Deep Learning and Image Recognition

Chaoyang Li[1]*, Xiaohan Li[2], Manni Chen[1], Xinyao Sun[1]
[1]School of Computer Science and Engineering, Huizhou University, Guangdong 516007, China
[2]Department of News, Northeastern University, Liaoning 110006, China
*Corresponding author: Chaoyangli@hzu.edu.cn

2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT) | 979-8-3503-9905-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEICT57916.2023.10245041

***Abstract*—Over the next decade, we can expect artificial intelligence (AI) to have a profound impact on our society. As AI technologies continue to advance and become more integrated into various industries, we will witness significant changes that are shaping the human being's future for the better. This paper reviews the major deep learning algorithms that have accomplished latest achievements on various fields such as manufacturing robots, smart assistants, automated financial investing, social media monitoring, marketing chat-bots and self-driving car. These algorithms are trained on large datasets of labeled images, where each image is associated with a set of labels or categories. This paper provides an overview of popular techniques for image recognition. It discusses the convolutional neural network (CNN), which is made up of multiple convolutional layers followed by pooling layers, dropout layers, batch normalization lays and fully connected layers. The paper addresses other techniques used in deep learning for image recognition include transfer learning, where a pre-trained model is refined on new datasets, and generative adversarial networks (GANs), which can generate realistic images by training two neural networks against each other. This paper highlights that deep learning has revolutionized the field of image recognition and has opened up new possibilities for many applications.**

Keywords—deep learning, image recognition, convolutional neural network, artificial intelligence, natural language processing, object detection.

I. INTRODUCTION

Artificial intelligence (AI) is indeed a broader field that encompasses various techniques, methods, and approaches used to develop intelligent machines [1-2]. AI involves the creation of algorithms and computer programs that can perform tasks typically requiring human intelligence, such as learning, problem-solving, perception, reasoning, and decision-making. Some common techniques in AI include machine learning, natural language processing (NLP), computer vision, robotics and expert systems [3]. These are just a few examples of the many techniques used in artificial intelligence. The field is constantly evolving, with new breakthroughs and innovations being made all the time. As AI continues to advance, it has the potential to transform numerous industries and improve our daily lives in countless ways.

Deep learning is a machine learning methodology that exploits computers to mimic the human brain's functionalities or activities, usually through training neural network models with training datasets so that complex patterns could be recognized and predictions or decisions based on that data can be performed [4]. Deep learning can achieve excellent performance for classification tasks, therefore it is particularly suited for occasions that match input data to a learned type. Advanced deep learning algorithms have revolutionized image recognition, object detection, generative adversarial network, speech recognition software, Natural Language Processing (NLP), and many other applications [5-8].

Image recognition has many applications in various areas as follows:

1. Face Recognition, a technology that allows computers to identify and verify individuals by analyzing their facial features [9-10]. It involves capturing an image of a person's face and using algorithms to match it with a database of known faces. Face recognition has many practical applications, such as security systems, social media platforms, and mobile devices. However, there are also concerns about privacy and potential misuse of this technology.

2. Video surveillance analysis, which refers to the process of examining and interpreting video content for extracting information, identify patterns, and make decisions [11]. This field combines elements of computer science, image processing, and cognitive psychology, among others, to study video data. It's important to note that video surveillance analysis raises ethical concerns regarding privacy and surveillance. Proper regulations and oversight should be in place to guarantee that these technologies are exploited in a responsible and transparent fashion.

3. Industrial defect detection, which refers to the process of identifying and analyzing defects in a product or process in industrial settings [12-13]. This can include physical, chemical, or mechanical defects that may affect the performance, safety, or reliability of the product. Industrial defect detection is an important process that helps ensure the quality and reliability of products and processes in various industries.

4. 3D image vision, which refers to the technology and techniques used to create and analyze 3D images from visual data [14]. It involves the use of cameras, sensors, and software to capture and process depth information about objects or environments in real-time. 3D image vision has a diversity of applications such as computer-aided design (CAD) [15], virtual reality (VR) [16], healthcare [17], and construction [18].

5. Medically image diagnosis, which refers to the process of using many medical imaging techniques, such as PET scans, CT scans, Fluoroscopy, MRI, Mammogram, and ultrasound, to visualize the inside of the body's organs and tissues [19]. These images are then analyzed by medical professionals to help diagnose a variety of medical conditions, diseases, and abnormalities. Medically image diagnosis plays a critical role in modern medicine, allowing doctors to identify and treat illnesses more accurately and effectively.

6. Optical character recognition (OCR), which is a technique of text recognition or text extraction that transforms printed or handwritten text from images such as product labels, advertisement material, posters into machine-understandable text [20]. It uses image processing, pattern recognition [21], and artificial intelligence algorithms to identify and extract the text content. OCR technology can significantly reduce efforts for manual data entry when we extract text from documents

like articles, tables, reports, invoices, bank check and vehicle license plate. As smartphones and tablets become more prevalent, OCR technology is also becoming increasingly important in people's daily lives.

There are some techniques for image recognition as follows:

1. Convolutional Neural Networks (CNNs) [22]. CNNs are a type of deep learning algorithm that can train neural networks for classification and computer vision tasks. They are comprised of multiple node layers such as an input layer, an output layer, and one or multiple hidden layers including convolutional layers, pooling layers, and fully connected layers. CNNs are particularly effective for image recognition tasks because they can learn hierarchical representations of the input data.

2. Feature Extraction. It is employed to extract main signal characteristics from other unwanted or additional information in order to accomplish reliable classification. Feature extraction is the most important part in the process of image recognition since the classification performance heavily depends on the feature selection. Common feature extraction techniques include SIFT (Scale-Invariant Feature Transform) [23], SURF (Speeded-Up Robust Features) [24], and HOG (Histogram of Oriented Gradients) [25].

3. Machine Learning. Machine learning algorithms, such as linear regression, random forests, logistic regression, decision trees, Native Bayes, K-nearest neighbor (KNN), K-means and support vector machines (SVMs) [26], can be used for image recognition tasks. These algorithms are trained on a labeled dataset of images and their corresponding labels.

4. Deep Learning, a machine learning technique that teaches computer to learn by examples. Deep learning networks can contain as many as 150 hidden layers whereas traditional neural networks only contain 2-3 hidden layers. Deep learning models, such as ResNet [27] and Inception, have been highly successful in image recognition tasks due to their ability to learn features directly from the data without the requirement for manual feature extraction.

5. Transfer Learning, a technique where a trained model on one task is applied to a related task. For example, a trained CNN is fine-tuned on a new dataset for image recognition tasks. Reusing or transferring knowledge can be useful when there is not enough data usable for training a new model from scratch.

6. Data Augmentation, a technique of generating additional training data by adding slightly modified copies of the original dataset. This can help to reduce overfitting and improve the accuracy of the model by exposing it to more variations of the input data. These are just a few of the many techniques used for image recognition.This paper focuses on the fundamentals and principles for deep learning in image recognition area.

II. Deep Learning Platform

Deep learning platforms are software tools and libraries helping to implement functionalities needed to build, train, and deploy deep learning models. These platforms typically provide many features such as data preprocessing, model training, deployment, and monitoring.

Some popular deep learning platforms include Tensor Flow, PyTorch, Keras. These platforms are designed to be user-friendly and accessible to developers of all levels, from beginners to experts.

A. TensorFlow

TensorFlow is an end-to-end open-source platform developed by the Google Brain team for internal Google use in research and production that offers numerous tools for building and deploying machine learning models. It supports multiple programming languages such as Python, C++, and Java. Tensor Flow has a comprehensive, flexible ecosystem of tools, libraries, and community resources for machine learning and artificial intelligence, but it does come with its own set of challenges and limitations. For example, Tensor Flow can be complex to learn, especially for beginners who are not familiar with programming or machine learning concepts. Moreover, Tensor Flow releases new versions frequently, which can make it hard for users to follow the latest features and compatibility issues.

B. Keras

Keras is an open-source software library that offers Python interface for building and training deep learning models, it is a powerful and easy-to-use free platform for developing and evaluating deep learning models for beginners. One drawback is that Keras relies heavily on external libraries such as Tensor Flow, which can sometimes cause compatibility issues. Additionally, it does not have as many built-in features or functionalities compared to other deep learning frameworks such as TensorFlow or PyTorch.

C. PyTorch

PyTorch is another open-source platform developed by Facebook that is particularly well-suited for research and development. It offers a flexible and dynamic environment that allows developers to experiment with new ideas and architectures quickly. PyTorch uses GPU to accelerate the Tensor computation, similar to Numpy. It allows for dynamic computation graphs, which means that the model can be changed during runtime without recompiling the entire codebase. This paper uses PyTorch as deep learning platform.

III. Neural Network

A neural network is also called an artificial neural network which uses interconnected nodes or neurons in a layered structure to simulate a human brain. Neural networks can be trained to perform various tasks where conventional algorithms fail to succeed, such as speech recognition, language translation, and decision-making. As shown in Fig. 1 [28], the neural network consists of an input layer, an output layer and multiple hidden layers. The solid lines stand for weights which connect the neurons. A neural network's algorithm is employed to obtain the ideal weights so that the prediction of the entire network is the best.

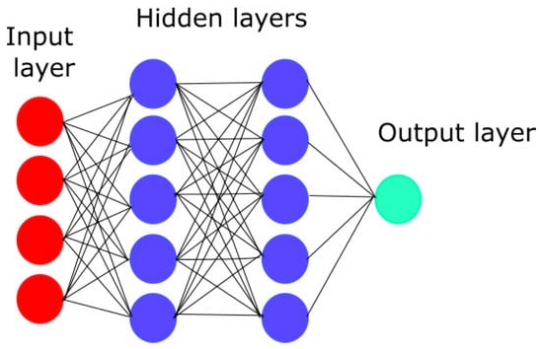

Fig. 1. Artificial neural network model

A. Forward Propagation

Forward propagation is a fundamental concept in machine learning and computer vision. It is one of the core process during learning phase and a critical component of some deep learning algorithms such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). It refers to the process of passing information from one layer of a neural network to another layer, with the goal of generating an output or prediction.

The feedforward neural network was the first and simplest ANN human invented where data moves in only one direction that is from input layer through possible hidden lays, then to output layer, without cycles or loops. The information is processed by each layer of neurons, with each layer transforming the input data into a higher-level representation. The output layer produces a prediction or classification based on the activations of its neurons.

The forward propagation algorithm involves computing the activation of each neuron in the network, using the weights and biases of the connections between them. These activation functions are then used to calculate the activation of the next layer, until the final outcome is produced.

B. Backpropagation

Backpropagation (also known as gradient descent) is an algorithm exploited to train neural networks. It is a supervised learning technique that employs the gradient of the error function with respect to the network's parameters to update those parameters so that the error is minimized.

During backpropagation, the difference (the error) between the predicted output of the network and the actual output is calculated. This quantity is then propagated backwards through the network, using the weights and biases of each layer. The parameters are updated based on the gradient of the error, which indicates how much each parameter needs to be adjusted in order to reduce the error.

The process of backpropagation can be visualized as follows:

1. An input sample is fed into the network.
2. The network processes this input and produces an output.
3. The difference between the output from the output layer and the true output is calculated as the error.
4. The error is propagated backwards through the network, using the weights and biases of each layer.
5. The weights and biases are updated based on the gradient of the error, which indicates how much each parameter needs to be adjusted in order to reduce the error.
6. The process repeats for several iterations, until the error is minimized.

IV. Convolutional Neural Network

A convolutional neural network (CNN) is a kind of artificial neural network which is particularly well-suited for image recognition tasks. It was first introduced in 2012 by Yann LeCun, a researcher at Facebook AI Research, and has since become one of the most widely used deep learning models for computer vision applications.

CNNs use a technique called convolution to extract features from images as shown in Fig. 2 [28]. Convolution involves sliding a small kernel or filter over an image, which applies a mathematical operation to each pixel in the image based on its location within the kernel. This process can be thought of as a way to determine patterns or features within the image that are relevant to the task at hand.

After applying the convolution operation, the pooling layers is employed to reduce the dimensions of the feature maps, allowing the neural network to focus on important areas. After pooling layer, the data is then fed into a flattening layer. Flattening layer is an intermediate layer in the CNN that flattens the input tensor to a one-dimensional vector. This is done by applying a non-linear activation function (such as ReLU) to each element of the tensor and then summing them up. The resulting feature map is then fed into a fully connected layer, which uses a mathematical function to classify the image based on the features it has learned. CNNs are often trained using large datasets of labeled images, allowing them to learn to recognize complex patterns and features in images with high accuracy. Some common applications of CNNs include image classification, object detection, and facial recognition. They have also been exploited in fields like medical imaging, finance, and natural language processing.

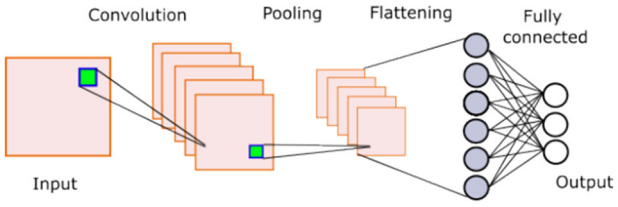

Fig. 2. CNN model

V. Object Detection

Object detection is used to accomplish a computer vision task by identifying and localizing objects within an image or video. It is a subfield of computer vision that aims to automatically detect and classify objects of interest in an image or video, such as people, cars, animals, or other objects. Object detection typically conducts feature extraction firstly by extracting relevant features from the input image or video using a convolutional neural network (CNN) [29]. The features are typically regions of interest (ROIs) that correspond to the objects of interest as shown in Fig. 3 [30]. Once the features have been extracted, they are exploited to classify the objects in the image or video. This can be done using a classifier such as a support vector machine (SVM) or a deep learning model like ResNet. Next localization is performed by determining the location of the detected objects within the image or video. This can be done using techniques such as bounding boxes or polygons.

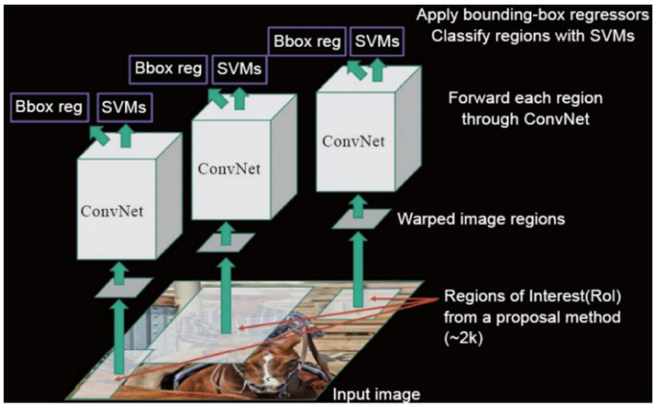

Fig. 3. Object detection process

Object detection has various practical applications, including self-driving cars, surveillance systems, medical imaging, and robotics. It is a challenging task due to the variability of object appearance and the need for accurate localization. However, with the developments in deep learning and computer vision, object detection has become increasingly accurate and reliable in recent years.

VI. Segmentation

A. Semantic Segmentation

Semantic segmentation is a computer vision technique that involves dividing an image into different regions or pixels, each of which stands for a different object or region of interest in the image as shown in Fig. 4 [31]. The aim of semantic segmentation is to distribute a label or class to each pixel in the image, indicating its corresponding object or region. This allows for more detailed and accurate analysis of the content of the image, as well as the ability to perform tasks such as object detection, tracking, and classification.

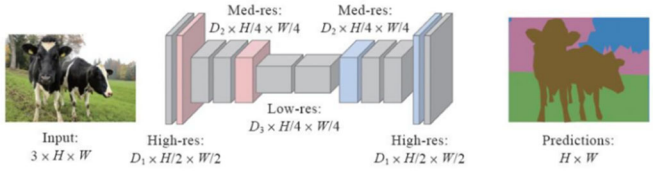

Fig. 4. Semantic segmentation and CNN architecture

In Fig. 4, subsampling and upsampling are combined to implement semantic segmentation. In the process, subsampling is accomplished through pooling and adjustment of the stride of convolution. Upsampling are conducted by Unpooling and Deconvolution, which is virtually the inverse process of subsampling.

B. Instance Segmentation

Instance segmentation is used to identify and segment individual objects within an image or video. It focuses on the process of dividing an image into many regions, each representing a distinct object or instance. The goal of instance segmentation is to produce a binary image where each region represents a unique object or instance. To achieve this goal, image filtering, edge detection, and feature extraction are conducted to prepare the image for segmentation. Then several algorithms are utilized for instance segmentation by identifying patterns in the image and separating the objects into separate regions. Finally, post-processing techniques can be applied to refine the results and ensure that each segment represents a valid object or instance. In paper [32], MaskR-CNN is used to implement instance segmentation as shown in Fig. 5.

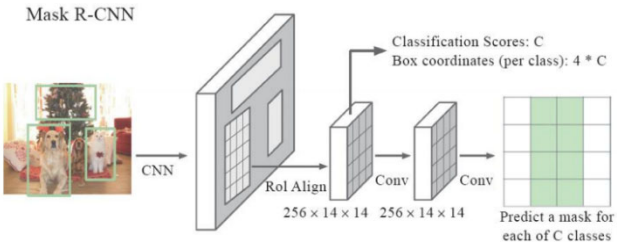

Fig. 5. The network architecture in Mask R-CNN

In conclusion, instance segmentation is a powerful tool for analyzing and understanding visual data, and has numerous real-world applications in areas such as robotics, and healthcare, autonomous driving, medical imaging, and surveillance.

VII. Generative Adversarial Networks

Machine learning can be classified as supervised learning, unsupervised learning, reinforcement learning where a large amount of data is shown to a machine. The machine then learn, make decision, identify patterns or perform classification tasks. We call a machine learning as supervised learning since the machine is "supervised" while it is working. We feed the machine with information to help it learn and outcome will be the labeled data. On the other hand, unsupervised learning accomplishes training based on unlabeled data. Unsupervised learning is employed to find unknown patterns or structure in the raw data without any predefined labels. Common unsupervised learning algorithms include auto encoders, and generative adversarial networks (GANs). Reinforcement learning exploits an agent interacting with an environment in order to maximize a reward signal. The agent learns by trial and error, taking actions that lead to higher rewards and adjusting its behavior accordingly. Common reinforcement learning algorithms include Q-learning, SARSA (SARSA is an acronym for State Action Response Score), Deep Q-Networks (DQN), and actor-critic methods.

A. Generative Model

A generative model can produce new data based on the feathers learned from the training data. In other words, it can be employed to create new text or data that is similar to the original data. Generative models are often exploited to accomplish tasks such as text production, speech synthesis, and image synthesis. Some popular generative models include Markov chains, recurrent neural networks (RNNs), and variation auto encoders (VAEs).

B. Discriminative Model

A discriminative model can classify new data into pre-defined types. It is trained to recognize patterns in the data and use those patterns to make predictions about new data. Based on conditional probability, it can be exploited for classification or regression.

C. Generative Adversarial Networks

While generative models can create new data based on learned patterns, discriminative models can classify existing data into pre-defined categories. The generative network uses a random noise vector as input and produces new data samples that resemble the training data. The goal is to

produce images that can be easily distinguished from the real images by the discriminator network. The discriminator network uses an image as input and produces a probability distribution over the possible classes that the image belongs to. It then uses this distribution to determine whether the image is real or fake.

During training, the generator network tries to fool the discriminator network by producing images that are difficult for it to distinguish from the real images. The discriminator network also tries to improve its performance by becoming more accurate at detecting fake images. This process is repeated until both networks converge to a point where they can accurately classify images as real or fake as shown in Fig. 6 [30].

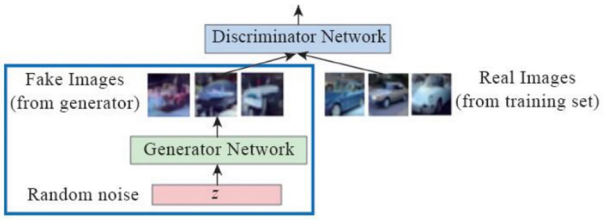

Fig. 6. The training architecture for GAN

One of the key advantages of GANs is that they can generate high-quality images that are difficult to distinguish from the real images. However, they also have some limitations, such as the risk of generating biased or inappropriate images if not carefully designed and monitored during training.

VIII. Conclusion

In this paper, we review the fundamentals of deep learning and image recognition. Some applications of computer vision are explained and multiple deep learning platforms are discussed. Basic algorithms for neural networks such as forward propagation and backpropagation are illustrated. We also analyze convolutional neural networks and address several image recognition techniques such as object detection, semantic segmentation, instance segmentation, Generative model, Discriminative model and Generative adversarial networks.

References

[1] C. Alonso, A. Berg, S. Kothari, C. Papageorgiou,, and S. Rehman, "Will the AI revolution cause a great divergence?" in Journal of Monetary Economics, 127, 18-37.

[2] F. Olan, S. Liu, J. Suklan, U. Jayawickrama, and E. O. Arakpogun, "The role of artificial intelligence networks in sustainable supply chain finance for food and drink industry," in International Journal of Production Research, 60(14), 4418-4433.

[3] M. Madani, H. Motameni, and H. Mohamadi, "Fake news detection using deep learning integrating feature extraction, natural language processing, and statistical descriptors," in Security and Privacy, 5(6), e264.

[4] Q. Ji, Y. Jiang, L. Qu, Q. Yang, and H. Zhang, "An image diagnosis algorithm for keratitis based on deep learning," in Neural Processing Letters, 54(3), 2007-2024.

[5] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," in Journal of big Data, 8, 1-74.

[6] T. Liu, J. Zhang, G. Gao, J. Yang, and A. Marino, "CFAR ship detection in polarimetric synthetic aperture radar images based on whitening filter," in IEEE Transactions on Geoscience and Remote Sensing, 58(1), 58-81.

[7] X. Qian, , X. Cheng, G. Cheng, X. Yao, and L. Jiang, "Two-stream encoder GAN with progressive training for co-saliency detection," in IEEE Signal Processing Letters, 28, 180-184.

[8] N. Yu, H. Ren, T. Deng, and X. Fan, "A lightweight radar ship detection framework with hybrid attentions," in Remote Sensing, 15(11), 2743.

[9] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," in IEEE Transactions on Affective Computing, 13(4), 2132-2143.

[10] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever,, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," Advances in neural information processing systems, 29.

[11] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," IEEE Transactions on Industrial Electronics, 66(12), 9692-9702.

[12] D. Bacioiu, G.Melton, M.Papaelias, and R.Shaw, "Automated defect classification of Aluminium 5083 TIG welding using HDR camera and neural networks," Journal of Manufacturing Processes, 45, 603-613.

[13] S. Mei, H.Yang, and Z.Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," IEEE Transactions on Instrumentation and Measurement, 67(6), 1266-1277.

[14] I. Sa, C. Lehnert, A. English, C. McCool, F. Dayoub, B. Upcroft, and T. Perez, "Peduncle detection of sweet pepper for autonomous crop harvesting—combined color and 3-D information," IEEE Robotics and Automation Letters, 2(2), 765-772.

[15] N. Cobetto, C. E. Aubin, S. Parent, S. Barchi, I. Turgeon, and H. Labelle, "Effectiveness of braces designed using computer-aided design and manufacturing (CAD/CAM) and finite element simulation compared to CAD/CAM only for the conservative treatment of adolescent idiopathic scoliosis: a prospective randomized controlled trial," European spine journal, 25, 3056-3064.

[16] S. Zhang, J. Liu, T. K. Rodrigues, and N. Kato, "Deep learning techniques for advancing 6G communications in the physical layer," IEEE Wireless Communications, 28(5), 141-147.

[17] T. Rose, C. S. Nam, and K. B. Chen, "Immersion of virtual reality for rehabilitation-Review," Applied ergonomics, 69, 153-161.

[18] T. Xia, J. Yang, and L. Chen, "Automated semantic segmentation of bridge point cloud based on local descriptor and machine learning," Automation in Construction, 133, 103992.

[19] A. M. Obaid, A. Turki, H. Bellaaj, and M. Ksontini, "Detection of Biliary Artesia using Sonographic Gallbladder Images with the help of Deep Learning approaches," In 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT) (Vol. 1, pp. 705-711). IEEE.

[20] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of post-OCR processing approaches," ACM Computing Surveys (CSUR), 54(6), 1-37.

[21] A. B. Ajiboye, and R. F. Weir, "A heuristic fuzzy logic approach to EMG pattern recognition for multifunctional prosthesis control," IEEE transactions on neural systems and rehabilitation engineering, 13(3), 280-291.

[22] F. S. Botros, A. Phinyomark, and E. J. Scheme, "Electromyography-based gesture recognition: Is it time to change focus from the forearm to the wrist?" IEEE Transactions on Industrial Informatics, 18(1), 174-184.

[23] P. Prasanna, K. J. Dana, N. Gucunski, B. B. Basily, H. M. La, R. S. Lim, and H. Parvardeh, "Automated crack detection on concrete bridges," IEEE Transactions on automation science and engineering, 13(2), 591-599.

[24] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," IEEE Transactions on Robotics, 37(6), 1874-1890.

[25] Q. Li, G. Qu, and Z. Li, "Matching between SAR images and optical images based on HOG descriptor."

[26] S. Ismail, H. Reza, K. Salameh, H. Kashani Zadeh, and F. Vasefi, "Toward an intelligent blockchain ioT-enabled fish supply chain: A review and conceptual framework," Sensors, 23(11), 5136.

[27] S. J. Pan, and Q. Yang, "A survey on transfer learning," IEEE transactions on knowledge and data engineering. 22 (10), 1345.

[28] D. Lopez-Bernal, D. Balderas, P. Ponce, M. Rojas, and A. Molina, “Implications of artificial intelligence algorithms in the diagnosis and treatment of motor neuron diseases—A review,” in Life, 13(4), 1031.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[30] L. Fei-Fei, J. Justin, Y. Serena, CS231n：Convolutional Neural Networks for Visual Recognition.

[31] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).