# Review of Research Progress in Computer Vision

Bohan Ma[1], Fuzhu Li[2],Yi Ren[2*]

[1] Beijing Bayi School,Beijing,100080,China,mabohan815@Yeah.net

[2] Jiangsu University,Zhenjiang,212013,China,lifuzhu@ujs.edu.cn

*Corresponding author

E-mail address: renyi_ujs@163.com

**Abstract.** Computer vision, being a pivotal element of artificial intelligence, has witnessed significant advancements in theoretical methodologies, key technologies, and practical applications after years of development, which has shown great application prospects in medical, industrial, transportation, military, public safety, and other fields. Therefore, it is necessary to review the current development and application of computer vision technology to comprehensively evaluate its future development trend and potential. Firstly, this paper reviews the development history of computer technology. Then, aiming at the specific tasks of computer vision, the typical methods and models based on depth learning in recent years are summarized, including image classification, target detection, image segmentation, etc. Then, the application and research progress of computer vision technology in various vital fields are introduced respectively. Finally, the current challenges and critical research problems in computer vision technology are analyzed.

**Keywords:** Computer vision, Convolutional neural networks, Image Classification, Target Detection.

## 1. Introduction

Computer Vision (CV) stands as a foundational technology within the realm of artificial intelligence. Its primary focus lies in the acquisition, manipulation, and interpretation of visual data encompassing images and videos from the physical environment [1, 2]. It endows machines with the ability to perceive vision information and covers many significant CV technical fields such as image recognition, face recognition, edge detection, motion detection, optical character recognition, and machine vision [3]. The research on computer vision technology began as early as the 1950s, and the publication of Vision in 1982 penned by David C. Marr heralds that computer vision becomes an independent discipline [4, 5]. After the 21st century, computer vision has realized intelligent applications in multiple scenes and fields through cross-integration with artificial intelligence technologies such as big data analysis and machine learning [6, 7]. Based on statistics from reputable organizations, over the past decade, computer vision technology has witnessed a remarkable surge in accuracy, soaring from a modest 50% to an impressive 99%, with its business application fields and market scale increasing day by day [8]. By the end of 2020, the global market size of computer vision is 9.45 billion US dollars, which is expected to reach about 41 billion US dollars by 2030 [9].

Computer vision is mainly divided into two research fields, that is, 2D vision and 3D vision. The research on 2D vision includes target recognition, target tracking, video content understanding, etc. The research on 3D vision includes 3D reconstruction based on images, 3D pose estimation of objects, etc. [10, 11]. In the past decade, computer vision technology has achieved a series of leaps from low-level tasks (such as boundary recognition) to high-level ones (such as scene understanding) mainly due to three factors. Namely, (1) the maturity of deep learning (DL) technology [12]; (2) the improvement of localization computing ability of graphics processing unit (GPU) [13]; (3) the open source of large tag data sets for training algorithms [14]. During this period, the establishment of ImageNet large-scale visual recognition challenge (ILSVRC) provided a powerful platform for CV technology promotion, which gathered a large number of DL researchers and consolidated them to compete and cooperate with each other, greatly enhancing the technology of various CV tasks [15, 16]. The most important threshold faced by early CV technology is feature engineering. In other words, it is necessary to find suitable features to characterize the research objects and combine suitable classifiers. However, this process requires abundant manpower and time costs, so it is difficult to extend to various fields [17, 18]. As a fundamental component of DL, convolutional neural networks (CNN) has pioneered a new situation for CV researchers [19]. The concept of CNN was mentioned as early as the 1980s, but the effect of shallow CNN is not as good as the above-mentioned feature engineering + classifier. It is not until the development and enrichment of computing power and data that deep CNN becomes possible. In 2012, on the ILSVRC platform, a CNN-based method achieved 83.6% of the top 5 classification accuracy, which was about 10% higher than that in previous years [20]. Nowadays, CNN is the main Backbone architecture in the industry. By transforming and optimizing the classic Backbone architecture, algorithm engineer can quickly adapt to business and migrate the model to its own vertical domain, which has been instrumental in technical tasks like image classification detection, image registration, image retrieval, image reconstruction and enhancement [21, 22].

CNN technology relies heavily on supervised learning, in which the optimal model is obtained by supervised training on the generated data based on the data set. This method needs to use the data set containing data points (such as images) and data labels (such as object classification). However, considering the sparsity and low openness of data in many fields, a transfer learning algorithm is proposed, which will first train in large and unrelated data sets (such as ImageNet), and then fine-tune in data sets to be applied (such as medicine and transportation), thus promoting the efficiency of learning and training [19, 23]. Additional, to mitigate the expenses associated with data collection and annotation, some research is developing data synthesis technologies, such as data enhancement, generating adversarial networks (GANs) [24], and crowdsourcing image annotations to generate effective algorithms [25]. In recent years, some studies have proposed a self-supervised learning technology [26] that can extract implicit tags from data points to train algorithms, which makes this field completely unsupervised learning without data tags. These technologies will reduce obstacles in developing and deploying data sets in practical fields.

On the other hand, the ILSVRC platform reported the first method of driving DL by GPU in 2012 [27]. CV can only perform addition, subtraction, multiplication, and division linearly one by one in the face of matrix operation and pixel block convolution from the previous pure CPU calculation, which greatly limits the throughput speed. At the birth of GPU, to meet the processing scene conditions of images and videos, it is necessary to process the mathematical calculations of each pixel or pixel block in parallel in design, which naturally enables thousands of arithmetic logic units (ALU). Thus, it provides convenience to process independent mathematical calculations and greatly accelerates the matrix operations that often occur in the calculation and reasoning of deep learning [28, 29].

## 2. Typical Network Model Under Specific Tasks

### 2.1 Image Classification

Image classification, as a fundamental task in computer vision, involves assigning appropriate labels to images from a predefined set of categories. Traditional image classification algorithms usually include low-level feature learning, feature coding, spatial constraint, classifier design, model fusion, etc. [31] as shown in Figure 1. Alex Krizhevsky [31] won the championship in the ILSVRC 2012 for proposing the CNN model, which is called AlexNet.

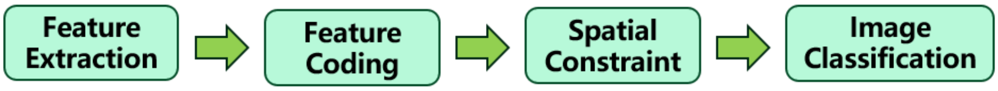


Figure 1 Analysis of Traditional Image Classification Algorithm

- AlexNet[31]

The emergence of AlexNet is of milestone significance to the development of deep learning. Figure 2 shows the network structure of AlexNet. This network include one input layer, five convolutional layers (Conv1, Conv2, Conv3, Conv4, Conv5), two fully connected layers (FC6, FC7), and one output layer.The input is 224 × 224 × 3 images with RGB channels, and each convolution layer includes a convolutional kernel, bias term, ReLU activation function, and local response normalization (LRN) module. Additionally, after Conv1, Conv2, and Conv5, there is a subsequent maximum pooling layer, with softmax serving as the final output layer, which converts the network output into probability value for predicting the image category.

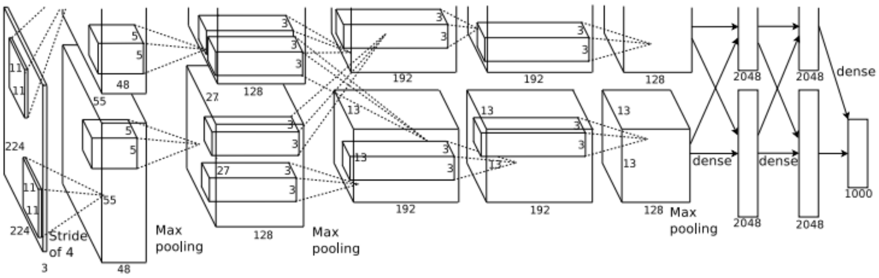
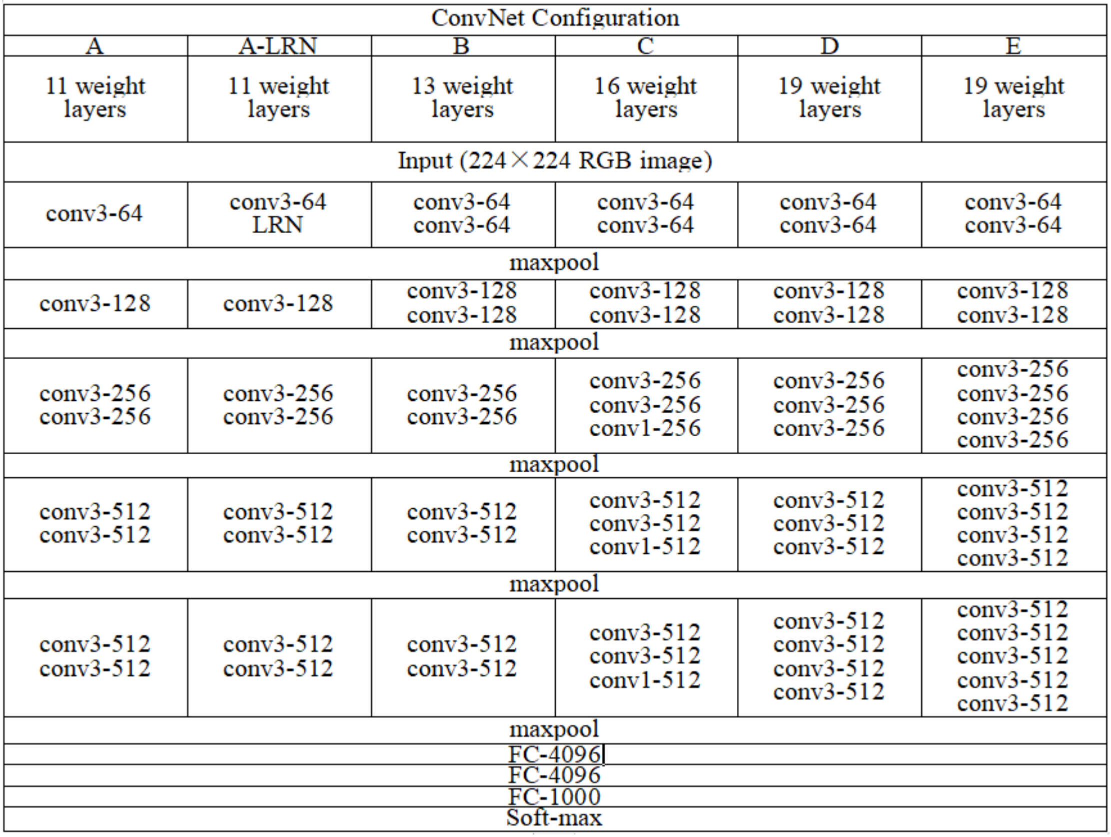


Figure 2 AlexNet Network Architecture

- VGG[32]

The VGG network was proposed by the Visual Geometry Group (VGG) at the University of Oxford. One notable improvement of VGG compared to AlexNet is the substitution of larger convolutional kernels (11x11, 5x5) with a series of consecutive 3x3 convolutional kernels. This choice is motivated by the fact that, for a given receptive field, employing stacked smaller convolutional kernels is more advantageous than using a single large kernel. This approach allows for increased network depth through multiple nonlinear layers, facilitating the learning of more complex patterns while minimizing parameter usage and associated costs. The architecture of the VGG network is presented in Figure 3.

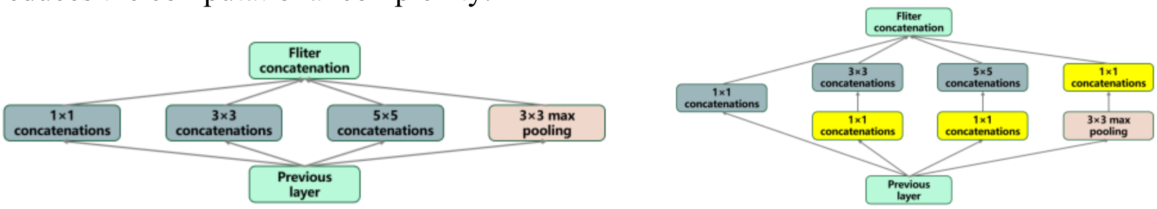


| ConvNet Configuration | | | | | |
|---|---|---|---|---|---|
| A | A-LRN | B | C | D | E |
| 11 weight layers | 11 weight layers | 13 weight layers | 16 weight layers | 19 weight layers | 19 weight layers |
| Input (224×224 RGB image) | | | | | |
| conv3-64 | conv3-64<br>LRN | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 |
| maxpool | | | | | |
| conv3-128 | conv3-128 | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 |
| maxpool | | | | | |
| conv3-256<br>conv3-256 | conv3-256<br>conv3-256 | conv3-256<br>conv3-256 | conv3-256<br>conv3-256<br>conv1-256 | conv3-256<br>conv3-256<br>conv3-256 | conv3-256<br>conv3-256<br>conv3-256<br>conv3-256 |
| maxpool | | | | | |
| conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>conv1-512 | conv3-512<br>conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>conv3-512<br>conv3-512 |
| maxpool | | | | | |
| conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>conv1-512 | conv3-512<br>conv3-512<br>conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>conv3-512<br>conv3-512<br>conv3-512 |
| maxpool | | | | | |
| FC-4096 | | | | | |
| FC-4096 | | | | | |
| FC-1000 | | | | | |
| Soft-max | | | | | |

Figure 3 VGG Network Architecture

- GoogLeNet[33]

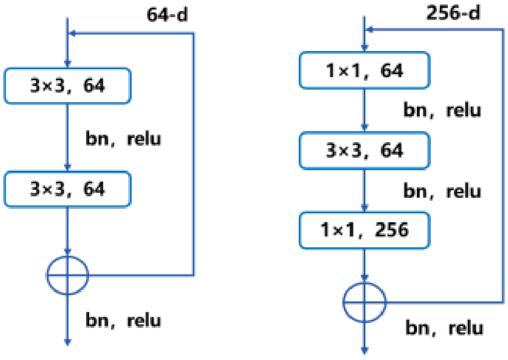
GoogLeNet is a cutting-edge deep learning architecture introduced by Christian Szegedy in 2014, and it emerged as the champion in the ILSVRC 2014 competition. The Inception module's fundamental structure is depicted in Figure 4. Figure 4a showcases the simplest design, which involves the concatenation of feature maps from 3 convolutional layers and 1 pooling layer. However, this design has a drawback: the pooling layer does not alter the number of feature channels, resulting in a substantial increase in feature channels when concatenated. Consequently, as multiple layers of such modules are stacked, the parameters and computations also escalate. To improve this shortcoming, three 1 × 1 convolutions are introduced in Figure 4b to reduce the dimension. Besides, superimposing more convolution in the receptive field of the same size can extract richer features. Meanwhile, the Inception module incorporates 1 × 1 convolutions to reduce dimensionality, which also reduces the computational complexity.



(a) Inception Simple Module (b) Inception with Dimension Reduction Module
Figure 4 Inception Module Structure

- ResNet[34]

Traditional deep CNNs often encounter the issues of vanishing or exploding gradients, which hinder the training of deeper networks. In 2015, Kaiming He and his colleagues introduced Residual Network(ResNet) as a solution to this problem. By incorporating residual blocks, ResNet effectively addresses the gradient issue and enables the successful training of deeper networks. ResNet achieved a groundbreaking milestone by surpassing 100 layers in neural network depth, with some networks even exceeding 1000 layers. The residual module, depicted in Figure 5, showcases two connection modes. Figure 5a illustrates the connection mode of the basic module, which consists of two 3x3 convolutions with an equal number of output channels. On the other hand, Figure 5b demonstrates the connection mode of the Bottleneck module. ResNet training converges quickly, which trains hundreds or even nearly thousands of convolutional neural networks.



(a) Basic Module Connection Mode (b) Bottleneck Module Connection Mode
Figure 5 Residual Module Structure

## 2.2 Target Detection

R-Recently, advancements in deep learning technology have revolutionized target detection algorithms. Traditional approaches, such as HOG, SIFT, and LBP, which rely on manual feature extraction, have been replaced by machine learning techniques on the basis of deep neural networks [35, 36, 39]. Since Girshick [37] put forward the R-CNN model in 2014, target detection has garnered significant attention within the field of cv. After R-CNN, Girshick's team successively introduced the Fast R-CNN [38] model.

- R-NCC[35]

S-CNN was born in 2013. Being a pioneer in integrating deep learning into target detection algorithms, it holds immense significance in the advancement of such algorithms. The R-CNN algorithm stands as a representative of the two-step approach. In other words, the initial step involves generating region proposals, followed by the utilization of CNN for recognition and classification purposes. Its target detection system is shown in Figure 6.

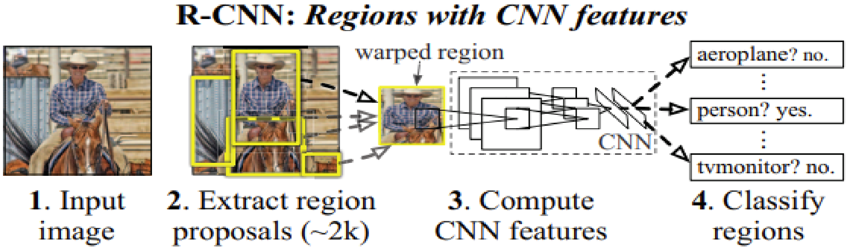


Figure 6 Overview of R-CNN Target Detection System

- SPP-Net[36]

To solve the problem of R-CNN, He Kaiming proposed the SPP-Net method, as depicted in Figure 7. The SPP-Net method incorporates a spatial pyramid pooling (SPP) layer between the convolutional layer and the fully connection layer of the network. This layer introduces spatial divisions to divide and resize the candidate regions before extracting convolutional features using the CNN. As a result, the input image size remains consistent throughout the network.

Figure 8 illustrates the network architecture that incorporates a SPP layer. Spatial pyramid pooling solves the inconsistent input candidate region sizes, but its more crucial significance lies in reducing repeated calculations in R-CNN, which greatly improves the efficiency of the algorithm.

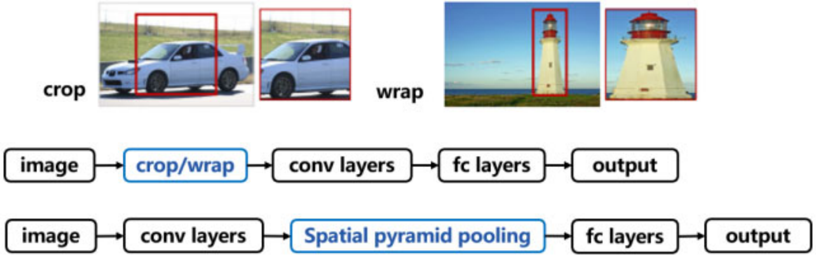


Figure 7 SPP Network Structure

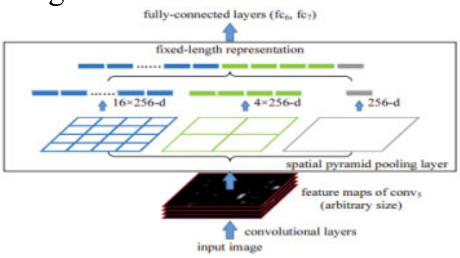

fully-connected layers (fc6, fc7)
fixed-length representation
16×256-d
4×256-d
256-d
spatial pyramid pooling layer
feature maps of conv5
(arbitrary size)
convolutional layers
input image

Figure 8 SPP Layer Network Structure

- Fast R-CNN[38]

In 2015, Girshick introduced an enhanced version of the R-CNN method, known as Fast R-CNN, to address the issue of detection speed. This improvement was achieved by incorporating shared convolutional feature extraction. Figure 9 illustrates the architecture of the Fast R-CNN model. The method's design includes a more efficient pooling layer structure, which resolves the need for the R-CNN methodto resize and scale image regions to a uniform size..

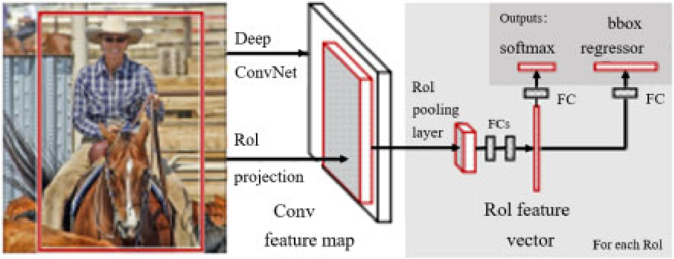


Figure 9 Fast R-CNN Structure

- Mask R-CNN[40]

In 2017, He Kaiming et al. further advanced the R-CNN method with the introduction of the Mask R-CNN. The Mask R-CNN approach builds upon Faster R-CNN by introducing additional branches for predicting target masks and split masks in parallel with the existing target detection box regression.

This extension allows for simultaneous object detection and segmentation within each region of interest (RoI). Specifically, binary classification semantic segmentation is performed within each RoI, enabling the identification of object boundaries. This branch operates concurrently with the classification and object detection box regression branch, as illustrated in the figure below.

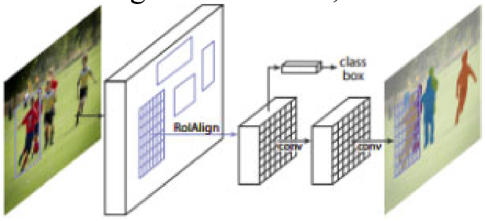


Figure 10 Mask R-CNN Architecture

### 2.3 Image Segmentation

Image segmentation is the task of semantic segmentation of images, which requires all pixels of the whole picture to be classified into one of predefined categories. It plays a vital role in diverse fields, including target localization, object boundary extraction, and image enhancement and editing.

- U-Net[41]

U-Net achieved remarkable success in the ISBI cell tracking competition in 2015, where it secured several first prizes. Initially designed for the specific task of cellular-level segmentation. The algorithm incorporates skip connections, represented by the gray arrows in Figure 14, to establish connections between the up-sampled output and the corresponding output of the sub-module in the encoder that matches the resolution of the input for the subsequent sub-module in the decoder. This strategic integration of high- and low-resolution information within U-Net yields significant benefits in the context of medical image segmentation.

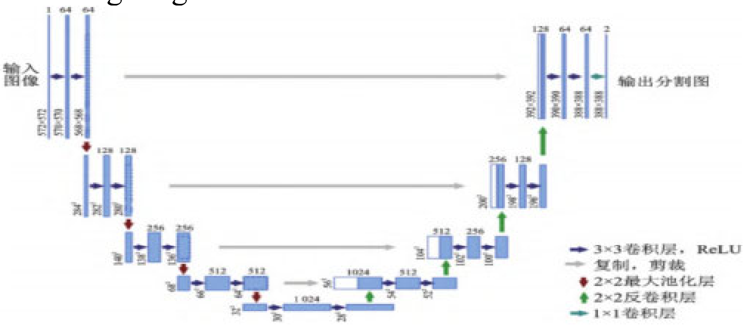


Figure 11 Schematic Diagram of U-Net Structure

- DeepLab[42]

The design of the pooling layer and up-sampling layer in deep CNN architectures can significantly impact the effectiveness of image segmentation. The parameters are not learnable and pooling will trigger the loss of spatial information and internal data structure of pixels, which in turn hampers the ability of up-sampling to accurately reconstruct small object details. Consequently, image segmentation' development has reached a bottleneck. To address this issue, DeepLab introduced the concept of dilated convolutions in 2016, which mitigates the information loss associated with pooling layers. Figure 12 illustrates the structure of this architecture.

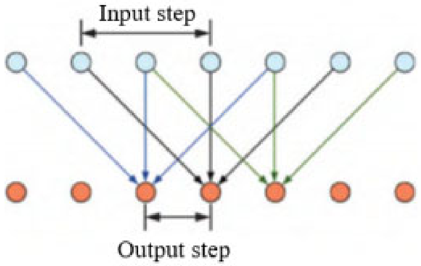


Figure 12 Hole convolution schematic

- SegNet[43]

The main motivation of SegNet proposed by Cambridge University in 2017 is to design a pixel-level image segmentation network for road and indoor scene understanding. Meanwhile, high efficiency in memory and computing time should be ensured. SegNet adopts the full convolutional structure of “encoder-decoder”, and the encoding network adopts the convolutional layer of VGG16. The decoder obtains the maximum pooled index from the corresponding encoder and then upsamples it to generate sparse feature mapping. Multiplexing pooled index reduces the number of parameters and improves the boundary division. As for the road scene segmentation data set, that is, CamVid 11 Road Class Segmentation, SegNet has a mIoU of 60.1% and a Boundary F1 score (BF) of 46.84%. As for the indoor scene segmentation data set of SUN RGB-D Indoor Scenes, almost all the deep network structures at that time did not perform well, but SegNet still surpassed other networks in most indicators. Figure 13 illustrates the structure of this architecture.

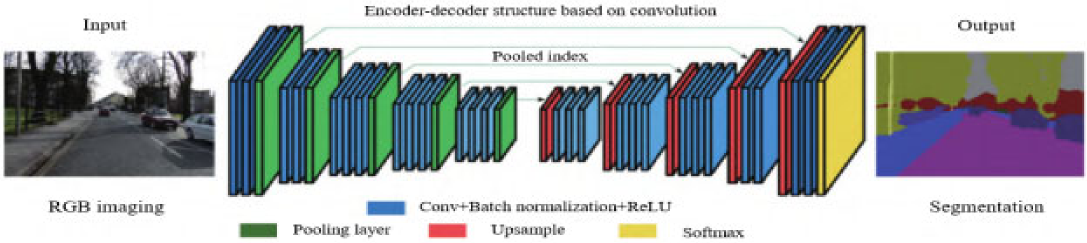

Input
Encoder-decoder structure based on convolution
Pooled index
Output
RGB imaging
Conv+Batch normalization+ReLU
Segmentation
Pooling layer
Upsample
Softmax

Figure 13 SegNet Structure schematic

## 3. Application

### 3.1 Biomedicine

- Medical Image Analysis

Medical image is an indispensable auxiliary means of the modern medical care system. Deep CNNs (DCNNs) have been extensively utilized for medical image segmentation tasks, computer-aided diagnosis, disease detection and classification, medical image retrieval, and other technical fields for its powerful learning ability and portability [44, 45]. The capability of DL models in medical image segmentation is heavily influenced by two key factors: training dataset' size and the quality of image annotation. In many medical image analysis tasks, especially in 3D scenes, great difficulty comes from building a large and high-quality training data set. Zhou [46] and Donahu [47] et al. pointed out that transferring learning technology can solve the insufficient training and further improve the quality of images.

- Assist Wound Assessment

Chronic trauma is a public health problem with wide influence in the world. The evaluation of wound healing is established by observing the epithelialization process and measuring wound contraction. Wound imaging, facilitated by computer vision technology, offers doctors a direct and comprehensive means of obtaining crucial information [48]. In the aspect of 2D model feature extraction, Wang et al. [49] proposed a cascade two-stage classifier based on support vector machines (SVM) to accurately determine diabetic foot ulcers' area. To achieve this, they employed the SLIC method for superpixel segmentation of the wound image. The color and texture features of these superpixels were then utilized as input for the classifier. The first stage of the classifier performed binary classification to distinguish between injured and uninjured areas. Instances that were misclassified by the first stage were further refined by the second stage of the SVM classifier for more accurate classification. Finally, the detection boundary is refined by a conditional random field. Song et al. [50] employed a combination of K-means clustering, edge detection, threshold segmentation, and region-growing techniques in both gray and RGB images to extract 49 informative features from wound images. These features encompassed various aspects such as area, centroid, minimum and maximum intensity, etc. The extracted features were then filtered and organized into a comprehensive feature vector, which served as the input for training multilayer perceptron (MLP) and radial basis function (RBF) neural networks. In the aspect of 3D model feature extraction, Anghel et al. [51] introduced a novel device called the 3D Wound Measurement (3DWM) device. It's designed specifically for accurately measuring various wound parameters such as depth, and volume. To achieve this, they employed an interactional graphics cutting method for wound segmentation. The largest rectangular package, representing the wound boundary, was then utilized for precise length and width measurements. Additionally, a depth map consisting of pixel position depth values was employed for accurate volume measurement. In addition to this research, several systems and tools have also been developed for 3D wound measurement. Nixon [52] et al. mentioned a 3D system. It incorporates the Silhouette Ettestar™ (3D camera), Silhouette Teconnect™ (3D structure creation framework), and Silhouette Ettecentral™ (Internet-based recording framework) as its key components. This advanced device enables accurate measurement of wound volume, depth, and circumference. Woundvision [53] has pioneered a multimodal imaging method known as Scout. It enables limited 3D measurements of wound images. Wound visualization analysis based on computer vision enhances the accuracy of clinical diagnosis, reduces the economic burden, and improves the quality and efficiency of medical diagnosis.

### 3.2 Intelligent Transportation

Intelligent Transportation System (ITS) is integral to intelligent city technology, which aims to promote traffic safety, mobility, and sustainable development of transportation to increase productivity [54, 55]. With various application forms, ITS includes highway cooperative maneuver, road safety information sharing, traffic signal optimization, automatic driving, etc. [56, 57] Different deep neural network models are employed for various tasks. For instance, the deep belief network has shown effectiveness in face recognition [58]. Stacked autoencoder framework have proven to be effective in tasks like object detection [59]. The deep learning method based on YOLO (You Only Look Once) has been used for target detection tasks [60]. In the field of ITS, variants of CNN networks are widely employed in cv research. Among these, RNNs(Recurrent Neural Networks) are particularly powerful deep learning methods capable of modeling sequences of inputs and outputs. RNNs have found broad applications in tasks such as lane detection [61, 62]. One notable variant of RNN is the LSTM network. It can excels in sequence prediction tasks by capturing order dependence. Convolutional LSTM network has been used for video anomaly detection and autonomous vehicle application [63, 64]. CV technology has a pioneering potential in transportation systems. With the further development of DL, it will further improve the intelligent transportation system and become the leading research in the future.

### 3.3 Structural Deformation and Damage Monitoring

- Structural Deformation

Cv-based structural deformation monitoring technology has been extensively utilized in the field of structural health monitoring (SHM) Furthermore, numerous laboratories and sites have established a foundational framework that aids in the development of cv-based structural displacement measurement systems capable of sustained and reliable monitoring in real-world conditions [65]. Recently, the primary objective of SHM has been to capture the real-time behavior of structures using sensors and subsequently evaluate their performance. However, the practical implementation of this technology in engineering applications has been hindered by the demanding and costly requirements associated with sensor network maintenance and data collection systems [66, 67]. With the progress of computational advancements, optical sensing capabilities, and image analysis techniques,, information on structural strain, inclination angle, and displacement can be obtained by image analysis on the computer through various algorithms. Moreover, crucial dynamic properties such as oscillation patterns, frequencies, rates of change, and damping coefficients can be obtained through subsequent analysis. The procedure of monitoring structural deformations, employing computer-based methodologies, can be outlined as follows. (1) Align artificial or natural targets using camera and lens configurations to capture visual data. (2) Conduct camera calibration. (3) Extract distinctive attributes from the initial image frame and trace these characteristics throughout succeeding frames. (4) Compute structural displacement [65].

Currently, cv-based monitoring techniques have been utilized in various aspects of SHM, encompassing identification of loose bolts [68], quantification of disaster effects [69], cable stress monitoring [70], modal frequency monitoring [71], two-dimensional [72] and three-dimensional [73] structural deformation measurement, etc.

- Structural Damage

In the realm of visual perception, fractures are usually recognized as sudden shifts in pixel intensity, frequently manifesting as slender, shadowy lines on solid surfaces. Conventional techniques in image analysis, including edge detection, morphological operations, model-based approaches, and more, can be employed for the identification of cracks in images. However, recently, models based on CNN architecture have shown better detection quality and results than traditional processing methods [74]. Ali [74] et al. have conducted an extensive analysis and synthesis of CNN-enabled fracture detection models for infrastructure systems. The study encompasses image pre-processing techniques, software utilities, hardware components, datasets etc. The focus lies in the categorization and partitioning of fracture images using CNN frameworks, as well as the exploration of fracture detection in structural contexts. Li [75] et al. put forward a crack detection method based on GoogLeNet's VGG-16 network. In this approach, a fusion of VGG-16 and RNN is employed to differentiate between mild and severe fractures. Kortman [76] et al. also investigated the limitations of methods for detecting road deterioration that satisfy the prerequisites of autonomous driving systems. They examined the structure of environment perception systems and the development of current methods for road damage detection.

### 3.4 Security Monitoring

As the number of  monitoring devices in urban areas continues to rise, an overwhelming amount of  visual data is being recorded constantly. Manual monitoring alone is physically incapable of analyzing and comprehending the vast content within these videos. Consequently, the utilization of cv-based recognition methods for biological feature recognition has reached a mature stage, attracting interest from various industries. These features include facial features and more subtle features such as iris, palmprint, voiceprint, and even gait of human eyes [77]. Initially, techniques primarily relied on handcrafted spatio-temporal characteristics and conventiona image processing methods. However, in recent years, there has been a shift towards employing more sophisticated approaches, such as DL, for tasks such asvideo data classification and clustering, as well as anomaly detection, etc. Zaheer [78] et al. reviewed the methods developed for video anomaly detection from 2015 to 2018, and classified them according to the network structure classification and the data sets used. It was found that in video surveillance camera anomaly detection, the DL-based method has demonstrated exceptional performance even in challenging environmental conditions. The utilization of deep neural networks with hierarchical feature representation learning surpasses the manual feature extraction techniques employed in traditional architectures. Additionally, the disclosure of crime data sets such as USCD, UMN, and UCF further promotes the efficiency of video stream anomaly detection [79]. Although the communication between machines and organisms fails to break through mechanical means, and the input means of various information and data need to be optimized, machines can detect whether people in images are living through visual recognition, biometric detection, and other technologies. In addition, they can identify the people's information in images with the help of background data comparison and even recognize the trajectory and posture of people in images, which undoubtedly provides new ideas and methods for urban security management.

## 4. Challenges and Prospects

Deep learning and the new round of artificial intelligence development have promoted the development of computer vision, and its technology application has been continuously penetrated into all walks of life. As a subfield of artificial intelligence, cv necessitates continuous development and enhancement alongside advancements in AI software and hardware, while this development boundary cannot be completely predicted at present. Considering the recent advancements in cv and artificial intelligence technology, this paper puts forward some valuable research directions [80].

### 4.1 Vision Research Under New Imaging Conditions

The new imaging technology, represented by computational photography, enables researchers to recover the essential information of the target scene from the reconstructed high-dimensional and high-resolution optical signals, including geometry, material, motion, and interaction, etc., which solves the missing information from the 2D scene to the 3D scene in computer vision research at present. Moreover, it enables machines to have a more comprehensive perception and understanding of physical space and the objective world. In recent years, new computational imaging devices have emerged constantly, such as light field cameras, event cameras, depth cameras, infrared cameras, TOF cameras, high-speed cameras, polarization cameras, etc. These cameras with a wide range of applications have advantages that traditional cameras do not have in some aspects. One exemplary specialized solution is the light field camera, which excels at accurately focusing and capturing clear images even in scenarios involving high-speed image motion and low light intensity. When the event camera detects motion, a very high refresh rate will be rendered on a per-pixel basis. The image data produced by these cameras are different from the traditional images, which are sampled from different parts of the light field in space. The research on visual theory and algorithms under these images will be a new direction in the future.

### 4.2 Research on Bio-Inspired Computer Vision

Computer vision as a subject with strong application has made great achievements in recent decades and has been applied to many fields, the ability of computer vision systems is far from the ability of human beings to complete similar tasks for complex problems. The biological vision system is the most powerful and perfect visual system known to human beings, whose structural characteristics and operation mechanisms have enlightening significance for the computer vision model. Bio-inspired computer vision studies how to introduce the structure, function, and mechanism of human brain visual pathway into the modeling and learning of computer vision, so as to solve the current difficult problems in computer vision research. From the aspect of imitating biology, there have been many successful cases of exploring computer vision inspired by biology. For example, the Gabor filter formally simulates the information coding mode of cells in the primary visual cortex, a classic case in cv.

## 5. Conclusions

Computer vision is a rapidly advancing and extensively utilized field within artificial intelligence. Serving as the "eyes" of AI, it enables the acquisition and analysis of vast amounts of visual information across diverse domains. Apart from reviewing the development history of cv, typical architectures, and algorithms for specific tasks, this paper also summarizes research progress and application in various fields, which finally concludes the current challenges and important research issues in computer vision technology in the future. The author believes that with the changing algorithms, upgrading hardware computing power, and the data explosion, the application of computer vision will have more potential.

## References

[1] Lu, H. T. & Zhang, Q. C. (2016). Applications of deep convolutional neural network in computer vision. Journal of Data Acquisition and Processing, 31(1), 1-17.

[2] Pan, Y. (2016). Heading toward Artificial intelligence 2.0. Engineering, 2(4), 409-413.

[3] Litjens, G., Kooi, T., Bejnordi, E. B., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42.

[4] Ping, J. N., Zhang, C. P., Li, D. Y. et al. (2020). “Digital space government” research program—Government reforming caused by the fourth industrial revolution. Journal of Public Management, 17(1), 1-17.

[5] David, M., Shimon, U., A. T P. (2010). Vision: A computational investigation into the human representation and processing of visual information. The MIT Press.

[6] Wang, W. Y., Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. Journal of Database Management, 30(1), 61-79.

[7] Bengio, Y. (2009). Learning deep architectures for AI. Boston: Now Publishers Inc.

[8] Keith, M., Javan, C. (2022). A review of synthetic image data and its use in computer vision. Journal of Imaging, 8(11).

[9] Liu, Y. J. (2023). Advancement and prospect—Market prospect of computer vision security in 2023. China Security & Protection, (Z1), 60-64.

[10] William, S. (2022). Guest editorial: Special issue on computer vision from 2D to 3D. International Journal of Computer Vision, 131(2), 405-405.

[11] Matteo, P. B., T. M. (2021). Computer vision for 3D perception and applications. Sensors, 21(12), 3944-3944.

[12] Lecun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[13] (2010). CUDA by example: an introduction to general-purpose GPU programming. Scitech Book News, 34(4).

[14] Deng, J., Dong, W., Socher, R. et al. (2009). ImageNet: A large-scale hierarchical image database. 248-255.

[15] Russakovsky, O., Deng, J., Su, H. et al. (2014). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.

[16] Daisuke, K. & Shumpei, I. (2019). Machine learning approaches for pathologic diagnosis. Virchows Archiv: an international journal of pathology, 475(2):131-138.

[17] Cui, J., Zhang, J., Sun, G. et al. (2019). Extraction and research of crop feature points based on computer vision. Sensors, 19(11), 2553-2553.

[18] Wang, J. K. & Song, X. J. (2022). A survey on computer vision application. Computer Era, (10), 1-4+8.

[19] Athanasios V,Nikolaos D,Anastasios D, et al. (2018). Deep Learning for Computer Vision: A Brief Review. Computational Intelligence and Neuroscience, 7068349.

[20] Krizhevsky, A., Sutskever, I. & Hinton, E. G. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[21] Pei, Y., Huang, Y., Zou, Q., Zhang, X. et al. (2021). Effects of image degradation and degradation removal to CNN-based image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4), 1239-1253.

[22] CătălinaLucia, C., Răzvan, C. U. & Daniel, A. S. (2023). Evolutionary image registration: A review. Sensors, 23(2), 967-967.

[23] Gyunyeop, K. & Sangwoo, K. (2022). Effective transfer learning with label-based discriminative feature learning. Sensors, 22(5), 2025-2025.

[24] Fang, F. & Bao, S. (2022). FragmGAN: Generative adversarial nets for fragmentary data imputation and prediction.

[25] Rting, S. N., Doyle, A., Hilten, A. V. et al. (2020). A survey of crowdsourcing in medical image analysis. Human Computation, (7), 1-26.

[26] 2019 Index IEEE transactions on pattern analysis and machine intelligence. 41. (2020). IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(1).

[27] Romain, B., M. F. D. G. (2012). Simulating spiking neural networks on GPU. Network (Bristol, England), 23(4), 167-182.

[28] Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. CoRR, 9351, 234-241.

[29] Ekhmd, A., Dufort, P., Forsberg, D. et al. (2014). Medical image processing on the GPU— Past, present and future. Medical Image Analysis, 17(8), 1073-1094.

[30] Qin, L. & Gao, W. (2009). Scene image categorization based on content correlation. Journal of Computer Research and Development, 46(07), 1198-1205.

[31] Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks.Communications of the ACM, 60(6): 84-90.

[32] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.

[33] Szegedy, C., Wei, L., Jia, Y. et al. (2014). Going deeper with convolutions. CoRR, abs/1409. 4842.

[34] He, K., Zhang, X., Ren, S. et al. (2015). Deep residual learning for image recognition. CoRR, abs/1512.03385.

[35] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91-110.

[36] Dalai, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings of Computer Vision and Pattern Recognition(CVPR). San Diego, USA: IEEE, 1: 886-893.

[37] Girshick B R,Donahue J,Darrell T, et al. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, abs/1311.2524.

[38] Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 1440-1448.

[39] He, K., Zhang, X., Ren, S. et al. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 1904-1916.

[40] He, K., Gkioxari, G., Dollar, P. & Girshick, R. (2020). Mask R-CNN. IEEE Transactions on Pattern Analysis & Machine Intelligence, 42(2), 386-397.

[41] Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 234-241.

[42] Chen, L. C., Papandreou, G., Kokkinos, I. et al. (2016). Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2016-06-07)[2022-01-20]. Retrieved from https://arxiv.org/abs/1412.7062.

[43] Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481-2495.

[44] Yabo, F., Yang, L., Tonghe, W. et al. (2021). A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica, 85, 107-122.

[45] Abhishek, C. & Divakar, S. (2013). Comparative study of different technique for medical image segmentation: A survey. International Journal of Computers & Technology, 11(1), 2169-2174.

[46] Zhou, Z., Sodha, V., Siddiquee, R. M. M. et al. (2019). Models genesis: Generic auto-didactic models for 3D medical image analysis. CoRR, abs/1908.06912.

[47] Donahue, J., Jia, Y., Vinyals, O. et al. (2013). DeCAF: A deep convolutional activation feature for generic visual recognition. CoRR, abs/1310.1531.

[48] M. D. A, Chuanbo, W., Behrouz, R. et al. (2021). Image-based artificial intelligence in wound assessment: A systematic review. Advances in Wound Care.

[49] Lei, W., C. P. P., Emmanuel, A. et al. (2017). Area determination of diabetic foot ulcer images using a cascaded two-stage SVM-based classification. IEEE Transactions on Bio-medical Engineering, 64(9), 2098-2109.

[50] Bo, S. & Ahmet, S. (2012). Automated wound identification system based on image segmentation and artificial neural networks. 2012 IEEE International Conference on Bioinformatics and Biomedicine.

[51] L. E. A., Anagha, K., E. T. B. et al. (2016). The reliability of a novel mobile 3-dimensional wound measurement device. Wounds: a compendium of clinical research and practice, 28(11), 379-386.

[52] Mark, N., Christine, M. & Aranzmedical. (2023). Evidence-based wound surveillance [EB/OL]. (2014-06-14)[2023-08-10]. Retrieved from https://www.aranzmedical.com/wp-content/uploads/2014/06/14/woundmanagement- white-paper/website-full-white-paper.pdf

[53] Scout, WoundVision. The WoundVision Scout [EB/OL], (2020-02-17)[2023-08-10]. Retrieved from https://www.woundsource.com/product/ woundvision-scout.

[54] Esma, D. & Murat, D. (2023). Computer vision applications in intelligent transportation systems: A survey. Sensors, 23(6), 2938.

[55] Lin, Y., Wang, P. & Ma, M. (2017). Intelligent transportation system(ITS): Concept, challenge and opportunity. 2017 IEEE 3rd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS). IEEE, 2017.

[56] Porter, M. C. (2021). Towards safe and equitable intelligent transportation systems: Leveraging stochastic control theory in attack detection. University of Michigan.

[57] Wang, Y., Zhang, D., Liu, Y. et al. (2018). Enhancing transportation systems via deep learning: A survey. Transportation Research Part C, 99:144-163.

[58] Babu, K., Kumar, C. & Kannaiyaraju, C. (2022). Face recognition system using deep belief network and particle swarm optimization. Intelligent Automation & Soft Computing, 33(1), 317-329.

[59] Maria, J., Amaro, J., Falcao, G. et al. (2016). Stacked auto-encoders using low-power accelerated architectures for object recognition in autonomous systems. Neural Processing Letters, 43(2), 445-458.

[60] Jessica, F., M. J. C., Vanessa, F. et al. (2021). Robust real-time traffic surveillance with deep learning. Computational Intelligence and Neuroscience.

[61] Rateke, T. & Wangenheim, V. A. (2020). Passive vision road obstacle detection: A literature mapping. International Journal of Computers and Applications, 44(4), 1-20.

[62] Raza, A., Huang, J. C., Abu, S. M. T. et al. (2022). Structural crack detection using deep convolutional neural networks. Automation in Construction, 133.

[63] Chen, S., Zhang, S., Shang, J. et al. (2019). Brain-inspired cognitive model with attention for self-driving cars. IEEE Transactions on Cognitive and Developmental Systems, 11(1), 13-25.

[64] L. J., Mei, X., Prokhorov, V. D. et al. (2017). Deep neural network for structural prediction and lane detection in traffic scene. IEEE Trans. Neural Netw. Learning Syst., 28(3), 690-703.

[65] Yizhou, Z., Weimin, C., Tao, J. et al. (2022). A review of computer vision-based structural deformation monitoring in field environments. Sensors, 22(10), 3789.

[66] Feng, D., Feng, Q. M., Ozer, E. et al. (2015). A vision-based sensor for non-contact structural displacement measurement. Sensors, 15(7), 16557.

[67] Feng, D. & Feng, Q. M. (2018). Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection–A review. Engineering Structures, 156:105-117.

[68] Ramana, L., Choi, W. & Cha, Y. (2019). Fully automated vision-based loosened bolt detection using the viola—Jones algorithm. Structural Health Monitoring, 18(2), 422-434.

[69] Engineering—Civil engineering: Investigators from State University of New York release new data on civil engineering (image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization). Journal of Engineering, 2019.

[70] Kim, S. & Kim, N. (2013). Dynamic characteristics of suspension bridge hanger cables using digital image processing. NDT and E International, 59, 25-33.

[71] Dong, C., Ye, X. & Jin, T. (2018). Identification of structural dynamic characteristics based on machine vision technology. Measurement, 126,405-416.

[72] Park, J., Lee, J., Jung, H. et al. (2010). Vision-based displacement measurement method for high-rise building structures using partitioning approach. NDT and E International, 43(7), 642-647.

[73] XiaoWei, Y., Tao, J., Peng, A. et al. (2021). Computer vision-based monitoring of the 3D structural deformation of an ancient structure induced by shield tunneling construction. Structural Control and Health Monitoring, 28(4).

[74] Raza, A., Huang, J. C., Abu, S. M. T. et al. (2022). Structural crack detection using deep convolutional neural networks. Automation in Construction, 133.

[75] Li, S. & Zhao, X. (2018). Convolutional neural networks-based crack detection for real concrete surface. Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring.

[76] Data from Leuphana University Luneburg provide new insights into mathematics (watch out, pothole. (2022). Featuring Road Damage Detection In an End-to-end System for Autonomous Driving). Network Daily News.

[77] Chen, Z. H. & Wang, M. X. (2021). Application of computer vision in intelligent security. Telecommunications Science, 37(08), 142-147.

[78] Zaheer, M. Z., Lee, J. H., Lee, S. et al. (2019). A brief survey on contemporary methods for anomaly detection in videos. 2019 International Conference on Information and Communication Technology Convergence (ICTC).

[79] Thakur, D. & Biswas, S. (2020). Smartphone based human activity monitoring and recognition using ML and DL: a comprehensive survey. Journal of Ambient Intelligence and Humanized Computing (prepublish).

[80] Dark Blue College. (2020). Important research issues on the prospect of the computer vision. Retrieved from 2023-08-15, https://zhuanlan.zhihu.com/p/26942231 5.