet + 3DresConv and 3DresNet. Based on 2D, the convolution kernel of 3DresNet increases one dimension, adding parameters T (in channels, out channels, T). These effectively extract temporal information of the input and features of the diagram by considering the timing.

In 3DConv and 2DConv, the major difference is that inputs and features have become temporal, in that one dimension has been enhanced. Thereafter, 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.

6.3 Improved dataset

To ascertain the physical-to-virtual connection of the tower crane digital twin, a critical step in this research is to base tower crane operational mode recognition on deep learning. This means there is a need to improve recognition performance and role to validate the effortless development of future research. Improvement methods based on the 2D tower crane image characteristics will be recommended in this part of the document, following this circumstance as stated. In the operational mode recognition process, the original 3DesNet restrictions and limitations will be stated.

6.3.1 Edge extraction

Edge extraction, as stated in Section 5.4.2, will be employed in operational mode recognition of the lower crane. The Sobel operator will be used to process the divided tower crane. The operator, sometimes called the Sobel–Feldman operator or Sobel filter, is used in image processing and computer vision, particularly within edge detection algorithms where it creates an image emphasizing edges. The values of the brightness and darkness of an image can be calculated by the operator by combining the3 Gaussian smoothing, the 2-D convolution operator and the differential derivation. Edge direction information and noise smoothing effect

have greater accuracy. Normally, when there is no requirement for a high level of accuracy, edge detection tends to use the Sobel operator. The tower crane image processed by the Sobel operator and the original image of the tower crane is shown in the following Figure 6.3. The noise in the picture can be removed, the edge contour heightened, and the image sharpened by edge extraction. Edge extraction is conducted in five stages by the Sobel operator.

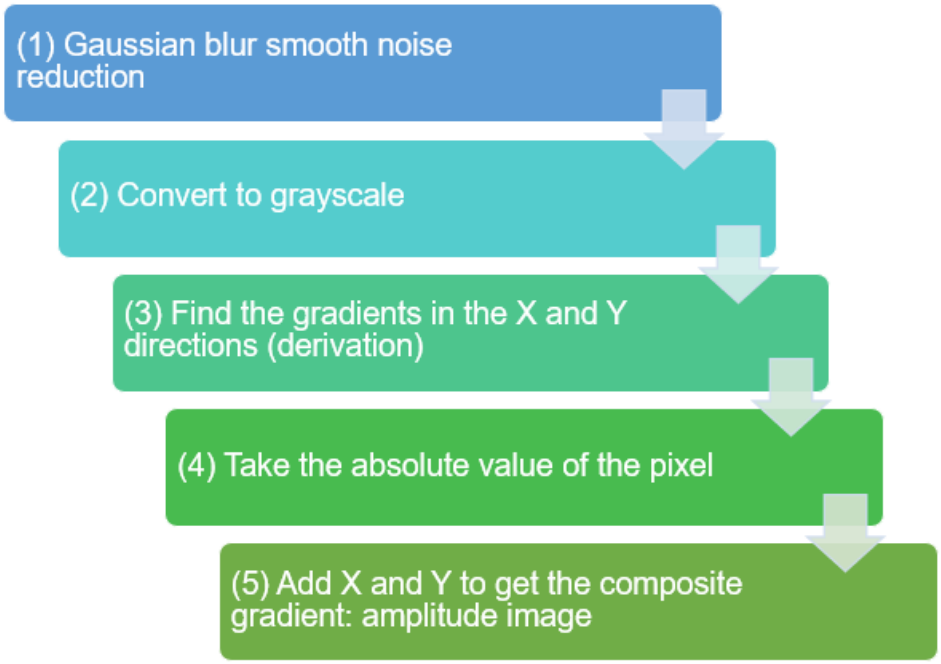


Figure 6.3: Steps of sobel operator to extract edge

6.3.2 Data augment

A total of 1373 groups of image folders are sorted once the divided tower crane images have been sorted. Twenty frames of tower crane operational images are included in each folder. A division is made into 1167 sets of training set data, 119 sets of Dev datasets and 87 sets of test datasets from the 1373 folders. The precision after machine learning is not ideal as the total volume of dataset is comparatively small. As a prerequisite for applying deep learning networks, there is a need for a large-volume dataset. To enhance the accuracy of the deep learning model data augmentation is needed.

Data augmentation methods are presented in Chapter 3.7. For the trained model to have a greater generalizability, the dataset needs to be as diverse as possible, where the data augment can increase the training dataset. By preventing the network from incorporating irrelevant features, data augmentation can improve the pertinent data, and encourage improved data-

related performance. This will greatly improve the overall performance. A data augmentation method employing rotated images was applied in 3DesNet. Any angle rotated around a set point describes image rotation, where the image centre point is the default, and the transformation matrix is:

$$H = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{Eq (6-4)}$$

In the model training of tower crane operation mode recognition. The tower crane was rotated 10ºand 20º clockwise and counter clockwise. Figure 6.4 shows that the crane was rotated by +10º and -10º and the original image by +20º and -20º. The number of datasets was increased from 1373 to 6865 by this operation. Subsequently, this will greatly improve the overall performance of the model.

Figure 6.4: Dataset after data augmentation and edge extraction

6.4 Experiment verification results

This section uses the tower crane image dataset from chapter 3 and the dataset optimization strategy introduced in Section 6.3 and analyzes the improved 3DResnet to verify the mode recognition capability of the tower crane.

6.4.1 Experimental setup

Three groups of experiments were conducted to compare and test the superiority of the improved operation mode recognition algorithm using 1,373 sets and 27,460 pieces of annotated tower crane image data. In the first set of experiments, the ResNet34, LSTM+CNN, CNN+LSTM, Optical flow+RNN, and 3DResNet34 models were trained separately to find the most suitable one for future training. In the second set of experiments, four 3DResNet algorithms with different depths (3DResNet18, 3DResNet34, 3DResNet50, 3DResNet101) were compared. The third group of experiments selected the best two 3DResNet algorithms, compared them with the improved algorithm with different improvement strategies, and analyzed the impact of the improved strategy on the model's performance through ablation experiments. The optimizer's weight decay was set to 0.0005, the initial learning rate was 0.002, the number of iterations was 500, and the batch size was 16.

6.4.2 Experimental environment

The following experimental hardware was used for the present study: Intel(R) Core (TM) i9-11900F@2.50GHz CPU; 64.0 GB memory; and an NVIDIA GeForce GTX 3090 24G graphics card. Python 3.7 was used as the programming language, TensorFlow GPU as the deep learning framework, Cuda 10.2 for GPU acceleration, and OpenCV4.0 for image preprocessing.

The study adopted the idea of transfer learning and used a pre-trained model to improve the convergence speed of the 3DResnet tower crane operational mode recognition algorithm. Training was conducted using the approximate joint method. Different learning rates and iterations were selected depending on the depth of the model. Since 3DResNet18 had the lowest depth, the number of iterations selected was relatively small. Given the large size of the image dataset, the appropriate learning rate had to be selected; this increased the speed of the model training. The use of a pre-trained deep learning model for inference operation ordinarily comprises the steps outlined below (Figure 6.5).

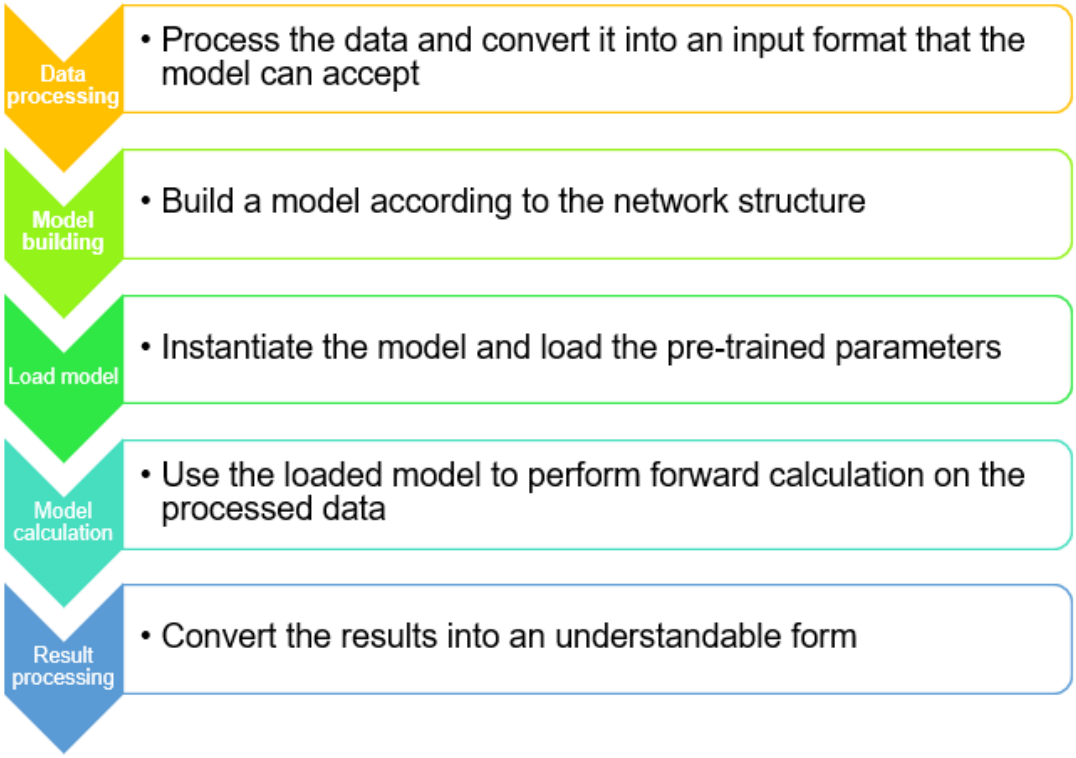


Figure 6.5: steps of pre-trained deep learning model

6.4.3 Evaluation indicators: Accuracy, loss, dev loss

Accuracy (see Section 5.5.3) and loss function and dev loss (loss in dev dataset) were used for the improved tower crane operation mode recognition algorithm:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6\text{-}5)$$

6.4.4 Results and analysis of the experiments

6.4.4.1 Comparison of candidate algorithms

After completing the object detection of the tower crane, we analyzed its operation mode. Because there were no algorithms available to do this, we proceeded as follows (and the algorithm with the highest recognition accuracy was selected for further research). A depth of 34 layers was selected for the traditional ResNet algorithm, and based on the initial dataset, ResNet34, LSTM+CNN, CNN+LSTM, Optical flow+RNN, and 3DResNet34 were trained separately.

For Model 1, 2DResNet with 34 layers was selected. Model 2 used LSTM as input for 2DResNet34. This allowed us to extract the long-distance features of the tower crane image dataset and obtain information about the time series before and after fusion. We then used 2DResNet34 to extract local features and inputted the fully connected layer after transformation. Model 3 used 2DResNet34 to extract the local features of the tower crane image, then used LSTM to extract their long-distance features. Finally, these features were transformed and input to the fully connected layer. Model 4 used optical flow to process the dataset and obtain the optical flow image of the movement of the tower crane. Model 5 used 3DResNet 34 to train the tower crane operational mode dataset 3DResnet34 = 2DResNet34 + 3DConv, and 3DResNet34 used a residual structure to connect each 3DConv module. The convolution kernel of 3DResNet34 increased one dimension on the basis of 2D and added parameters T [in_channels, out_channels, T].

Table 6.1: Comparison of different types of algorithms

Model	Method	Precision
Model 1	ResNet34	0.40
Model 2	LSTM+CNN	0.57
Model 3	CNN+LSTM	0.48
Model 4	Optical flow+RNN	0.55
Model 5	3DResNet 34	0.75

Table 6.1 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%); 2DResnet does not have 3DConv layers, unlike 3DResNet. When dealing with spatial and temporal order problems, the characteristics of tower crane motion cannot include learning, so the accuracy of this model was relatively low. Model 3 used 2DResNet34 to extract the local features of tower crane images, then used LSTM to extract their long-distance features. However, the accuracy of Model 3 was 48%. Thus, 2DResNet has a poor effect on the recognition of tower crane rotation, and the long-distance feature collection of local characteristics of the latter could not be used to recognize operation overall. These algorithms did not, therefore, learn the operational characteristics of the tower crane. The training detection precision of Model 2 with LSTM+CNN or Model 4 with optical flow+RNN were 57% and 55% respectively, indicating that the algorithms learned some of the characteristics

of the tower crane's operations, but the error detection and missed detection were quite serious. In future research, we would use optical flow as input for 3D CNN to achieve the best model performance. Finally, the accuracy rate (75%) of Model 5 (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, notwithstanding certain enhancements and the addition of deeper layers.

6.4.4.2 Comparison of different depth of 3Dresnet

Convolutional neural networks with 3D convolution kernels have recently received considerable attention, and they have begun to outperform traditional 2D CNNs in large video datasets. A 3-D CNN can extract spatiotemporal features from a video more intuitively and effectively. Resnet is one of the most successful architectures in image classification because it allows the signal to bypass the previous layer and move to the next layer in sequence. Since these connections flow through the gradients of the network from later layers to earlier layers, Resnet can be used for very deep network training.

In the previous section, we tried different algorithms and compared five models: ResNet34, LSTM+CNN, CNN+LSTM, Optical flow+RNN, and 3DResNet. We eventually chose 3DRseNet as the algorithm for tower crane mode operation detection. The accuracy of 3DResNet increases as the depth increases to 152 layers but not after the depth reaches 200 layers. In addition to the experiments outlined above, we compared 3DResnet algorithms with different depths; 3DResNet18, 3DResNet34, 3DResNet50, 3DResNet101 datasets were used separately in training the model to select the most appropriate depth for optimization. Because 3DResNet 152 was too deep, and the existing dataset may have led to overfitting, it was not selected. Table 6-2 displays the results of the comparisons.

Table 6.2: Comparison of Resnet algorithm with different depth

Method	Precision	Training time
3DResNet18	0.64	192
3DResNet34	0.75	158
3DResNet50	0.82	280
3DResNet101	0.86	293

Table 6.2 shows that the deeper the network depth of the 3DResNet structure, the higher the accuracy. The recognition accuracy of Resnet18 was only 0.64, while that of 3DResNet 101 reached 86%. Compared with 3DResNet 34, the recognition precision of 3DResNet50 and 3DResNet101 improved by 9.33% and 14.7%, respectively; both reached more than 80% overall. From Figure 6.6, we can see that when the depth of layers of 3DResNet increases, the accuracy of the model improves; the loss of the models was around 0.1-0.2 as a result. Subsequent experiments involve the use of the 3DResNet50 and the 3DResNet101 algorithms to train datasets processed by data augmentation and edge extraction.

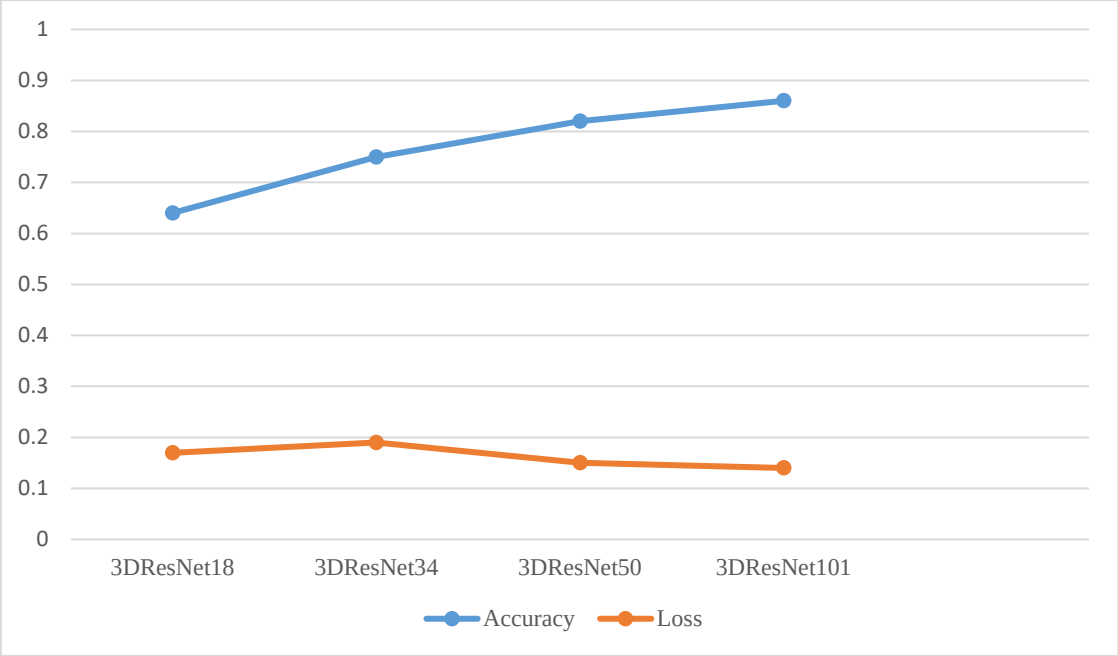


Figure 6.6: Accuracy and loss of different depth of 3DResNet

6.4.4.3 Augment dataset

The 3DResNet algorithms 3DResNet50 and 3DResNet101 achieved precision rates of 82% and 86%, respectively. Section 6.4 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 dataset to 6,865 sets and were verified with the new 565 dev dataset. In the present section, we measure accuracy, loss, and dev loss to determine the pros and cons of the model. The Table 6.3 and Figure 6.7 below present the precision and loss of the 3DResNet50 and 3DResNet101 and a histogram comparing the accuracy of the previous

and augmented datasets. From them we can see that the loss was around 0.10–0.15; however, the loss of 3DResNet101 with the augmented dataset was 0.24. This may have been due to a lack of data and the increase in depth of the layers (thus resulting in the overfitting of the model). The best accuracy rate was 87% (with the augmented dataset in 3DResNet 50 and 3DResNet101).

Table 6.3: Comparison of precision of different depth of 3DResNet

Method	Accuracy	Loss	Training time
3DResNet50	0.82	0.15	280
3DResNet50（augmented）	0.87	0.11	1024
3DResNet101	0.86	0.14	293
3DResNet101（augmented）	0.87	0.24	1117

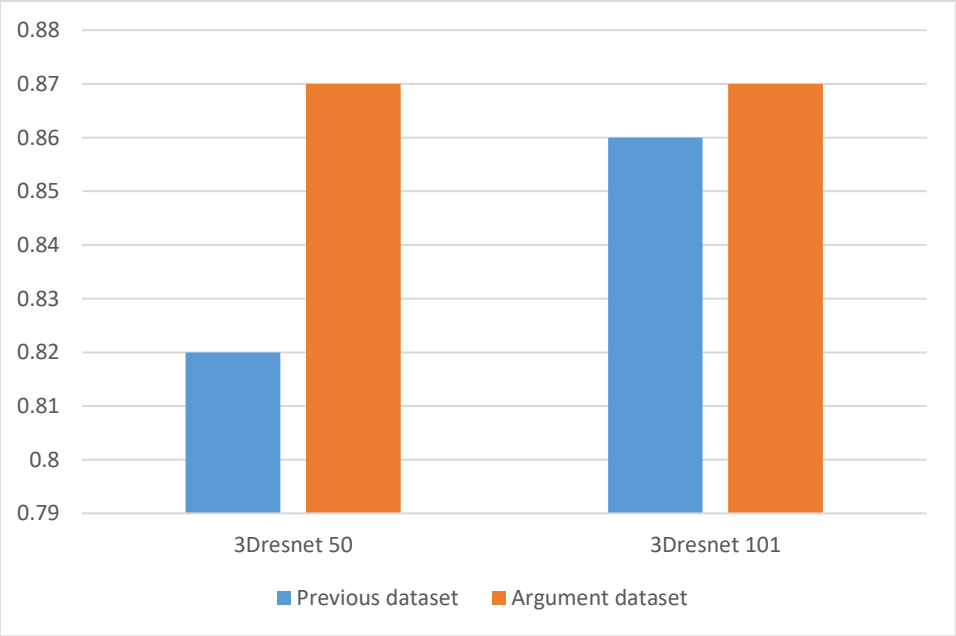


Figure 6.7: Histogram of accuracy comparison of previous and augmented dataset

Figure 6.8 shows the accuracy curve of the algorithms of different depths (3DResnet18, 3DResnet34, 3DResnet50, 3DResnet101, improved 3DResnet50, and improved 3DResnet101) and indicates that the augmented dataset can significantly improve detection accuracy. The recognition accuracy of the ResNet50 with the original dataset was 82%, but

after using the augmented dataset, it reached 87%. Meanwhile, the recognition accuracy of the original data set of ResNet101 was 86%, and after augmenting the dataset, it was 87%. After taking into consideration the longer training and processing time needed by 3DResNet100, the improved 3DResNet50 was chosen.

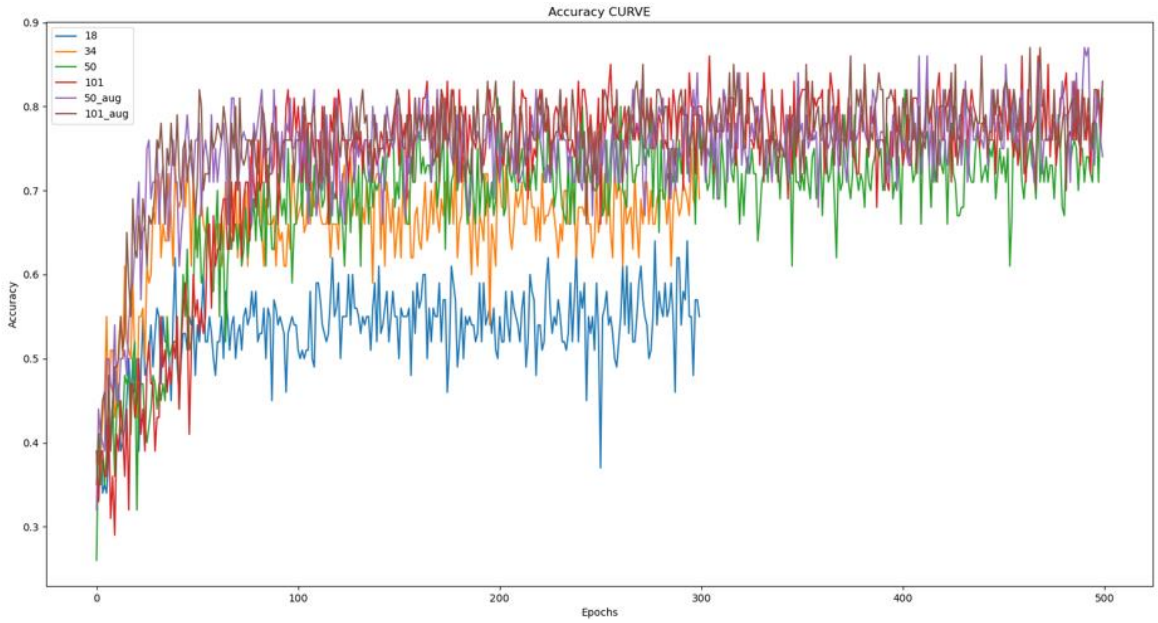


Figure 6.8: Accuracy curve of different depth of 3DresNet

Figure 6.9 shows the fit curve of the accuracy of the above algorithms of different depths. In general, compared with Figure 6.8, 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.

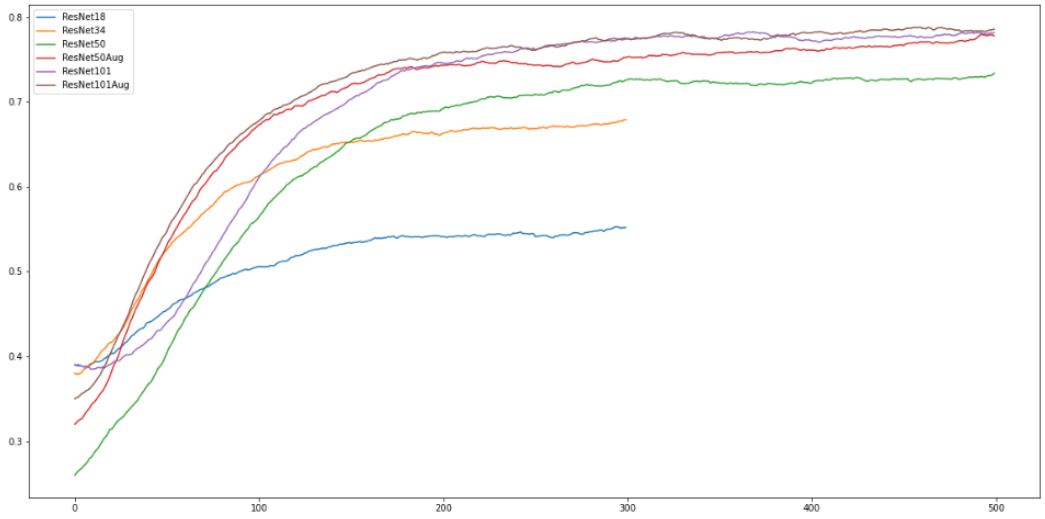


Figure 6.9: Fit curve of accuracy of different depth ResNet

From Figure 6.10, we can see that the loss oscillation is very apparent, and the loss curve is extremely erratic. There are two possible explanations for these phenomena. First, the loss curve of 3DResNet18 increased in the 300 epochs, which indicates that the model was underfitted and 18 layers were insufficient (and so deeper layers should be added). Second, the loss curve of 3DResNet50 and 3DResNet 101 was quite obvious. These models were overfitting to some extent, but some solutions may be suggested. The first is to determine the optimal model by early termination: during the training process, it may be the case that the final accuracy of the training is not as high as the previous epoch, so we can directly terminate the training and use the previous model as the best option. The second method is to use regularization to constrain the model. This would involve the weight decay method of the optimizer; that is, in the later stage of training, the gradient of the weight is rendered slower and slower through the decay factor. The third method would be to adjust the network structure. Last but not least, the volume of training data could be increased because overfitting may result when the dataset is too small or there is no data augmentation.

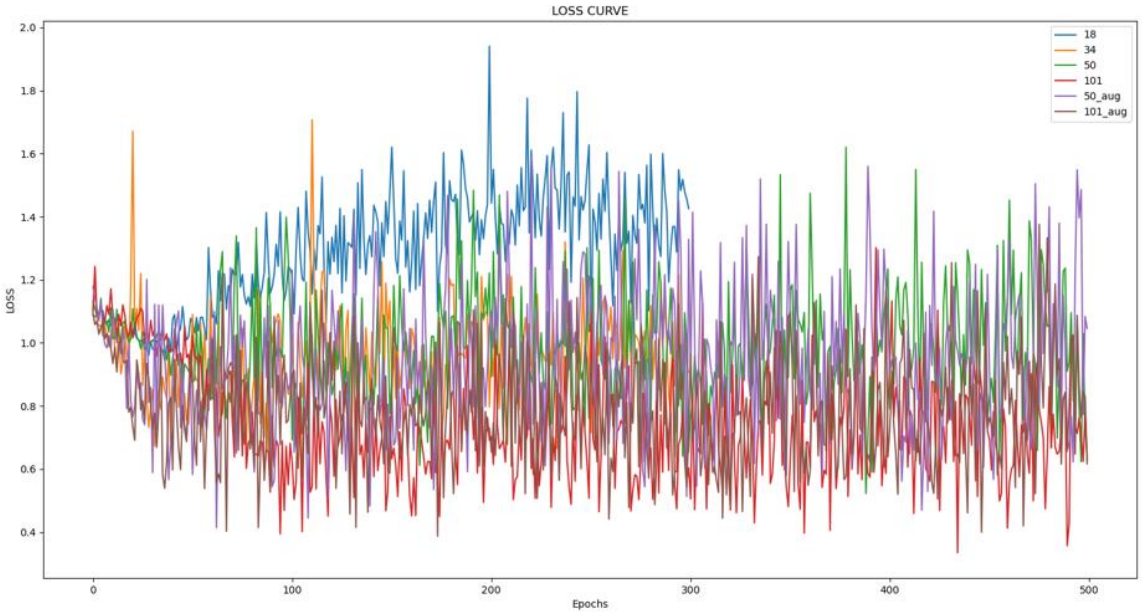


Figure 6.10: Dev loss curve of different depth of 3DResNet

The Figure 6.11 shows the training loss of the 3DResNet50 with original dataset (before data augmentation) and the augmented dataset (after data augmentation). It can be seen from the figure that the final training loss of these two situations is about 0.20, but when using the original dataset (before augmentation), the overall training loss is fluctuating with some irregular rising and falling. What's more, after using augmented dataset, the curve of training loss tended to be stable earlier after 50 epochs compared with original dataset.

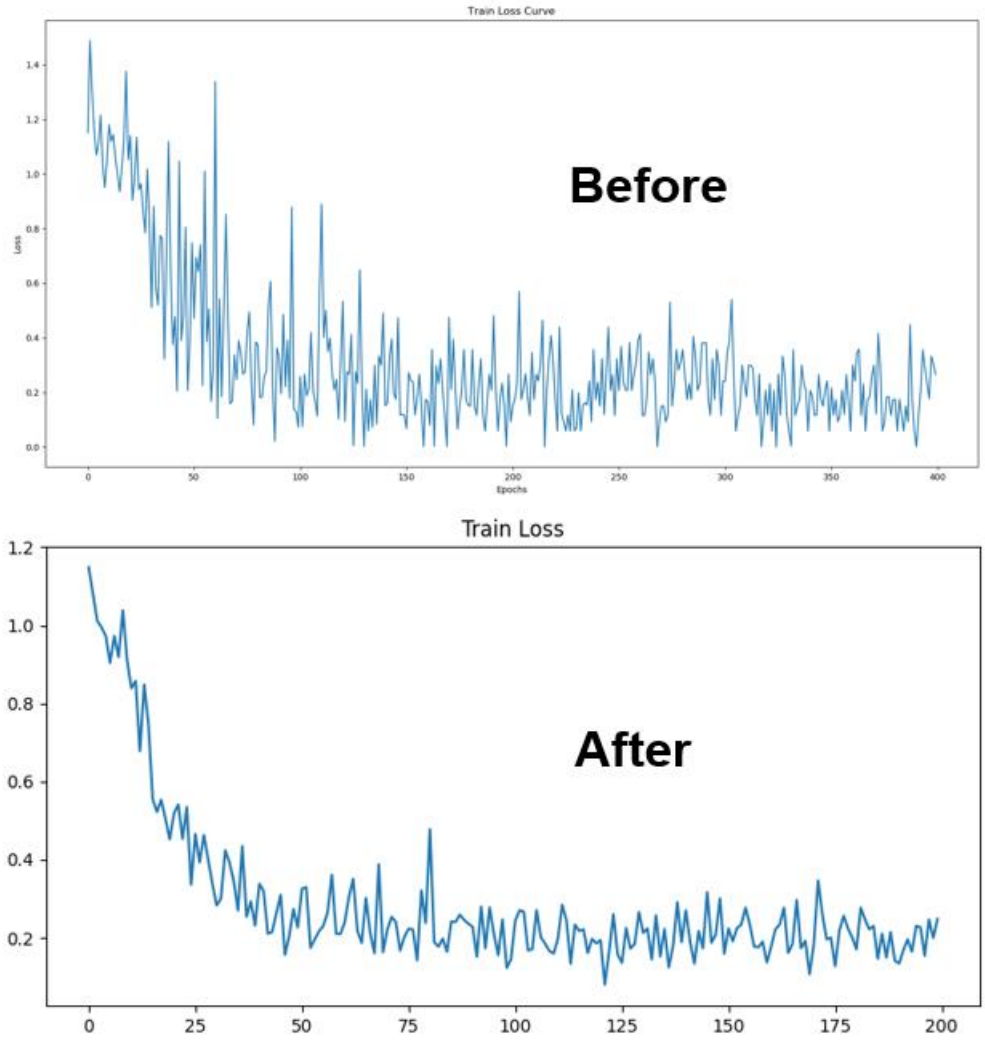


Figure 6.11: Training loss of the 3DResNet50 before and after data augmentation

6.4.4.4 Ablation experiment

To better analyze and verify the effectiveness of the 3DResNet improvement strategy used in the present chapter, we selected 3DResNet50 and used ablation experiments to compare the impact of data augmentation and edge extraction on the tower crane mode recognition dataset. The results under different improvement strategies are described below in Table 6.4.

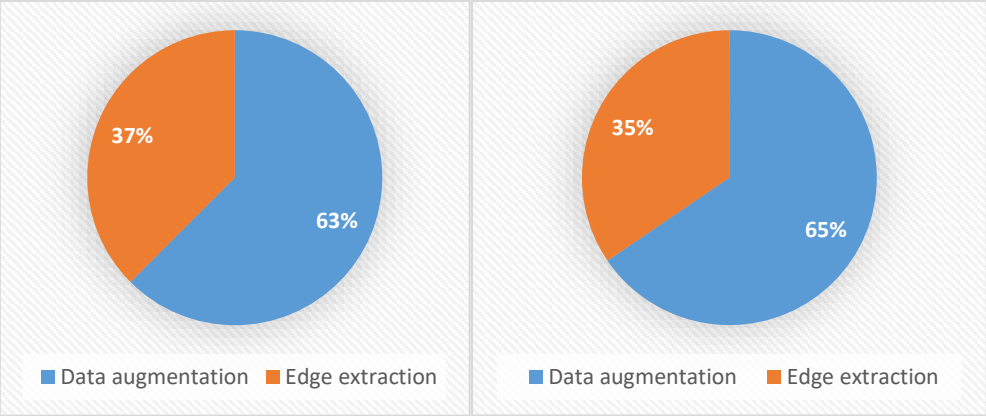
Table 6.4: Detection results of Yolov5 under different improvement strategies

	Model 1	Model 2	Model 3	Model 4
Data augment	×	√	×	√
Edge extraction	×	×	√	√
Accuracy	0.8235	0.8655	0.8487	0.8739
Loss	0.1521	0.1168	0.1335	0.1060

Model 1 represents the 3DResNet50yolov5x model; Model 2 represents an improved model that augmented the dataset by rotating the tower crane images; Model 3 represents a model that only performs edge extraction on images; and Model 4 represents the improved 3DResNet50 model that was proposed above.

In the case of Model 1, the original 3DResNet50 had an accuracy of 82.35% for tower crane operation recognition, with a loss of 0.1521. For Model 4, the results were 87.39% and 0.1060, respectively, which indicated that both methods can improve tower crane operational mode recognition. For Models 1 and 2, the precision improved by a substantial 5.1%. After dataset augmentation, the accuracy of the models increased, and the loss decreased. The results from Models 1 and 3 suggest that edge extraction can reduce the noise in the image, thus allowing the algorithms to better identify the operation mode of the tower crane.

Finally, the present study examined the impact of different improvement strategies on the final results from Model 4 based on the above table. Figure 6.12 shows the contribution rates of different optimization strategies to model accuracy and loss. The figure below demonstrates that from the perspective of accuracy increments, the contribution of data augmentation is relatively high, and the corresponding accuracy rate under the single strategy reaches 5.1%, which accounts for 63% of the overall improvement in accuracy. Under the edge extraction strategy, the corresponding accuracy rate increased by 3.1% and the overall accuracy of the station by 37%. This shows that in the case of the 3DResNet algorithm, the use of an augmented dataset can improve accuracy. In terms of loss, the effect of data augmentation is also greater than that of edge extraction, accounting for 65% and 35% of the improvement in recall rates, respectively. Therefore, this method appeared to improve the generalizing capacity of the model and reduce loss.



(a)Contribution rate for accuracy (b) Contribution rate for recall

Figure 6.12: Contribution of the improvement under different strategies

Through comparative analysis, it has been found that the 3DResNet50 improvement strategies described herein can improve the detection accuracy of the model and that the comprehensive improvement strategy is more effective than the single improvement strategy.

6.5 Summary

In the tower crane digital twin, the present study has completed the physical-to-virtual connection through pre-modeling and tower crane image recognition. The operation mode of the tower crane now needs to be reflected in the virtual model. We used machine learning to evaluate the tower crane in real-time operation. The present chapter has discussed the development of gesture recognition based on the characteristics of tower crane operation video and explained the detection methods and principles of LSTM, CNN, and ResNet algorithms. Because 3DResNet's accuracy rate was insufficient, a series of optimization strategies have been proposed.

The ResNet34, LSTM+CNN, CNN+LSTM, Optical flow+RNN, and 3DResNet34 algorithms were used to train the initial tower crane operation mode dataset. The precision of 3DResNet reached 75%; the algorithm will be optimized in future studies. Four different depth network structures of 3DResNet (i.e., 3DResNet18, 3DResNet34, 3DResNet50, and 3DResNet101) were then trained, and the best was finally selected. Thereafter, two methods were proposed to improve the recognition accuracy of 3DResNet50 and reduce the loss of the model. The first method involved augmenting the dataset by rotating the picture clockwise and counter

clockwise by 10° and 20°, extending the new dataset by five times that of the previous datasets. The second method eliminated the noise in the image using the Sobel operator in edge extraction and compared the model accuracy of the original and improved datasets after 3DResNet50 and 3DResNet101 training. Finally, through an ablation experiment, the contribution of these two improvements to accuracy and loss was assessed.

In conclusion, the single image detection speed of 3DResNet50 was the fastest and most accurate. The 3DResNet50 of the augmented data set was therefore selected as the final tower crane operation mode recognition algorithm.

7. Conclusion and prospects

Digitalization has become an integral part of contemporary society, but the construction industry has been slow to recognize it. Additionally, the cost of applying the BIM remodeling method is high, it is not intelligent enough, and its capacity to generalize is poor. Introducing the concept of the digital twin into construction engineering may help accelerate the process of digitalization in engineering and ensure that operations are conducted safely. The present study involved the application of the concept of the digital twin to a tower crane on a construction site. The main elements of the study and conclusions are presented below:

(1) A literature review of the concept and history of the digital twin was carried out. The application of the digital twin in different fields was examined and the enabling technologies for realizing digital twins were summarized. The feasibility of applying the concept to multidisciplinary collaborations with tower cranes was explored.

(2) A framework for a tower crane digital twin was proposed. It comprised five elements: the physical entity and environment; the virtual entity and environment; the physical-to-virtual connection; the virtual-to-physical connection; and the digital twin process. It was proposed that the tower crane digital twin should feature real-time and autonomous characteristics.

(3) Because the target recognition algorithm and the target operating state recognition algorithm require a large number of datasets for parameter training and performance testing and that open source, large-scale, well-labeled tower crane image and video datasets are currently lacking in engineering, it was proposed that tower crane datasets should be generated and optimized using image preprocessing, image annotation, and image dataset augmentation. Subsequently, a tower crane dataset including more than 500 videos and 45,588 images was created. A total of 1,373 sets of data were obtained, including 27,460 tower crane images. 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. The new separate image included one crane as a whole.

(4) The target detection algorithms were reviewed, the development of computer vision and deep learning was discussed, and the state of the art of target detection algorithms was briefly considered. The improved yolov5 algorithm was proposed for tower crane object detection. The yolov5 series algorithms were presented in terms of related content (i.e., bounding box,

anchor boxes, intersection over union and loss), the yolov5 framework, and CIOU loss function and nms. Based on the understanding of the above concepts, it was proposed that DIoU_nms be used to improve the prediction accuracy of the yolov5 algorithm in situations where tower cranes overlap. The initial accuracy rate of yolov5x was 93.85%. The concept of edge extraction, which was used to reduce the noise in the tower crane image in the present study, was introduced. An ablation experiment was designed to judge the superiority of the two improved methods. The final accuracy rate was 95.45%.

(5) After comparing the LSTM and RNN algorithms, 3DResNet was used to identify the motion state of the tower crane on the sorted operating images; after data augmentation, the final accuracy rate reached 87%. Through the above method, the operation of the tower crane can finally be expressed and simulated in the virtual space.

Although the present study yielded some useful results, it has three limitations, and therefore further researches are proposed to address these limitations accordingly.

(1) The study was confined to the application of digital twin physical-to-virtual connections. Future researchers might therefore 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 other types of tower cranes, thus building a lifecycle digital twin of the construction site as a whole.

(2) The data for tower crane attitude recognition was insufficient. Additional scenarios (as well as tower cranes) and modified algorithms might be added to increase the detection precision rate beyond 95%. As was noted above, the tower crane operation mode detection of the 3DResNet model was 87%. In the future work, we can obtain high-quality, labelled image data and develop some algorithms with higher recognition accuracy. What's more, the residual network can be deepened to improve the quality of 3DResNet algorithm itself.

(3) The data collected by the computer vision method was limited. From the perspective of tower crane safety, a sensor can reflect some abnormal operational status in real-time. Therefore, in future studies, more sensors could be added to monitor quantitatively the safe operation of the tower crane. Other ideas could be pursued. For example, a digital twin with

sensors could be used to simulate the operation of the tower crane in the virtual model while predicting the development of future operations. This predictive ability could be used to forewarn managers of potential accidents.

Reference

1. Lu SCY, Li D, Cheng J, Wu CL, Int Inst Prod Engn RES, Int Inst Prod Engn RES, editors. A model fusion approach to support negotiations during complex engineering system design. 47th General Assembly of CIRP - Manufacturing Technology; 1997 Aug 24-30; Tianjin, Peoples R China. 3001 BERN: Hallwag Publishers; 1997.
2. Maropoulos PG, Ceglarek D. Design verification and validation in product lifecycle. CIRP Ann-Manuf Technol. 2010;59(2):740-59.
3. Grieves MW, editor Virtually Indistinguishable Systems Engineering and PLM. 9th IFIP WG 51 International Conference on Product Lifecycle Management (PLM); 2012 Jul 09-11; Univ Quebec, Ecole Technologie Superieure, Montreal, CANADA. BERLIN: Springer-Verlag Berlin; 2012.
4. Glaessgen EH, Stargel DS. The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles. 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference - Special Session on the Digital Twin.
5. Tao F, Sui FY, Liu A, Qi QL, Zhang M, Song BY, Guo ZR, Lu SCY, Nee AYC. Digital twin-driven product design framework. Int J Prod Res. 2019;57(12):3935-53.
6. Rosen R, von Wichert G, Lo G, Bettenhausen KD, editors. About The Importance of Autonomy and Digital Twins for the Future of Manufacturing. 15th IFAC Symposium on Information Control Problems in Manufacturing; 2015 May 11-13; Ottawa, CANADA. AMSTERDAM: Elsevier Science Bv; 2015.
7. Zhang H, Liu Q, Chen X, Zhang D, Leng JW. A Digital Twin-Based Approach for Designing and Multi-Objective Optimization of Hollow Glass Production Line. IEEE Access. 2017;5:26901-11.
8. Grieves M, Vickers J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In: Kahlen F-J, Flumerfelt S, Alves A, editors. Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches. Cham: Springer International Publishing; 2017. p. 85-113.
9. Padovano A, Longo F, Nicoletti L, Mirabelli G, editors. A Digital Twin based Service Oriented Application for a 4.0 Knowledge Navigation in the Smart Factory. 16th IFAC Symposium on Information Control Problems in Manufacturing (INCOM); 2018 Jun 11-13; Bergamo, ITALY. AMSTERDAM: Elsevier Science Bv; 2018.
10. Zheng Y, Yang S, Cheng HC. An application framework of digital twin and its case study. J Ambient Intell Humaniz Comput. 2019;10(3):1141-53.
11. Wang XJ, Zhang Y, Zhang J, Fu CL, Zhang XL. Progress in urban metabolism research and hotspot analysis based on CiteSpace analysis. J Clean Prod. 2021;281:12.
12. Zhang XL, Zhang Y, Wang YF, Fath BD. Research progress and hotspot analysis for reactive nitrogen flows in macroscopic systems based on a CiteSpace analysis. Ecol Model. 2021;443:10.
13. Azam A, Ahmed A, Wang H, Wang YE, Zhang ZT. Knowledge structure and research progress in wind power generation (WPG) from 2005 to 2020 using CiteSpace based scientometric analysis. J Clean Prod. 2021;295:22.
14. Tuegel EJ, Ingraffea AR, Eason TG, Spottswood SM. Reengineering Aircraft Structural Life Prediction Using a Digital Twin. International Journal of Aerospace Engineering. 2011;2011:154798.
15. Schleich B, Anwer N, Mathieu L, Wartzack S. Shaping the digital twin for design and production engineering. CIRP Ann-Manuf Technol. 2017;66(1):141-4.
16. Wang JJ, Ye LK, Gao RX, Li C, Zhang LB. Digital Twin for rotating machinery fault diagnosis in smart manufacturing. Int J Prod Res. 2019;57(12):3920-34.

17. Liu Q, Zhang H, Leng JW, Chen X. Digital twin-driven rapid individualised designing of automated flow-shop manufacturing system. Int J Prod Res. 2019;57(12):3903-19.

18. Fedorko G, Molnar V, Vasil M, Salai R. Proposal of digital twin for testing and measuring of transport belts for pipe conveyors within the concept Industry 4.0. Measurement. 2021;174:14.

19. Bhatti G, Mohan H, Singh RR. Towards the future of smart electric vehicles: Digital twin technology. Renew Sust Energ Rev. 2021;141:18.

20. Graessler I, Poehler A, editors. Intelligent control of an assembly station by integration of a digital twin for employees into the decentralized control system. 4th International Conference on System-Integrated Intelligence - Intelligent, Flexible and Connected Systems in Products and Production; 2018 Jun; Leibniz Univ Hannover, Hannover Ctr Prod Technol, Hanover, GERMANY. AMSTERDAM: Elsevier Science Bv; 2018.

21. Nikolakis N, Alexopoulos K, Xanthakis E, Chryssolouris G. The digital twin implementation for linking the virtual representation of human-based production tasks to their physical counterpart in the factory-floor. Int J Comput Integr Manuf. 2019;32(1):1-12.

22. Shengli W. Is Human Digital Twin possible? Computer Methods and Programs in Biomedicine Update. 2021;1:100014.

23. Kruize JW. Digital twins in farm management: illustrations from the FIWARE accelerators Smart AgriFood and Fractals. 2018.

24. Pylianidis C, Osinga S, Athanasiadis IN. Introducing digital twins to agriculture. Computers and Electronics in Agriculture. 2021;184:105942.

25. Lopes MR, Costigliola A, Pinto R, Vieira S, Sousa JMC. Pharmaceutical quality control laboratory digital twin - A novel governance model for resource planning and scheduling. Int J Prod Res. 2020;58(21):6553-67.

26. Leng JW, Zhang H, Yan DX, Liu Q, Chen X, Zhang D. Digital twin-driven manufacturing cyber-physical system for parallel controlling of smart workshop. J Ambient Intell Humaniz Comput. 2019;10(3):1155-66.

27. Eckhart M, Ekelhart A, Assoc Comp M, editors. Towards Security-Aware Virtual Environments for Digital Twins. 4th ACM Workshop on Cyber-Physical System Security (CPSS) / 5th ACM Asia Public-Key Cryptography Workshop (APKC); 2018 Jun 04; Incheon, SOUTH KOREA. NEW YORK: Assoc Computing Machinery; 2018.

28. Pan Y, Zhang LM. A BIM-data mining integrated digital twin framework for advanced project management. Autom Constr. 2021;124:15.

29. VanderHorn E, Mahadevan S. Digital Twin: Generalization, characterization and implementation. Decis Support Syst. 2021;145:11.

30. Tao F, Cheng JF, Qi QL, Zhang M, Zhang H, Sui FY. Digital twin-driven product design, manufacturing and service with big data. Int J Adv Manuf Technol. 2018;94(9-12):3563-76.

31. Schroeder GN, Steinmetz C, Pereira CE, Espindola DB, editors. Digital Twin Data Modeling with AutomationML and a Communication Methodology for Data Exchange. 4th IFAC Symposium on Telematics Applications (TA); 2016 Sep 06-09; Porto Alegre, BRAZIL. AMSTERDAM: Elsevier; 2016.

32. Catelani D. From the Digital-Twin to the Cyber Physical System Using Integrated Multidisciplinary Simulation: Virtualization of Complex Systems. 2021. p. 18-39.

33. Cheng JF, Chen WH, Tao F, Lin CL. Industrial IoT in 5G environment towards smart manufacturing. J Ind Inf Integr. 2018;10:10-9.

34. Deng TH, Zhang KR, Shen ZJ. A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. J Manag Sci Eng. 2021;6(2):125-34.

35. White G, Zink A, Codeca L, Clarke S. A digital twin smart city for citizen feedback.

Cities. 2021;110:11.

36. Fuller A, Fan Z, Day C, Barlow C. Digital Twin: Enabling Technologies, Challenges and Open Research. IEEE Access. 2020;8:108952-71.

37. Jiang F, Ma L, Broyd T, Chen K. Digital twin and its implementations in the civil engineering sector. Autom Constr. 2021;130:16.

38. Miller A, Alvarez R, Hartman N. Towards an extended model-based definition for the digital twin. Computer-Aided Design and Applications. 2018;15:1-12.

39. Jiang HF, Qin SF, Fu JL, Zhang J, Ding GF. How to model and implement connections between physical and virtual models for digital twin application. J Manuf Syst. 2021;58:36-51.

40. Hu L, Nguyen N-T, Tao W, Leu MC, Liu XF, Shahriar MR, Al Sunny SMN. Modeling of Cloud-Based Digital Twins for Smart Manufacturing with MT Connect. Procedia Manufacturing. 2018;26:1193-203.

41. Tong X, Liu Q, Pi SW, Xiao Y. Real-time machining data application and service based on IMT digital twin. J Intell Manuf. 2020;31(5):1113-32.

42. Um J, Weyer S, Quint F, editors. Plug-and-Simulate within Modular Assembly Line enabled by Digital Twins and the use of AutomationML. 20th World Congress of the International-Federation-of-Automatic-Control (IFAC); 2017 Jul 09-14; Toulouse, FRANCE. AMSTERDAM: Elsevier Science Bv; 2017.

43. Kritzinger W, Karner M, Traar G, Henjes J, Sihn W, editors. Digital Twin in manufacturing: A categorical literature review and classification. 16th IFAC Symposium on Information Control Problems in Manufacturing (INCOM); 2018 Jun 11-13; Bergamo, ITALY. AMSTERDAM: Elsevier Science Bv; 2018.

44. Ghosh AK, Ullah A, Teti R, Kubo A. Developing sensor signal-based digital twins for intelligent machine tools. J Ind Inf Integr. 2021;24:18.

45. Wang YC, Tao F, Zhang M, Wang LH, Zuo Y. Digital twin enhanced fault prediction for the autoclave with insufficient data. J Manuf Syst. 2021;60:350-9.

46.Kutzke DT, Carter JB, Hartman BT. Subsystem selection for digital twin development: A case study on an unmanned underwater vehicle. Ocean Eng. 2021;223:15.

47. Liu SM, Lu YQ, Li J, Song DQ, Sun XM, Bao JS. Multi-scale evolution mechanism and knowledge construction of a digital twin mimic model. Robot Comput-Integr Manuf. 2021;71:17.

48. Yu G, Wang Y, Mao ZY, Hu M, Sugumaran V, Wang YK. A digital twin-based decision analysis framework for operation and maintenance of tunnels. Tunn Undergr Space Technol. 2021;116:19.

49. Lu YJ, Zhao ZC, Wei W, Kui Z. Digital twin product lifecycle system dedicated to the constant velocity joint. Comput Electr Eng. 2021;93:13.

50. Zhou GH, Zhang C, Li Z, Ding K, Wang C. Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing. Int J Prod Res. 2020;58(4):1034-51.

51. Zhuang CB, Liu JH, Xiong H. Digital twin-based smart production management and control framework for the complex product assembly shop-floor. Int J Adv Manuf Technol. 2018;96(1-4):1149-63.

52. Tao F, Zhang M. Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing. IEEE Access. 2017;5:20418-27.

53. Jones D, Snider C, Nassehi A, Yon J, Hicks B. Characterising the Digital Twin: A systematic literature review. CIRP J Manuf Sci Technol. 2020;29:36-52.

54. Bottani E, Assunta C, Murino T, Vespoli S. From the Cyber-Physical System to the Digital Twin: the process development for behaviour modelling of a Cyber Guided Vehicle in M2M logic2017.

55. Macchi M, Roda I, Negri E, Fumagalli L, editors. Exploring the role of Digital Twin for Asset Lifecycle Management. 16th IFAC Symposium on Information Control Problems in Manufacturing (INCOM); 2018 Jun 11-13; Bergamo, ITALY. AMSTERDAM: Elsevier Science Bv; 2018.

56. Qi QL, Tao F. Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison. IEEE Access. 2018;6:3585-93.

57. Luo WC, Hu TL, Zhang CR, Wei YL. Digital twin for CNC machine tool: modeling and using strategy. J Ambient Intell Humaniz Comput. 2019;10(3):1129-40.

58. Blanke M, Kinnaert M, Lunze J, Staroswiecki M. DIAGNOSIS AND FAULT-TOLERANT CONTROL, by M. Blanke, M. Kinnaert, J. Lunze and M. Staroswiecki, with contributions by Jochen Schröder, Springer, Berlin, 2003, hardback, xv + 571 pp., ISBN 3-540-01056-4 (£77.00). Robotica. 2004;22(3):347-8.

59. Badihi H, Zhang YM, Hong H. Fault-tolerant cooperative control in an offshore wind farm using model-free and model-based fault detection and diagnosis approaches. Appl Energy. 2017;201:284-307.

60. He R, Chen GM, Dong C, Sun SF, Shen XY. Data-driven digital twin technology for optimized control in process systems. ISA Trans. 2019;95:221-34.

61. Zhang C, Zhou GH, He J, Li Z, Cheng W, editors. A data- and knowledge-driven framework for digital twin manufacturing cell. 11th CIRP Conference on Industrial Product-Service Systems; 2019 May 29-31; Peoples R China. AMSTERDAM: Elsevier; 2019.

62. Heng A, Zhang S, Tan ACC, Mathew J. Rotating machinery prognostics: State of the art, challenges and opportunities. Mech Syst Signal Proc. 2009;23(3):724-39.

63. Haag S, Anderl R. Digital twin – Proof of concept. Manufacturing Letters. 2018;15:64-6.

64. Ayani M, Ganeback M, Ng AHC, editors. Digital Twin: Applying emulation for machine reconditioning. 51st CIRP Conference on Manufacturing Systems (CIRP CMS); 2018 May 16-18; Stockholm, SWEDEN. AMSTERDAM: Elsevier Science Bv; 2018.

65. Talkhestani BA, Jazdi N, Schloegl W, Weyrich M, editors. Consistency check to synchronize the Digital Twin of manufacturing automation based on anchor points. 51st CIRP Conference on Manufacturing Systems (CIRP CMS); 2018 May 16-18; Stockholm, SWEDEN. AMSTERDAM: Elsevier Science Bv; 2018.

66. Shen W, Yang C, Gao L. Address business crisis caused by COVID-19 with collaborative intelligent manufacturing technologies. IET Collaborative Intelligent Manufacturing. 2020;2(2):96-9.

67. Semeraro C, Lezoche M, Panetto H, Dassisti M. Digital twin paradigm: A systematic literature review. Comput Ind. 2021;130:23.

68. Shao GD, Helu M. Framework for a digital twin in manufacturing: Scope and requirements. Manufacturing Letters. 2020;24:105-7.

69. Aheleroff S, Xu X, Zhong RY, Lu YQ. Digital Twin as a Service (DTaaS) in Industry 4.0: An Architecture Reference Model. Adv Eng Inform. 2021;47:15.

70. Scime L, Singh A, Paquit V. A scalable digital platform for the use of digital twins in additive manufacturing. Manufacturing Letters. 2022;31:28-32.

71. Zhang J, Deng TM, Jiang HF, Chen HJ, Qin SF, Ding GF. Bi-level dynamic scheduling architecture based on service unit digital twin agents. J Manuf Syst. 2021;60:59-79.

72. Qi QL, Tao F, Zuo Y, Zhao DM, editors. Digital Twin Service towards Smart Manufacturing. 51st CIRP Conference on Manufacturing Systems (CIRP CMS); 2018 May 16-18; Stockholm, SWEDEN. AMSTERDAM: Elsevier Science Bv; 2018.

73. Lu YQ, Liu C, Wang KIK, Huang HY, Xu X. Digital Twin-driven smart

manufacturing: Connotation, reference model, applications and research issues. Robot Comput-Integr Manuf. 2020;61:14.

74. Leng JW, Zhou M, Xiao YX, Zhang H, Liu Q, Shen WM, Su QY, Li LZ. Digital twins-based remote semi-physical commissioning of flow-type smart manufacturing systems. J Clean Prod. 2021;306:15.

75. Rocha AD, Barata J. Digital twin-based optimiser for self-organised collaborative cyber-physical production systems. Manufacturing Letters. 2021;29:79-83.

76. Fan YP, Yang JZ, Chen JH, Hu PC, Wang XY, Xu JC, Zhou B. A digital-twin visualized architecture for Flexible Manufacturing System. J Manuf Syst. 2021;60:176-201.

77. Leng JW, Wang DW, Shen WM, Li XY, Liu Q, Chen X. Digital twins-based smart manufacturing system design in Industry 4.0: A review. J Manuf Syst. 2021;60:119-37.

78. Uhlemann THJ, Lehmann C, Steinhilper R, editors. The Digital Twin: Realizing the Cyber-Physical Production System for Industry 4.0. 24th CIRP Conference on Life Cycle Engineering (CIRP LCE); 2017 Mar 08-10; Kamakura, JAPAN. AMSTERDAM: Elsevier Science Bv; 2017.

79. Zhu ZX, Xi XL, Xu X, Cai YL. Digital Twin-driven machining process for thin-walled part manufacturing. J Manuf Syst. 2021;59:453-66.

80. Fang YL, Peng C, Lou P, Zhou ZD, Hu JM, Yan JW. Digital-Twin-Based Job Shop Scheduling Toward Smart Manufacturing. IEEE Trans Ind Inform. 2019;15(12):6425-35.

81. Dong YF, Tan RH, Zhang P, Peng QJ, Shao P. Product redesign using functional backtrack with digital twin. Adv Eng Inform. 2021;49:17.

82. Lu YJ, Zhao ZC, Wei W, Kui Z. Digital twin product lifecycle system dedicated to the constant velocity joint. Comput Electr Eng. 2021;93.

83. Maschler B, Braun D, Jazdi N, Weyrich M. Transfer learning as an enabler of the intelligent digital twin. Procedia CIRP. 2021;100:127-32.

84. Botkina D, Hedlind M, Olsson B, Henser J, Lundholm T, editors. Digital Twin of a Cutting Tool. 51st CIRP Conference on Manufacturing Systems (CIRP CMS); 2018 May 16-18; Stockholm, SWEDEN. AMSTERDAM: Elsevier Science Bv; 2018.

85. Dotoli M, Fay A, Miskowicz M, Seatzu C, editors. A Survey on Advanced Control Approaches in Factory Automation. 15th IFAC Symposium on Information Control Problems in Manufacturing; 2015 May 11-13; Ottawa, CANADA. AMSTERDAM: Elsevier; 2015.

86. Chakraborty S, Adhikari S. Machine learning based digital twin for dynamical systems with multiple time-scales. Comput Struct. 2021;243:15.

87. Uhlemann THJ, Schock C, Lehmann C, Freiberger S, Steinhilper R, editors. The Digital Twin: Demonstrating the potential of real time data acquisition in production systems. 7th Conference on Learning Factories (CLF); 2017 Apr 04-05; Darmstadt, GERMANY. AMSTERDAM: Elsevier Science Bv; 2017.

88. Mukherjee T, DebRoy T. Mitigation of lack of fusion defects in powder bed fusion additive manufacturing. J Manuf Process. 2018;36:442-9.

89. Opoku DJ, Perera S, Osei-Kyei R, Rashidi M. Digital twin application in the construction industry: A literature review. J Build Eng. 2021;40:15.

90. Boje C, Guerriero A, Kubicki S, Rezgui Y. Towards a semantic Construction Digital Twin: Directions for future research. Autom Constr. 2020;114:16.

91. Khajavi SH, Motlagh NH, Jaribion A, Werner LC, Holmstrom J. Digital Twin: Vision, Benefits, Boundaries, and Creation for Buildings. IEEE Access. 2019;7:147406-19.

92. Zhang L, Zhou LF, Horn BKP. Building a right digital twin with model engineering. J Manuf Syst. 2021;59:151-64.

93. Shanbari HA, Blinn NM, Issa RR. LASER SCANNING TECHNOLOGY AND BIM IN CONSTRUCTION MANAGEMENT EDUCATION. J Inf Technol Constr. 2016;21:204-

17.

94. Zhang GC, Vela PA, Karasev P, Brilakis I. A Sparsity-Inducing Optimization-Based Algorithm for Planar Patches Extraction from Noisy Point-Cloud Data. Comput-Aided Civil Infrastruct Eng. 2015;30(2):85-102.

95. Kaewunruen S, Lian Q. Digital twin aided sustainability-based lifecycle management for railway turnout systems. J Clean Prod. 2019;228:1537-51.

96. Lu RD, Brilakis I. Digital twinning of existing reinforced concrete bridges from labelled point clusters. Autom Constr. 2019;105:16.

97. Angjeliu G, Coronelli D, Cardani G. Development of the simulation model for Digital Twin applications in historical masonry buildings: The integration between numerical and experimental reality. Comput Struct. 2020;238:20.

98. Tran H, Nguyen TN, Christopher P, Bui DK, Khoshelham K, Ngo TD. A digital twin approach for geometric quality assessment of as-built prefabricated facades. J Build Eng. 2021;41:13.

99. Buckova M, Skokan R, Fusko M, Hodon R, editors. Designing of logistics systems with using of computer simulation and emulation. 13th International Scientific Conference on Sustainable, Modern and Safe Transport (TRANSCOM); 2019 May 29-31; Novy Smokovec, SLOVAKIA. AMSTERDAM: Elsevier; 2019.

100. Greif T, Stein N, Flath CM. Peeking into the void: Digital twins for construction site logistics. Comput Ind. 2020;121:9.

101. Ozturk GB. Digital Twin Research in the AECO-FM Industry. J Build Eng. 2021;40:12.

102. Tchana Y, Ducellier G, Remy S, editors. Designing a unique Digital Twin for linear infrastructures lifecycle management. 29th CIRP Design Conference (CIRP Design); 2019 May 08-10; Povoa de Varzim, PORTUGAL. AMSTERDAM: Elsevier; 2019.

103. Karve PM, Guo YL, Kapusuzoglu B, Mahadevan S, Haile MA. Digital twin approach for damage-tolerant mission planning under uncertainty. Eng Fract Mech. 2020;225:19.

104. Lydon GP, Caranovic S, Hischier I, Schlueter A. Coupled simulation of thermally active building systems to support a digital twin. Energy Build. 2019;202:19.

105. Jiang F, Ding YL, Song YS, Geng FF, Wang ZW. Digital Twin-driven framework for fatigue life prediction of steel bridges using a probabilistic multiscale model: Application to segmental orthotropic steel deck specimen. Eng Struct. 2021;241:13.

106. Love PED, Matthews J. The 'how' of benefits management for digital technology: From engineering to asset management. Autom Constr. 2019;107:15.

107. Lu QC, Xie X, Parlikad AK, Schooling JM. Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. Autom Constr. 2020;118:16.

108. Braglia M, Gabbrielli R, Frosolini M, Marrazzini L, Padellini L, Ieee, editors. Using RFID technology and Discrete-Events, Agent-Based simulation tools to build Digital-Twins of large warehouses. IEEE International Conference on RFID Technology and Applications (RFID-TA); 2019 Sep 25-27; Pisa, ITALY. NEW YORK: Ieee; 2019.

109. Cai MN, Wang SY, Wu QX, Jin YJ, Shen XL, Ieee, editors. DTCluster: A CFSFDP Improved Algorithm for RFID Trajectory Clustering Under Digital-twin Driven. 20th Asia-Pacific Network Operations and Management Symposium (APNOMS); 2019 Sep 18-20; Matsue, JAPAN. NEW YORK: Ieee; 2019.

110. Cho DJ, editor Study on method of new digital watermark generation using QR-code. 8th IEEE International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA); 2013 Oct 28-30; Univ Technol Compiegne, Compiegne, FRANCE. NEW YORK: Ieee; 2013.

111. Chow YW, Susilo W, Baek J, Kim J, editors. QR Code Watermarking for Digital Images. 20th World Conference on Information Security Applications (WISA); 2019 Aug 21-24; Korea Inst Informat Secur & Cryptol, SOUTH KOREA. CHAM: Springer International Publishing Ag; 2020.

112. Grigoropoulos N, Lalis S, editors. Simulation and Digital Twin Support for Managed Drone Applications. IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT); 2020 Sep 14-16; Electr Network. NEW YORK: Ieee; 2020.

113. Yang T, Park S, Kim SH. Collaborative Reliable Event Transport Based on Mobile-Assisted Sensing in Urban Digital Twin. Electronics. 2022;11(10):13.

114. Cainelli G, Rauchhaupt L, Ieee, editors. Introducing resilience in industrial 5G systems using a digital twin approach. 17th IEEE International Conference on Factory Communication Systems (WFCS) - Communication in Automation; 2021 Jun 09-11; Johannes Kepler Univ, Linz, AUSTRIA. NEW YORK: Ieee; 2021.

115. Ramirez R, Huang CY, Liang SH. 5G Digital Twin: A Study of Enabling Technologies. Appl Sci-Basel. 2022;12(15):18.

116. Vering C, Mehrfeld P, Nurenberg M, Coakley D, Lauster M, Muller D, editors. Unlocking Potentials of Building Energy Systems' Operational Efficiency: Application of Digital Twin Design for HVAC systems. 16th Conference of the International-Building-Performance-Simulation-Association (IBPSA); 2019 Sep 02-04; Rome, ITALY. TORONTO: Int Building Performance Simulation Assoc-Ibpsa; 2020.

117. de Oliveira AS, Asato OL, Kaneshiro PJI, Nakamoto FY, editors. Proposed Virtual Architecture by using Multi-Agent Systems. 14th IEEE International Conference on Industry Applications (INDUSCON); 2021 Aug 15-18; Univ Sao Paulo, Escola Politecnica, ELECTR NETWORK. NEW YORK: Ieee; 2021.

118. Proos DP, Carlsson N, Ieee, editors. Performance Comparison of Messaging Protocols and Serialization Formats for Digital Twins in IoV. 19th IFIP Networking Conference (Networking); 2020 Jun 22-25; Electr Network. NEW YORK: Ieee; 2020.

119. Campolo C, Genovese G, Molinaro A, Pizzimenti B, editors. Digital Twins at the Edge to Track Mobility for MaaS Applications. IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT); 2020 Sep 14-16; Electr Network. NEW YORK: Ieee; 2020.

120. Li L, Xu WJ, Liu ZH, Yao BT, Zhou ZD, Pham DT, Asme, editors. DIGITAL TWIN-BASED CONTROL APPROACH FOR INDUSTRIAL CLOUD ROBOTICS. 14th ASME International Manufacturing Science and Engineering Conference; 2019 Jun 10-14; Erie, PA. NEW YORK: Amer Soc Mechanical Engineers; 2019.

121. Mavromatis I, Piechocki RJ, Sooriyabandara M, Parekh A, Ieee, editors. DRIVE: A Digital Network Oracle for Cooperative Intelligent Transportation Systems. 25th IEEE Symposium on Computers and Communications (ISCC); 2020 Jul 07-10; Rennes, FRANCE. NEW YORK: Ieee; 2020.

122. Luo D, Guan ZL, He C, Gong YM, Yue L. Data-driven cloud simulation architecture for automated flexible production lines: application in real smart factories. Int J Prod Res. 2022;60(12):3751-73.

123. Wei CC, Lee YT, Cao KS, Lee WC, Ieee, editors. Implementation of a Data Acquisition System for Heterogeneous Machines. IEEE/SICE International Symposium on System Integration (SII); 2017 Dec 11-14; Taipei, TAIWAN. NEW YORK: Ieee; 2017.

124. Angrish A, Starly B, Lee YS, Cohen PH. A flexible data schema and system architecture for the virtualization of manufacturing machines (VMM). J Manuf Syst. 2017;45:236-47.

125. Fahim M, Sharma V, Cao TV, Canberk B, Duong TQ. Machine Learning-Based Digital Twin for Predictive Modeling in Wind Turbines. IEEE Access. 2022;10:14184-94.
126. Liu T, Tang L, Wang WL, Chen QB, Zeng XP. Digital-Twin-Assisted Task Offloading Based on Edge Collaboration in the Digital Twin Edge Network. IEEE Internet Things J. 2022;9(2):1427-44.
127. Bhatti G, Singh RR, Ieee, editors. Intelligent Fault Diagnosis Mechanism for Industrial Robot Actuators using Digital Twin Technology. 2nd IEEE International Power and Renewable Energy Conference (IPRECON); 2021 Sep 24-26; Coll Engn Karunagappally, Karunagappally, INDIA. NEW YORK: Ieee; 2021.
128. Kim TW, Oh J, Min C, Hwang SY, Kim MS, Lee JH. An Experimental Study on Condition Diagnosis for Thrust Bearings in Oscillating Water Column Type Wave Power Systems. Sensors. 2021;21(2):20.
129. Chen Z, Daly S. Deformation twin identification in magnesium through clustering and computer vision. Mater Sci Eng A-Struct Mater Prop Microstruct Process. 2018;736:61-75.
130. Araujo C, Villanueva J, Medeiros I, Almeida R, editors. Digital Twin Design for Thermal Power Plant Cooling System using Fuzzy System. 14th IEEE International Conference on Industry Applications (INDUSCON); 2021 Aug 15-18; Univ Sao Paulo, Escola Politecnica, ELECTR NETWORK. NEW YORK: Ieee; 2021.
131. Ayo-Imoru RM, Ali AA, Bokoro PN, Ieee, editors. An enhanced fault diagnosis in nuclear power plants for a digital twin framework. IEEE International Conference on Electrical, Computer, and Energy Technologies (ICECET); 2021 Dec 09-10; Cape Town, SOUTH AFRICA. NEW YORK: Ieee; 2021.
132. Tai YH, Zhang LQ, Li Q, Zhu CS, Chang V, Rodrigues J, Guizani M. Digital-Twin-Enabled IoMT System for Surgical Simulation Using rAC-GAN. IEEE Internet Things J. 2022;9(21):20918-31.
133. Xu QH, Ali S, Yue T, Ieee Comp SOC, editors. Digital Twin-based Anomaly Detection in Cyber-physical Systems. 14th IEEE Conference on Software Testing, Verification and Validation (ICST); 2021 Apr 12-16; Electr Network. LOS ALAMITOS: Ieee Computer Soc; 2021.
134. Li M, Li Z, Huang XD, Qu T. Blockchain-based digital twin sharing platform for reconfigurable socialized manufacturing resource integration. Int J Prod Econ. 2021;240:14.
135. Lee D, Lee SH, Masoud N, Krishnan MS, Li VC. Integrated digital twin and blockchain framework to support accountable information sharing in construction projects. Autom Constr. 2021;127:9.
136. Shahbaz A, Jo KH. Deep Atrous Spatial Features-Based Supervised Foreground Detection Algorithm for Industrial Surveillance Systems. IEEE Trans Ind Inform. 2021;17(7):4818-26.
137. Vo XT, Tran TD, Nguyen DL, Jo KH, editors. Dynamic Multi-Loss Weighting for Multiple People Tracking in Video Surveillance Systems. 2021 IEEE 19th International Conference on Industrial Informatics (INDIN); 2021 21-23 July 2021.
138. Hou D, Liu T, Pan YT, Hou J, editors. AI on edge device for laser chip defect detection. 9th IEEE Annual Computing and Communication Workshop and Conference (CCWC); 2019 Jan 07-09; Univ Nevada, Las Vegas, NV. NEW YORK: Ieee; 2019.
139. Tang S, He F, Huang X, Yang J. Online PCB defect detector on a new PCB defect dataset.arXiv preprint arXiv:1902.06197, 2019.
140. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollar P, Zitnick CL, editors. Microsoft COCO: Common Objects in Context. 13th European Conference on Computer Vision (ECCV); 2014 Sep 06-12; Zurich, SWITZERLAND. CHAM: Springer

International Publishing Ag; 2014.
141. Lowe DG, editor Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision; 1999 20-27 Sept. 1999.
142. Viola P, Jones M. Robust Real-Time Face Detection. Int J Comput Vis. 2004;57:137-54.
143. Dalal N, Triggs B, editors. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005 20-25 June 2005.
144. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object Detection with Discriminatively Trained Part-Based Models. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1627-45.
145. Wang Z, Zheng L, Liu Y, Li Y, Wang S, editors. Towards Real-Time Multi-Object Tracking. Computer Vision – ECCV 2020; 2020 2020//; Cham: Springer International Publishing.
146. Zhang YF, Wang CY, Wang XG, Zeng WJ, Liu WY. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. Int J Comput Vis. 2021;129(11):3069-87.
147. Kumar A, Kaur A, Kumar M. Face detection techniques: a review. Artif Intell Rev. 2019;52(2):927-48.
148. Zhou Y, Liu L, Shao L, Mellor M, editors. DAVE: A Unified Framework for Fast Vehicle Detection and Annotation. 14th European Conference on Computer Vision (ECCV); 2016 Oct 08-16; Amsterdam, NETHERLANDS. CHAM: Springer International Publishing Ag; 2016.
149. Cao JL, Pang YW, Xie J, Khan FS, Shao L. From Handcrafted to Deep Features for Pedestrian Detection: A Survey. IEEE Trans Pattern Anal Mach Intell. 2022;44(9):4913-34.
150. Vo X-T, Jo K-H. A review on anchor assignment and sampling heuristics in deep learning-based object detection. Neurocomputing. 2022;506:96-116.
151. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective Search for Object Recognition. Int J Comput Vis. 2013;104(2):154-71.
152. He KM, Zhang XY, Ren SQ, Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904-16.
153. Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137-49.
154. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S, editors. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 21-26 July 2017.
155. Redmon J, Divvala S, Girshick R, Farhadi A, editors. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 27-30 June 2016.
156. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC, editors. SSD: Single Shot MultiBox Detector. Computer Vision – ECCV 2016; 2016 2016//; Cham: Springer International Publishing.
157. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg A. DSSD : Deconvolutional Single Shot Detector. 2017.
158. Lin TY, Goyal P, Girshick R, He KM, Dollar P. Focal Loss for Dense Object

Detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318-27.

159. Borghesi A, Bartolini A, Lombardi M, Milano M, Benini L, Aaai, editors. Anomaly Detection Using Autoencoders in High Performance Computing Systems. 33rd AAAI Conference on Artificial Intelligence / 31st Innovative Applications of Artificial Intelligence Conference / 9th AAAI Symposium on Educational Advances in Artificial Intelligence; 2019 Jan 27-Feb 01; Honolulu, HI. PALO ALTO: Assoc Advancement Artificial Intelligence; 2019.

Appendix of articles

1. Digital twin of construction crane and realization of the physical to virtual connection
   Enliu Yuan, Jian Yang, Mohamed Saafi, Jianqiao Ye (Under reviews)
   (Journal of Industrial Information Integration, submitted 25th January 2023)

2. Safety monitoring of construction equipment based on multi-sensor technology
   Zi-qing Yang, Jian Yang, Enliu Yuan (Published)
   (Proceedings of the 37th ISARC, Page 677-684, 2020)

3. Literature review and bibliometric analysis of literature of Digital twin
   Enliu Yuan, Jian Yang, Mohamed Saafi, Jianqiao Ye (To be submitted)
   (Buildings, December 2023)

4. Safety monitoring of historical building using the sensor technology.
   Enliu Yuan, Jian Yang, Mohamed Saafi, Jianqiao Ye (To be submitted)
   (Buildings, December 2023)

5. Tower crane operation mode detection using skeleton detection algorithms.
   Enliu Yuan, Jian Yang, Mohamed Saafi, Jianqiao Ye (To be submitted)
   (ISARC2024, July 2024)