Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Target Recognition and Detection Based on Convolutional Neural Network Deep Learning

Xiaoxuan Chen[a,*], Renlong Zhang

Qingdao University of technology, Qingdao, China

ARTICLE INFO

Article history:
Received 22 March 2021
Accepted 29 May 2021
Available online 3 June 2021

Keywords：
Deep learning
Convolutional neural network
Target recognition
Object detection

ABSTRACT

Convolutional neural network is a new artificial neural network method combining artificial neural network and deep learning technology. It has the characteristics of local perception region, hierarchical structure, feature extraction and global training combined with classification process. It has been widely used in the field of image recognition. Modern image recognition tasks require the classification system to be able to adapt to different types of recognition tasks. Depth network and its special case convolutional neural network are the research hotspots in the field of artificial neural network. The research on convolutional neural network and its application in different recognition tasks has important application value.

1. Introduction

In recent years, machine vision has been more and more applied to intelligent vehicles and intelligent transportation systems. The application of machine vision in intelligent vehicles mainly depends on the on-board system. By fixing the camera on the vehicle, it can identify the lane, pedestrian, vehicle, traffic sign and other targets. The application of machine vision in intelligent transportation system mainly depends on the camera fixed above the roadside to identify and detect traffic flow, vehicle license plate and other information.

Object detection is one of the basic tasks in the field of computer vision. There are many methods of target detection. For example, Collins and others use neural network to identify people and vehicles, Fung and others use feature point detection to estimate the shape of vehicles. In recent years, with the vigorous development of neural network, deep learning and other technologies, the target detection algorithm has changed from the traditional algorithm based on manual features such as hog and SVM to the detection technology based on deep neural network.

2. Theoretical basis of traffic signs

2.1 Definition of traffic signs

Traffic signs, road facilities that convey guidance, restriction, warning or indication information in words or symbols. Also known as road signs and road traffic signs. In traffic signs, it is generally safe to set eye-catching, clear and bright traffic signs, which is an important measure to implement traffic management and ensure road traffic safety and smoothness. There are many types of traffic signs, which can be divided into: main signs and auxiliary signs; Movable signs and fixed signs; Lighting signs, luminous signs and reflective signs; And variable information signs reflecting changes in driving environment.

2.2 Definition of some traffic signs

2.2.1 Warning Sign

It is mainly used for warning. The color feature of the sign is that the bottom plate is yellow, the three border patterns are black, and the contents marked inside are black. The shape features of the signs are regular triangles, such as paying attention to pedestrians, traffic lights, children, crosswind, roundabout, sharp left turn, etc. Some of the signs are shown in Fig. 1.

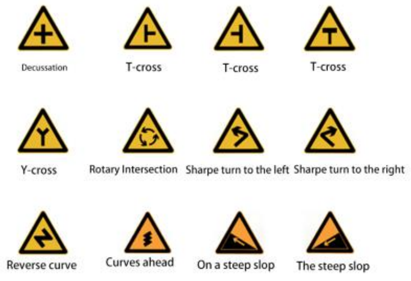

Fig.1. As shown in the figure, it is a warning sign. The color characteristics of the

[*] Corresponding author.
E-mail addresses: iclchenxiaox@163.com (X. Chen)

Doi:

sign are yellow for the base plate, black for the three border patterns, and black for the contents of the internal marks.

2.2.2 Mandatory Sign

Signs indicating the movement of vehicles and pedestrians. The color is blue background and white pattern; The shape is divided into circle, rectangle and square; Set near the road section or intersection where vehicles and pedestrians need to be indicated. Some signs are shown in Fig. 2.

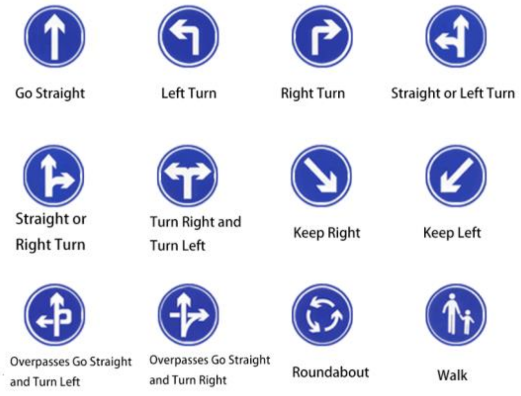

Fig.2. Some indicators are shown in the figure. This part of the sign can guide the unmanned vehicle to drive correctly.

In addition to the above two types of traffic signs, they also include tourism signs and road construction safety signs, which will not be studied in this paper. The distance between the traffic signs on the road and the road edge of the deep learning robot is 0.2m to 0.45m, and the height from its lower edge to the road surface is about 0.18m to 0.52m. In the image acquisition, the camera can be properly adjusted to obtain a high-definition image. In the detection and recognition, the background information at the bottom can be deleted according to the position of the signs in the image, so as to reduce the amount of calculation and save time, The accuracy of recognition is improved.

3. Overview of Yolo v5 convolutional neural network

3.1 Development history of Yolo

From Yolo v1 to Yolo v2, to Yolo 9000 and Yolo v3, Yolo has undergone several generations of changes. Yolo v1 creatively combines recognition and location; Yolo v2 improves the positioning accuracy and the recall rate; Yolo v9000 is a joint training method, which can detect more than 9000 types of models; Yolo v3 has two major improvements: using residual model and using FPN architecture to achieve multi-scale detection; Yolo v5 provides us with two optimization functions Adam and SGD, and both preset the training super parameters matching them. Yolo v5 is an excellent target detection algorithm. In the development process, while maintaining its speed advantage, Yolo v5 constantly improves the network architecture, absorbs various skills of other excellent target detection algorithms, and successively introduces the anchor box mechanism and FPN to achieve multi-scale detection, which improves the speed and accuracy of detection.

In terms of target detection, we choose to use the more advanced yolov5 model for the following reasons: markers are an important part of the traffic system and have an important significance in ensuring the safety of drivers' lives and property. However, in reality, due to the weather speed and the shielding damage of the markers, it will lay hidden dangers for drivers to drive safely. Therefore, the use of convolutional neural network has the characteristics of strong learning and classification ability, and the research on traffic sign recognition can significantly improve its recognition rate, so it has more important application value. And Yolov5 network is the smallest, with the lowest speed and the lowest AP accuracy. However, due to the small target detection, the pursuit of speed is a good choice. On this basis, the other three networks are deepening and widening, and the AP accuracy is also improving, but the speed consumption is also increasing. At present, the Yolov5 model is more than 10 MB in size, which is very fast, and the online production effect is considerable. Embedded devices can be used.

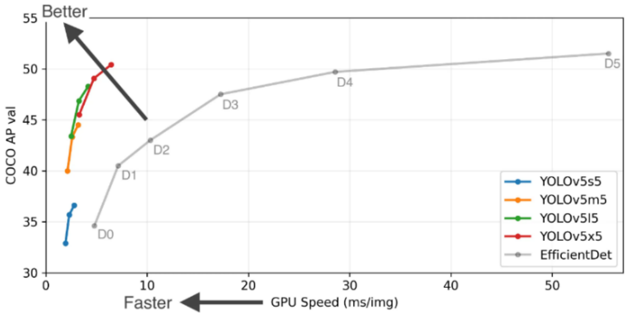

Fig.3. GPU recognition speed between Yolo models. It can be seen from the figure that the recognition speed of Yolo v5 model is improved a lot.

3.2 Composition of Yolo v5 model

Yolo v5 is mainly divided into four parts: input end, benchmark network, Neck network and head output end. Compared with Yolo v3 and Yolo v4, the four parts are improved in different aspects;

Input end: the optimization part mainly includes the main Mosaic data enhancement, adaptive anchor box calculation, and adaptive image scaling;

Benchmark network: Focus structure and slicing operation are added, and CSP structure is adopted. The Focus structure does not exist in Yolo v3&Yolo v4. The key is slicing. For example, in the above diagram, the 4 * 4 * 3 image is sliced into a 2 * 2 * 12 feature map. Taking the structure of Yolo v5s as an example, the original 608 * 608 * 3 image is input into the Focus structure. By using the slicing operation, it first becomes a feature map of 304 * 304 * 12, and then after a convolution operation of 32 convolution cores, it finally becomes a feature map of 304 * 304 * 32.

Neck network: The current Neck of Yolo v5 is the same as that of Yolo v4, which uses the FPN+PAN structure. However, when Yolo v5 first came out, only the FPN structure was used, and then the PAN structure was added. In addition, other parts of the network were also adjusted. Target detection network often inserts some layers between BackBone and the final Head output layer, and Yolo v5 adds FPN+PAN structure;

Head output layer: the anchor frame mechanism of the output layer is the same as that of Yolo v4, and the main improvement is the loss function GIOU during training_ Loss, and DIOU filtered by the prediction box_ Nms。

3.2 Training process of Yolo v5network

3.2.1 The Yolo detection system first divides the input image into S × S grids.

The network uses three scales to predict the output (52 × 52，26 × 26，13 × 13).Three different detection frames (box1, box2 and box3), which are the most common for such objects, are used to predict the objects in the image.

3.2.2 Predict the probability P_0 that the detection frame of each grid contains objects.

$$P_0 = P(object) * IOU \quad \text{......................} \quad (3.1)$$

If the center of an object falls in a grid, *P (object)* = 1; otherwise *P (object)* = 0.IOU is the overlap ratio between the detection box and the groundtruth. IOU = area of intersection part / area of union part. When the two boxes completely coincide, IOU = 1, and when they do not intersect, IOU = 0. If it is greater than 0.5, keep the box; if it is less than 0.5, delete the box; The definition and calculation method of IOU are shown in Fig.6.

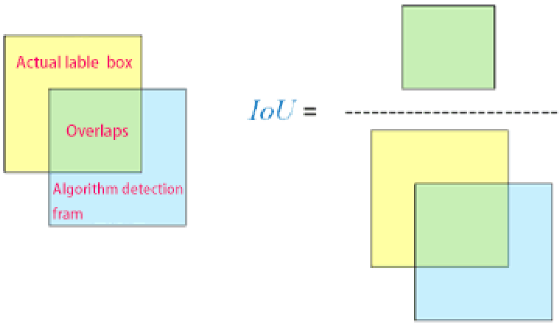

Fig.4. The definition and calculation method of the IOU are shown in Figure 6.

3.2.3 The network selects the box with the highest IOU value as the prediction box.

The following pictures (Fig.7, Fig.8) show the results of traffic sign recognition using the trained Yolo v5 network.

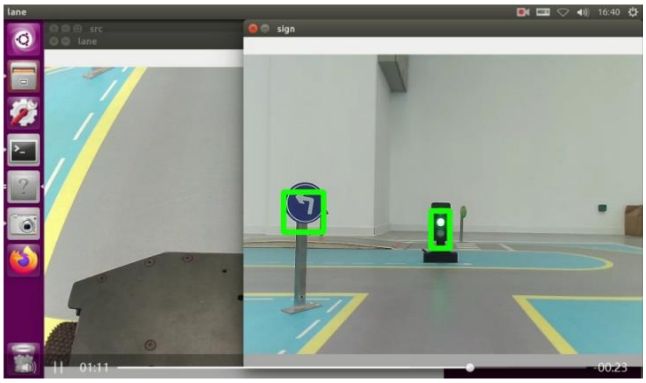

Fig. 5. As shown in the figure, the intelligent vehicle camera based on deep learning recognizes and detects crosswalk signs.

3.2.4 The network finally performs NMS non maximum suppression

In the target detection algorithm, when detecting candidate areas, there may be multiple candidate boxes. Check the highly overlapping candidate boxes for detecting targets of the same category, retain the candidate box with the highest target score, and remove the highly overlapping candidate box to reduce multiple detection of the same target, leaving only one optimal detection box. This process is called non maximum suppression (NMS).

First, the prediction boxes in the candidate area are sorted according to the classifier category score from high to low, the candidate box with the highest score is selected, and the IOU values of the other candidate boxes are calculated in turn; Then, set the threshold value. If the IOU value is greater than the set threshold value, delete the detection box with lower score; Finally, continue to select the detection box with the highest score from the unprocessed detection boxes, and repeat until all prediction boxes are processed.

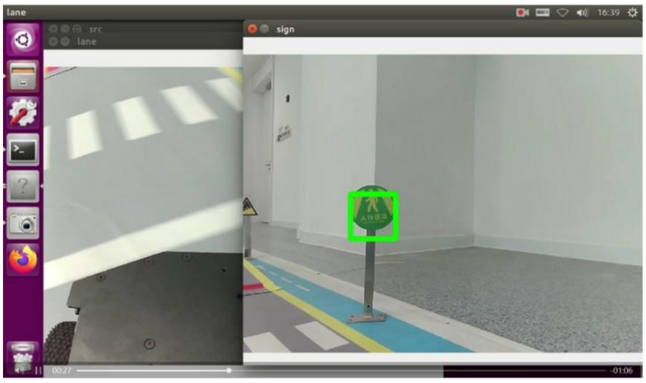

Fig.6. As shown in the figure, the smart car camera based on deep learning recognizes and detects traffic lights and left turn signs.

$$s_i = \begin{cases} s_i, iou(M, b_i) < N_t \\ 0, iou(M, b_i) \geq N_t \end{cases} \quad \text{......................} \quad (3.2)$$

In the above formula, S_i represents the possibility that the window may be filtered out, $iou(\ )$ represents the cross merge ratio of areas, M represents the candidate box with the highest confidence score in the ith screening, b_i represents the candidate box except M, and N_t represents the preset threshold. We set the threshold value to 0.45.

In the post-processing process of target detection, NMS operations are usually required for filtering many target boxes. Because CIOU_ Loss contains the influence factor v, which involves the information of groundtruth. When reasoning is tested, there is no groundtruth. So Yolo v4 is in DIOU_ DIOU is adopted based on Loss_ NMS, while Yolo v5 uses weighted NMS.

4. Realization principle of target recognition

4.1 Pretreatment of image data

In order to improve the accuracy of image recognition, the interference information in the image is removed. Before using depth learning to train the lane line extraction model, we use color separation to achieve the extraction under the most ideal light environment, so as to ensure that the car can automatically and smoothly travel, and keep collecting images.

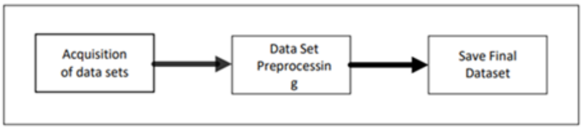

Fig.7. Data preprocessing flow chart. After the data is obtained, it will be preprocessed to obtain the dataset.

When collecting images, different vehicle speeds are set to collect images with different sharpness, so as to enhance the generalization ability of the vehicle's lane extraction model at different vehicle speeds. Then the image is preprocessed, and the processed data is sent to the network for training. The overall process of this part is shown in Fig.9.

4.2 RGB principle

RGB color space is also known as the three primary color model. It is an additive color model that superimposes red, green, and blue primary colors in different proportions to produce other colors. RGB color space can express all colors. Image acquisition, detection and display devices in electronic systems mostly use RGB color space, such as computer monitors, cameras, etc. In RGB color space, the three primary color components can be mapped to the three-dimensional space as the coordinate values of the Cartesian coordinate system in the European geometric space. The intensity of each color component ranges from 0 to 255. When the value of the three primary colors is the minimum, it means black. When the value of the three primary colors is the maximum, it means white. Although the three color components of RGB color space are independent channels, each channel contains the brightness information and color information of the image. There is a great correlation between the three components. Changing one of the three components will lead to changes in the perception of other components by the human eye. In the process of collecting lane line data in the laboratory, the light is always changing, and the brightness in the video screen also changes with the change of light. If RGB spatial data is used to input the CNN model, the segmentation effect will be affected because the brightness information and chroma information cannot be separated, so RGB spatial images will not be used.

4.3 HSV principle

HSV color space was proposed by American computer scientist A.R. Smith in 1978. HSV color space is closer to the intuitive characteristics of color, similar to the way human eyes perceive color. HSV stands for Hue, Saturation and Value, respectively. The HSV color space model is shown in the following figure. This inverted cone model is called the Hexagon Model. There are three channels in the HSV color space. The angle around the central axis represents hue H. Hue can be understood as the name of a color, and the value range is 0 to 360. Saturation S is expressed by the length from the central axis, and the value range is 0 to 1. The height along the central axis represents the brightness V, which is used to express the brightness of the color, and the value range is 0 to 1.

Compared with RGB color space, HSV color space can separate the component of luminance V, which greatly reduces the impact of external light intensity. The conversion between HSV color space and RGB color space is through nonlinear transformation. The formula is as follows：

$$h=\begin{cases}0 & \text{if } \max=\min\\ 60\times\dfrac{g-b}{\max-\min} & \text{if } \max=r \text{ and } g\geq b\\ 60\times\dfrac{g-b}{\max-\min}+360 & \text{if } \max=r \text{ and } g<b\\ 60\times\dfrac{g-b}{\max-\min}+240 & \text{if } \max=b\\ 60\times\dfrac{g-b}{\max-\min}+120 & \text{if } \max=g\end{cases} \quad \cdots\cdots(4.1)$$

$$s=\begin{cases}0 & \text{if } \max=0\\ \dfrac{\max-\min}{\max}=1-\dfrac{\min}{\max} & \textit{otherwise}\end{cases} \quad \cdots(4.2)$$

Since the pictures collected by the image acquisition device are based on RGB color space, the above nonlinear transformation is required for color space conversion when HSV color space is used.

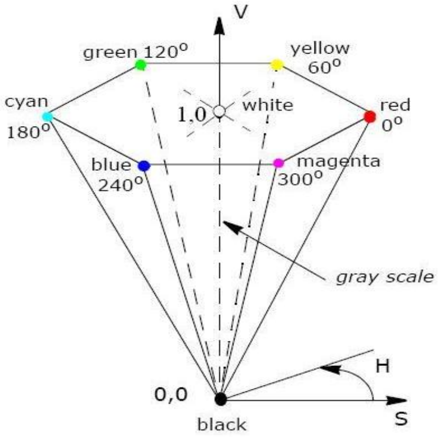

Fig.8. HSV color space model. The HSV color space model is shown in the figure. This inverted cone model is called the Hexagon Model

4.4 Gaussian filtering

Long time data collection leads to outdoor light hitting the track through the glass, which changes the picture. Here I use the preprocessing technology of Gaussian filtering (prior to RGB to HSV). Gaussian filtering is a kind of linear smoothing filtering, which is suitable for eliminating Gaussian noise and is widely used in the noise reduction process of image processing. Gaussian filtering is the process of weighted averaging the whole image. The value of each pixel is obtained by its own and other pixel values in the neighborhood after weighted averaging.

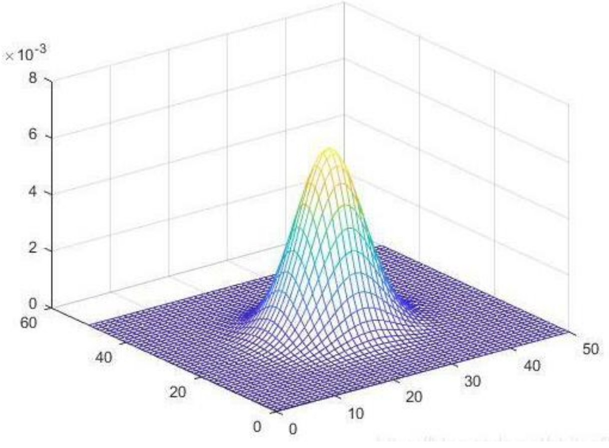

Fig.9. Two-dimensional Gaussian distribution. Gaussian filtering is a kind of linear smoothing filtering, which is suitable for eliminating Gaussian noise and is widely used in the noise reduction process of image processing.

The Gaussian filtering process of two-dimensional images is divided into two steps:
① The mask is calculated by two-dimensional Gaussian normal distribution function: the matrix is normalized by Gaussian changes. Gauss mask only needs to be calculated once. When this mask is

obtained, all pixels of the image are deconvoluted with the same set of masks.

② The convolution operation between the value of each position on the mask and the pixel value of the corresponding position on the image.

5. Overview of target identification and detection process

The process of target recognition and detection can be roughly divided into the following three processes. First, collect map information through handle manipulation. Second, the car motherboard converts photos into data points through in-depth learning, and the data points are collected to form a data set that can be recognized by the car. Third, the car carries out in-depth learning inside the CPU.

5.1 Manual collection of map information

The deep learning car based on convolutional neural network first revolves around the map under the operation of the wireless handle. The purpose is to collect all the information of the map, transmit the photos in various places in real time, and finally form a data photo database.

5.2 Transformation of photo set and data set

In the object detection algorithm, image annotation is a process of generating image text description by manually dividing the location information and category information of the target region in the image. Image annotation can be regarded as the cognitive behavior of machine imitating human learning process to acquire prior knowledge. Computer can recognize and distinguish the tagged images in advance, so as to continuously recognize image features and finally achieve autonomous recognition.

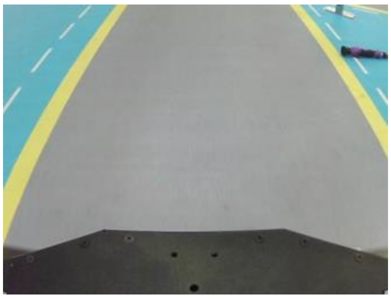

Fig.10. This picture is the original photo collected by the camera of the deep learning unmanned vehicle. It can be seen that the photo is greatly affected by the ambient light. In the process of unmanned driving, the ambient light is one of the factors most likely to affect the driving quality.

For the collection of markers, we adopt the method of placing the trolley on the track and repeatedly pushing the trolley at various angles before and after the markers. A total of five types of signs were collected, including crosswalk, uphill, speed limit 10, speed limit cancellation, and left turn, with about 2000 pieces of each type, to ensure that the number of labels in the data set is the same as far as possible. After manual filtering (some pictures that do not contain markers or have too few signs displayed are removed, and some pictures with appropriate size and distance and slight blurring are retained to enhance the robustness of the model and improve the prediction accuracy of the car in the real travel process), the wizard annotation assistant is used to annotate the image jpg file, and the annotated file is converted into xml format.

The in-depth learning intelligent vehicle processes the collected map information through relevant algorithms, such as extracting the lane line and relevant traffic sign information of each photo by means of photo brightness enhancement and gray processing. This kind of useful information is uniformly transformed into relevant data sets through internal algorithms, which is convenient for fast operation and improves operation efficiency.

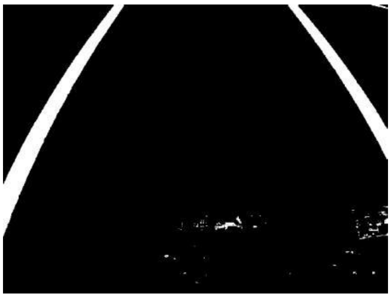

Fig.11. The picture is a lane line picture after in-depth learning. It can be seen from the picture that after algorithm operation, the computer will only recognize the relevant lane line, and the ambient light factor can be ignored, so as to improve the reliability and feasibility of unmanned driving.

5.3 CPU internal autonomous learning

After the CPU obtains the training data set, it needs to analyze and compare the network model, select the most suitable network model for traffic target detection, enter the experimental stage, use the configured network model to train on the specific data set, evaluate the obtained training model, and repeat the process.

6. Summary and Prospect

Through the deep learning intelligent vehicle based on convolutional neural network, this paper makes a theoretical analysis on the recognition and detection of traffic sign signals, and summarizes the architecture of convolutional neural network and the general process of deep learning. At the same time, the rapid development of Yolo V3 provides reliability and feasibility for unmanned driving, which is conducive to the in-depth development of unmanned driving technology in the future. Because unmanned driving is closely related to human life, the method of image target detection in the field of unmanned driving is not mature enough, and the detection accuracy needs to be improved.

Next, consider how to balance the detection accuracy of different targets, whether collecting more data or virtual some target images through image preprocessing, which will be a very important factor to improve the detection accuracy. Secondly, we consider generalizing the model so that it can be applied to more fields. Finally, if we can improve the network model, reduce the amount of calculation of network training and improve the speed of model training, it will be a great contribution to the research and experiment of target detection task.

References

Zhang Ting. Driverless strategy analysis based on Baidu Apollo 2.0 [J]. Modern economic information, 2019 (3): 394

Han Yi. Research on moving vehicle detection and tracking algorithm based on video image [D]. Harbin University of technology, 2015

A. delaEsealera, L. E. Moreno, M. A. Salichs, etal. Road traffic sign detection and classification[J]. IEEE Transactions on Industrial Electronies, 1997, 44(6): 848-859.

Mahanty Mohan and Bhattacharyya Debnath and Midhunchakkaravarthy Divya. A Review on Deep Learning-Based Object Recognition Algorithms[M].

Springer Nature Singapore, 2022 : 53-59.

Goel Rohini and Sharma Avinash and Kapoor Rajiv. Deep Learning Based Object Recognition in Real Time Images Using Thermal Imaging System[M]. 2021 : 348-354.

Rosso Marco Martino, Marasco Giulia, Aiello Salvatore, et al. Convolutional networks and transformers for intelligent road tunnel investigations[J]. Computers and Structures, 2023, 275

Xin Liu, Junhui Wu, Yiyun Man, et al. Multi-objective recognition based on deep learning[J]. Aircraft Engineering and Aerospace Technology, 2020, 92(8):1185-1193.

Chen Guosheng, Lian Wenjun, Hu Fudong, et al. Research on Intelligent Target Recognition Method Based on Pattern Recognition and Deep Learning[J]. SECOND TARGET RECOGNITION AND ARTIFICIAL INTELLIGENCE SUMMIT FORUM, 2020, 11427

Rosso Marco Martino, Marasco Giulia, Aiello Salvatore, et al. Convolutional networks and transformers for intelligent road tunnel investigations[J]. Computers and Structures, 2023, 275

Zhen Zhen Xing, Xing Zhen Zhen, . Image recognition algorithm based on spatial transform convolutional neural network[J]. Journal of Physics: Conference Series, 2020, 1651(1):012133-.

Zhichao Xian, . Survey of image recognition technology based on convolution neural network. 2020, 16

Gao Zhiyu, Liu Bailin, Gu Hongxian, et al. Research on the Application of Convolutional Neural Networks in the Image Recognition[J]. International Journal of Advanced Network, Monitoring and Controls, 2020, 5(2):31-38.

Hanqing Hu, Jin Lyu, Xiaolin Yin, et al. Research and Prospect of Image Recognition Based on Convolutional Neural Network[J]. Journal of Physics Conference Series, 2020, 1574(1):012161.

Cheng Richeng, . A survey: Comparison between Convolutional Neural Network and YOLO in image identification[J]. Journal of Physics: Conference Series, 2020, 1453:012139-012139.

Xin Liu, Junhui Wu, Yiyun Man, et al. Multi-objective recognition based on deep learning[J]. Aircraft Engineering and Aerospace Technology, 2020, 92(8):1185-1193.

Zhou Lijie, Yu Weihai, . Improved Convolutional Neural Image Recognition Algorithm based on LeNet-5[J]. Journal of Computer Networks and Communications, 2022, 2022

Wu Wenbo, Pan Yun, . Adaptive Modular Convolutional Neural Network for Image Recognition.[J]. Sensors (Basel, Switzerland), 2022, 22(15):5488-5488.

Ziyadinov Vadim, Tereshonok Maxim. Noise Immunity and Robustness Study of Image Recognition Using a Convolutional Neural Network[J]. Sensors, 2022, 22(3):1241-1241.

Liu Zhizhe, Sun Luo, Zhang Qian, et al. High Similarity Image Recognition and Classification Algorithm Based on Convolutional Neural Network.[J]. Computational intelligence and neuroscience, 2022, 2022:2836486-2836486.

Rana Ajay, Chauhan Kuldeep, Sun Y., et al. Computer vision and machine learning for image recognition: A review of the convolutional neural network (CNN) model[J]. Asian Journal of Multidimensional Research, 2021, 10(10):1023-1029.

Zhang Hanwen, Qin Zhen, Xie Hua, et al. Image recognition and detection based on fast area convolutional neural network[J]. Journal of Physics: Conference Series, 2021, 1976(1)

Li Zhongyu, Wang Huajun, Cao Yuhang, et al. Research on the identification of obstacle image based on convolutional neural network. 2021, 11848:1184807-1184807-7.

Wang Hao, Jiao Kaijie, . Blind guidance system based on image recognition and convolutional neural network[J]. IOP Conference Series: Earth and Environmental Science, 2021, 769(4)

Li Xiaohong, Lv Xiangfeng, . Research on Image Recognition Method of Convolutional Neural Network with Improved Computer Technology[J]. Journal of Physics: Conference Series, 2021, 1744(4):042023-.

Chen Xiaoxuan, male, 2019 undergraduate of Qingdao University of technology, majoring in electrical engineering and automation.

Zhang Renlong, male, 2018 undergraduate of Qingdao University of Technology, majoring in automation..