International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n3.01
Journal homepage: https://wepub.org/index.php/IJCSIT/index

The Integration of Generative Artificial Intelligence and Computer Vision in Industrial Robotic Arms

Chang Che[1,*], Chen Li[2], Zengyi Huang[3]

[1]Mechanical Engineering, The George Washington University, DC, USA
[2]Computer Science, The University of Texas at Dallas, Dallas, USA
[3]Applied Economics, The George Washington University, DC, USA
*Corresponding Author: cche57@gwmail.gwu.edu

ABSTRACT

Intelligent manufacturing has gradually become an important development trend in the industrial field. As an artificial intelligence technology, machine vision has been widely used in the field of automation. This paper discusses the development and application of robot arm intelligent picking system based on machine vision in the field of intelligent manufacturing. The system converts the target into the image signal through the image acquisition device, and sends it to the special image processing system for digital processing. Then, the image system performs various operations on the signal to extract the features of the target, and controls the action of the field equipment according to the discriminating results. With machine vision technology as the core, the system realizes automatic picking tasks and improves production efficiency and quality.

KEYWORDS

Machine Vision, Intelligent Manufacturing, Robotic Arms, Picking Robot Systems

1. INTRODUCTION

At present, robot technology is becoming more and more mature, and it is favored by contemporary researchers because of the integration of multi-disciplinary technologies such as computer, sensor, dynamics and bionics, and its application fields are increasingly extensive. At the same time, it is also an important sign to measure the level of industrial intelligent manufacturing in a country.

The development of robotics technology can be divided into three main stages: the first stage belongs to the era of "remote operators"; In the second stage [1], the robot needs to be controlled by the program pre-edited by professional and technical personnel to achieve its purpose, so that it can complete certain operations by itself; The third stage of robot technology has entered the era of intelligence, through the use of various external devices such as sensors, distance meters and other environmental information, and combined with intelligent technology for target recognition, semantic understanding, logical reasoning and decision-making. In the third stage, a robot is an intelligent machine that can independently operate a predetermined target action [2-3]. At present, robots are not simple substitutes for human labor, but intelligent mechanical equipment that integrates human strengths and their own advantages. In short, according to the current development trend, it is also the product of the evolution of human production activities. The robot not only has the ability to perceive and analyze the surrounding environment and respond quickly, but also the machine equipment has higher accuracy in the actual operation process and can better resist the harsher environment, and has the ability to carry out continuous labor operations for a long time. At the same

time, it is also one of the typical intelligent equipment in the field of advanced intelligent manufacturing technology.

In conclusion, with the continuous progress of Internet technology, the amount of data shows a large-scale growth, and more and more abundant data sets continue to emerge. In addition, thanks to the improvement of hardware capabilities, the computing power of computers is becoming more and more powerful. Researchers are constantly applying new models and algorithms to the field of computer vision. This has given birth to more and more rich model structure and more accurate accuracy, at the same time, the problems dealt with by computer vision are also more and more rich, including classification, detection, segmentation, scene description, image generation and style transformation, etc., and even not limited to two-dimensional pictures, including video processing technology and 3D vision, etc., the application field is also more and more extensive. At present, the mainstream computer vision tasks mainly include image classification, object detection, image segmentation, OCR, video analysis and image generation. Next we introduce each task and quickly practice it with the PaddleHub tool.

2. RELATED WORK

2.1. Computer vision

Computer Vision, also known as Machine Vision, is a discipline that lets machines learn how to "see" and is an important application field of deep learning technology, which is widely used in security, industrial quality inspection and automatic driving scenarios. Specifically, it is to let the machine to identify the object in the picture or video taken by the camera, detect the location of the object, and track the target object, so as to understand and describe the scene and story in the picture or video, in order to simulate the human brain visual system. Therefore, computer vision is also commonly referred to as machine vision, and the goal is to build artificial systems that can "sense" information from images or videos.

The development of computer vision begins with biological vision. For the origin of biological vision, the academic community has not yet formed a conclusion. Some researchers believe that the earliest biological vision was formed in jellyfish about 700 million years ago, and some researchers believe that biological vision emerged in the Cambrian period about 500 million years ago [4]. After decades of development, computer vision technology has been applied in many fields such as traffic (license plate recognition, road violation capture), security (face gate, community monitoring), finance [5] (face payment, automatic ticket recognition of counters), medical treatment (medical image diagnosis), industrial production (automatic detection of product defects), etc. [6] Affecting or changing People's Daily life and industrial production methods. In the future, with the continuous evolution of technology, more products and applications will emerge, creating greater convenience and broader opportunities for our lives.

Most computer vision tasks rely on image features (image information), and the quality of image features largely determines the performance of the vision system. Traditional methods usually use SIFT, HOG and other algorithms to extract image features, and then use SVM and other machine learning algorithms to further process these features to solve visual tasks. Pedestrian detection is to judge whether there is a pedestrian in the image or video sequence and to give accurate positioning, the earliest method is HOst. Sign extraction +SVM classifier, the detection process is as follows:

1. Use the sliding window to traverse the entire image to obtain the candidate region

2. Extract HOG features of candidate regions

3. Use SVM classifier to classify the feature map (to determine whether it is human)

4. Duplicate areas appear when sliding Windows are used. Use NMS(non-maximum) to filter the duplicate areas

Among them, Yann LeCun applied convolutional neural networks to the field of image recognition for the first time in 1998. Its main logic is to use convolutional neural networks to extract image features, predict the categories of images, and constantly adjust network parameters through training data. Finally, a set of network LeNet that can automatically extract and classify image features is formed [7]. This method was very successful in handwritten digit recognition tasks, but it did not develop well in the following time. On the one hand, the main reason is that the data set is not perfect, can only handle simple tasks, and it is easy to overfit on large-size data. On the other hand is the hardware bottleneck, when the network model is complex, the calculation speed will be particularly slow.

In 2012, Alex Krizhevsky et al. proposed AlexNet[8] and applied it to large-size image data set ImageNet, winning the champion of ImageNet Contest in 2012, which greatly promoted the development of convolutional neural networks in the field of computer vision.

2.2. Visual picking robot

In recent years, the target recognition and positioning technology of visual picking robots has been greatly developed [9-13]. Under natural conditions, the extraction of apple targets is affected by a series of factors such as illumination and occlusion. In improving the detection speed and recognition accuracy, it has been paid great attention by technicians. Since the target picture of apple is taken in the non-structured natural orchard environment, the branches and leaves are usually shielded from the apple, resulting in shape defects and uneven color distribution of the apple. In order to accurately identify and locate the target, Tien Thanh Nguyen et al. [14] innovatively proposed an apple recognition algorithm by using RGB color imaging space and three-dimensional environment shape information. The algorithm is based on color and shape features recognition and positioning, combined with RGB and three-dimensional information results, the camera can quickly detect and locate the target apple. The experimental results show that for the single apple and the target apple with occlusion under natural conditions, the recognition rate is 100% and 82% respectively, and the error of Cartesian coordinate system is less than 10mm, and the recognition and positioning process takes less than 1 second. [15] Experiments show that this algorithm has certain recognition accuracy and real-time detection ability, which can improve the ability of orchard apple picking and orchard monitoring.

Object detection algorithms have been developing rapidly in recent years. Since 2012, CNN has been widely concerned by technicians, and in 2015,Girshick et al. [16] made structural improvements on the basis of previous algorithms to give birth to Fast-RCNN. Fast-RCNN can partially reduce the number of candidate frames through the feature extraction of image information and the generation of candidate frames. In the same year, RenS et al. [17] further proposed the Faster-RCNN algorithm. This algorithm is the first real achievement of end-to-end object detection so far. Compared with the Fast-RCNN algorithm, the difference lies in the use of a candidate region generation Network (Rcgion Proposal Network) instead of the traditional selective algorithm to select the early candidate boxes. This algorithm greatly improves the speed of target detection.

In 2015, the powerful YOLO came out [18], which is an end-to-end one-stage object detection neural network. It is an algorithm that can return the target position and category information of the boundary box by passing in the image file. The whole calculation process omits the step of extracting candidate frame, which makes the detection speed increase one step. But at the same time, the rapid increase of monitoring speed brings about the lack of positioning accuracy, especially for small target objects, the recognition result is not ideal; The SSD(Single Shot MultiboxDetector) network was first announced at the ECCV2016 conference.

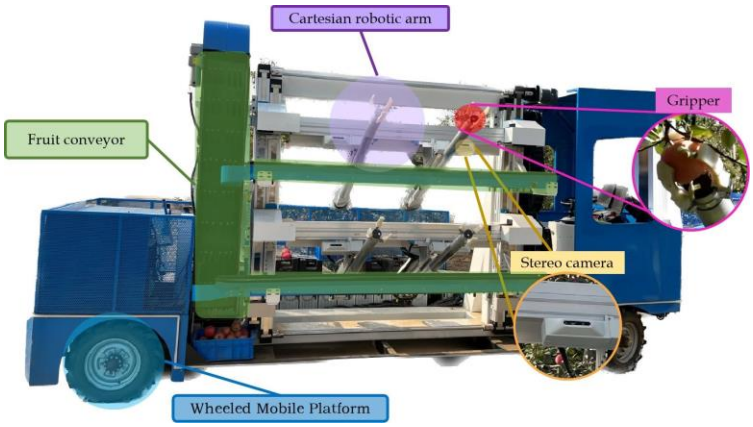

Figure 1. Robot picking field application diagram

From Figure 1 the SSD structure integrates the advantages of YOLO algorithm regression boundary box and the selection of anchor candidate box in FasterR-CNN algorithm structure, which improves the detection capability and achieves excellent results. Moreover, it is also an end-to-end target detection network. The structure of SSD algorithm is to integrate multiple layers of feature maps of different sizes and to return the results of candidate boxes of different sizes. Through the use of VOC 2007 public data for experimental verification, the test accuracy of different targets can reach 77.2% on average, and the speed in the detection process meets the performance requirements. In summary, the emergence of SSD networks has greatly promoted the development of neural networks in target detection practice and the application of various industries. Again, the structure based on SSD algorithm is better. This paper chooses to use SSD neural network, and on the basis of its structure optimization and improvement, to improve the ability of target detection.

2.3. Robot picking robotic arms

Robotics and machine vision differ in their areas of specialization. Robotics usually falls into the field of mechanical engineering and automatic control; Machine vision belongs to information engineering and electrical engineering. Through the cooperation of experts in these two fields, it is possible to give robots visual perception. It can be seen that robot vision is a highly integrated engineering technology, which detects people and objects in the environment through machine vision, calculates their position on the camera coordinate system, converts them to the robot arm coordinate system, and then drives the motor to drive the shaft joint operation goal is a seemingly simple process, but in fact contains complex computer operations.

How the robotic arm and camera (vision) fit together depends on the spatial relationship between the robotic arm and camera, also known as the hand-eye relationship.

Eye-to-hand means to hang the camera on the end axis of your arm. After the camera and visual recognition are completed, the driving arm clamps the workpiece; eye-to-hand means that the camera and the arm are fixed in two positions respectively, and the arm can move at the same time in the process of image recognition, so there is a good cycle time, but it must be used to ensure that the arm and the camera maintain a fixed relative relationship. If the relationship between the two changes, it needs to be recalibrated. As for the eye view hand structure, also known as secondary positioning, when the arm grips the workpiece, it moves into the camera field of view, compares the difference between the current position and the standard position, and then makes further postural compensation.

The integration of robotic arms and machine vision is not an easy task under the current industrial development. If the end customer does not have certain engineering capabilities, he still needs the help of a system integrator with specialized knowledge. For arms, the system integrator must first consider the length and load of the arm when selecting the right arm. Arm length ensures effective

working range. With regard to the load, work pieces such as end-effectors and fixtures need to be calculated. Whether it can meet the rated load range of arm operation.

On the other hand, there are many options for the integration of vision solutions. For situations where multiple cameras are required and the computing burden is high, a vision controller is often used. The hardware is essentially an industrial computer, typically supporting two to four industrial cameras, and has built-in image recognition software that allows users to write the visual recognition problems they want to solve. Another product is the smart camera, which itself is an embedded computing platform with a CCD/CMOS sensor. Users can choose the right lens according to their working field of view. The platform also contains vision processing software, but the computational performance is not as good as vision controllers, which are typically used for code reading and positioning. In addition, some system integrators, in order to save costs or increase flexibility, integrate commercial or free visualization libraries and develop specialized software.

In addition to considering the right arm model and vision solution, the most important metrics when evaluating the feasibility of an automation case are accuracy and cycle time. Sufficient precision guarantees the correctness of each process, and the expected production cycle can be used to assess whether production capacity increases with the introduction of automation, and to calculate return on investment (ROI). For the above accuracy part, if the object is positioned through vision, the factors affecting the overall accuracy include camera resolution, positioning algorithm, hand-eye relationship correction error, camera lens correction error, arm repetition accuracy, absolute accuracy, etc., which needs to rely on experienced machine vision technicians to effectively evaluate.

3. METHODOLOGY

In conclusion, with the modernization of agricultural production and the increasing demand for intelligence, machine vision technology is increasingly widely used in the agricultural field. Among them, the robot arm picking apples based on machine vision has become a research area of concern. The traditional apple picking process usually relies on manual labor, but due to the rising labor cost, labor shortage and other problems, the traditional picking method has been difficult to meet the needs of modern agricultural production. Therefore, the development of an intelligent picking system based on machine vision has become one of the urgent problems to be solved[26]. This research aims to use advanced machine vision technology combined with robotic arm control algorithm to realize automatic positioning and picking of apples, so as to improve picking efficiency, reduce production costs, and reduce manual labor pressure. This paper will introduce the research method and implementation process of this system in detail, as well as its application prospect in the field of intelligent agriculture.

3.1. Picking robot vision system establishment

There are many ways to define the imaging geometry model of camera, among which the pinhole model is one of the classic definition methods. The principle of image acquisition by camera is based on the ideal model of optical imaging, in fact the optical imaging model is a simplification of nonlinear lens imaging.

In a camera lens, it is assumed that parallel light is focused through the lens to a point on the imaging plane, but in fact these rays parallel to the main optical axis originate from the corresponding point on the space object. Therefore, in the pinhole camera model, the space object is focused to any position on the unique p(xy) of the imaging plane. After P(X,yZ) passes through the lens equivalent to a convex lens, the relationship between the object point and the image point on the imaging plane is the perspective projection, which also belongs to the central projection.

The basic camera imaging model is shown in Figure 2, where the vertical distance from the camera lens to the imaging plane is represented by the focal length f.

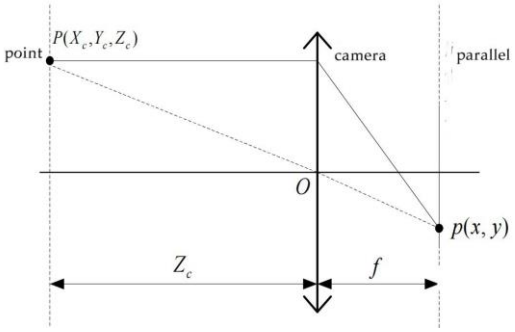

Figure 2. Basic model of pinhole camera imaging

According to the imaging model of pinhole camera as shown in Figure 2, the relationship is as follows:

$$\begin{cases} x = \dfrac{fX_c}{Z_c} \\ y = \dfrac{fY_c}{Z_c} \end{cases} \tag{1}$$

Where,(X,Y,Z.) represents the coordinates of object point P in the three-dimensional camera coordinate system with the optical center as the coordinate origin; (x,y represent the coordinates of the object point P in the two-dimensional image coordinate system with the center of the imaging plane as the coordinate origin; f is the focal length of the camera.

The accuracy of camera calibration mainly depends on the algorithm and hardware and other factors. Since the hardware parameters have been determined, it is necessary to make a judgment criterion for the camera calibration accuracy based on the algorithm process, including the smoothness of the checkerboard and the number of images collected. Because the checkerboard substrate is made of smooth aluminum with a processing accuracy of ±0.01mm, the impact on the calibration accuracy can be negligible. Therefore, controlling the number of acquired images is a major factor in evaluating the calibration accuracy of subsequent cameras.

According to the basic principle of camera calibration algorithm, each checkerboard image can obtain two relations including calibration parameters, so at least three checkerboard images with different directions or poses need to be collected to solve the internal and external parameter matrix of the camera. Too many images collected at the same time will greatly increase the calculation amount of the algorithm, which is not conducive to the rapid progress of the subsequent algorithm. Moreover, the presence of lens distortion does not reduce the calibration accuracy to zero. Accurate calibration parameters are the key to ensure the effective conduct of subsequent identification and positioning experiments, so it is necessary to use the control variable method to discuss the influence of the number of acquired images on the calibration accuracy.

3.2. Robot arm visual inspection experiment

The manipulator in the robot hardware is the key part to complete the apple picking action. The motion mode design of the manipulator is based on manual picking action data and path planning. The research team used the NOKOV Measurement motion capture system to collect data on the upper arm movements of pickers, surrounded by eight digital cameras with a resolution of 2048×1088 and a 3D accuracy of ±0.15mm. Data is collected by tracking marked points and transmitted to the host for real-time processing to calculate the coordinates, speed and acceleration of moving objects in space.

Small balls with fluorescent surfaces were labeled on the picker's shoulders, elbows, wrists, and fingertips (Figure 3). Since the structure of the manipulator is different from that of a human arm, only fingertip trajectory data was used for analysis.

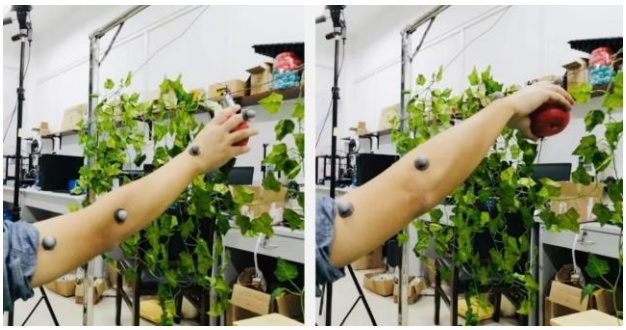

Figure 3. Simulated visual picking motion capture

The amplitude of the pick motion is small and the speed is fast, so the data acquisition frequency is set to 100 frames per second. At the same time, in order to maintain high tracking accuracy, the system ensures that at least three cameras track each marked point. The fingertip mark at the beginning of the movement is used as the initial position, and the height change is recorded for every 5mm horizontal displacement. The data collection experiment was carried out 10 times. After the test is completed, the mean value is fitted by polynomial.

3.3. The experimental results of robotic arm are discussed

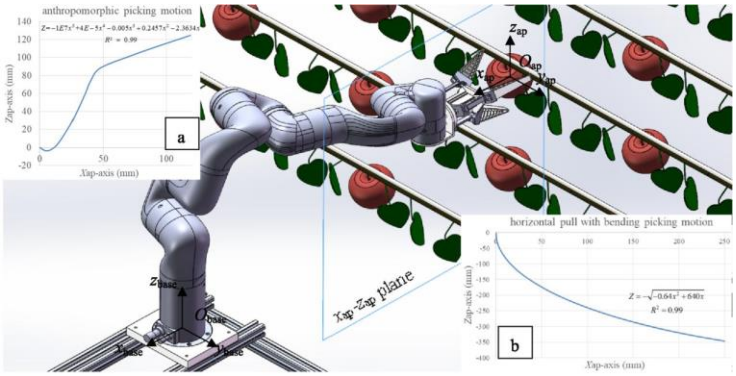

Figure 4. The statistical summary of the on-site evaluation of Apple harvest

The statistical summary of the on-site evaluation of Apple harvest is shown in Figure 4. Specifically, the success rate of apple picking using the anthropomorphic action was 80.17%, 2.76% lower than when using the "pull and bend horizontally" action (82.93%). In addition, in terms of time, the cycle time of the picking process using the "horizontal pull bending" movement is 12.53±0.53 seconds, which is 4.64 seconds less than the average time using the humanoid picking movement (17.17±0.36 seconds). The picking action itself took 1.14 seconds and 18.23% of the total cycle time, respectively.

The results show that the robot arm apple picking system based on machine vision technology has feasibility and significant advantages in modern agricultural production. By combining advanced motion capture systems and real-time processing technologies with robotic arm control algorithms, picking efficiency can be significantly improved and production costs reduced. The experimental results showed that compared with the human-like movement, the success rate of apple picking was slightly decreased (80.17% vs. 82.93%), but the efficiency was significantly improved, and the average cycle time was reduced by 4.64 seconds. In addition, the average picking cycle time using the "horizontal pulling" movement was 12.53 seconds, which was significantly shorter than the 17.17 seconds using the human-like picking movement, showing the potential to achieve significant efficiency improvements in agricultural production tasks. In addition, the actual picking action itself accounted for only about 18.23% of the total cycle time, suggesting that overall efficiency could be further improved by optimizing other aspects of the picking process, such as motion planning and trajectory optimization. Taken together, these findings highlight the promise of intelligent agricultural

systems in addressing the challenges of modern agricultural production, and provide an important reference for the automation and intelligent development of agricultural production.

4. CONCLUSION

In the development process of today's agricultural intelligence, the application of agricultural robots is more and more extensive, and artificial intelligence vision technology has been deeply involved in various machines and equipment. However, in terms of the current development level of agricultural intelligence in China, the full realization of agricultural mechanization still needs the unremitting efforts of researchers. Aiming at this situation, this paper studies the target recognition and picking position of apple fruit by using machine vision technology. Compared with the traditional algorithm, the detection results of apple growth characteristics are better in small targets and complex environments. However, in the process of Apple target detection, there are still many problems to be solved. First, due to the inevitable jitter caused by the wind direction during the picking process, the picking rate is reduced during the camera recognition process. Therefore, the current target recognition algorithm does not consider dynamic factors and has certain limitations. Secondly, in the actual environment, apple maturity pre-inspection and apple disease spot detection are also factors that we need to consider. In the follow-up research process, a series of pre-inspection operations can be added on the basis of the object recognition algorithm in this paper to reduce the workload of apple fruit quality classification after robot picking, so as to reduce a lot of unnecessary work in the later picking stage.

In summary, through the research of this paper, the research of apple target recognition and picking location based on machine vision technology has made some progress, but there are still some challenges and problems. Future research directions could include improving the target recognition algorithm to account for dynamic factors, increasing pre-inspection operations to improve the efficiency of apple picking consequence quality classification, and further exploring the application of other agricultural intelligent technologies to achieve the full realization of agricultural mechanization. These efforts will help promote the development of agricultural intelligence, improve agricultural production efficiency, reduce farmers' labor pressure, and promote sustainable agricultural development.

REFERENCES

[1] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[2] Che, Chang, et al. "Deep learning for precise robot position prediction in logistics." Journal of Theory and Practice of Engineering Science 3.10 (2023): 36-41.

[3] Yu, Liqiang, et al. "Stochastic Analysis of Touch-Tone Frequency Recognition in Two-Way Radio Systems for Dialed Telephone Number Identification." arXiv preprint arXiv:2403.15418 (2024).

[4] Hu, Hao, et al. "Casting product image data for quality inspection with xception and data augmentation." Journal of Theory and Practice of Engineering Science 3.10 (2023): 42-46.

[5] Yu, Liqiang, et al. "Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method." arXiv preprint arXiv:2401.06782 (2024).

[6] Che, Chang, et al. "Advancing Cancer Document Classification with R andom Forest." Academic Journal of Science and Technology 8.1 (2023): 278-280.

[7] Huang, Jiaxin, et al. "Enhancing Essay Scoring with Adversarial Weights Perturbation and Metric-specific AttentionPooling." arXiv preprint arXiv:2401.05433 (2024).

[8] Lin, Qunwei, et al. "A Comprehensive Study on Early Alzheimer's Disease Detection through Advanced Machine Learning Techniques on MRI Data." Academic Journal of Science and Technology 8.1 (2023): 281-285.

[9] Li, Chen, et al. "Enhancing Multi-Hop Knowledge Graph Reasoning through Reward Shaping Techniques." arXiv preprint arXiv:2403.05801 (2024).

[10] Song, T., Li, X., Wang, B., & Han, L. (2024). Research on Intelligent Application Design Based on Artificial Intelligence and Adaptive Interface.

[11] K. Xu, X. Wang, Z. Hu and Z. Zhang, "3D Face Recognition Based on Twin Neural Network Combining Deep Map and Texture," 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, 2019, pp. 1665-1668, doi: 10.1109/ICCT46805.2019.8947113.

[12] Shi, Peng, Yulin Cui, Kangming Xu, Mingmei Zhang, and Lianhong Ding. 2019. "Data Consistency Theory and Case Study for Scientific Big Data" Information 10, no. 4: 137. https://doi.org/10.3390/info10040137.

[13] Huang, Zengyi, et al. "Research on Generative Artificial Intelligence for Virtual Financial Robo-Advisor." Academic Journal of Science and Technology 10.1 (2024): 74-80.

[14] Che, C., Lin, Q., Zhao, X., Huang, J., & Yu, L. (2023, September). Enhancing Multimodal Understanding with CLIP-Based Image-to-
Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (pp. 414-418).

[15] Xu, Z., Gong, Y., Zhou, Y., Bao, Q., & Qian, W. (2024). Enhancing Kubernetes Automated Scheduling with Deep Learning and Reinforcement Techniques for Large-
Scale Cloud Computing Optimization. arXiv preprint arXiv:2403.07905.

[16] Huang, Zengyi, et al. "Application of Machine Learning-Based K-means Clustering for Financial Fraud Detection." Academic Journal of Science and Technology 10.1 (2024): 33-39.

[17] Xu, X., Xu, Z., Ling, Z., Jin, Z., & Du, S. (2024). Comprehensive Implementation of TextCNN for Enhanced Collaboration between Natural Language Processing and System Recommendation. arXiv preprint arXiv:2403.09718.

[18] Song, B., Xu, Y., & Wu, Y. (2024). ViTCN: Vision Transformer Contrastive Network For Reasoning. arXiv preprint arXiv:2403.09962.