# CNN-Based AI-Powered Gesture Recognition with Digital Twinning for Automotive Applications- A Survey

*Neethu P S[1*], S K Manju Bhargavi[2]*

[1]Postdoc Researcher, Lincoln University College, Malaysia; [*]Christ University Bangalore; [2]Jain Deemed to be University, Bangalore

ps.neethu@gmail.com

---

**Abstract:** Hand gesture recognition (HGR) is emerging as a transformational technique in human–machine interaction (HMI), particularly in the automobile sector. Convolutional Neural Networks (CNNs), acclaimed for their efficacy in image processing, have markedly enhanced the accuracy and real-time performance of gesture detection. The emergence of Digital Twins (DTs) has facilitated precise system modeling and real-time feedback, expanding prospects for integration and performance enhancement. This review emphasizes recent advancements in CNN-based gesture detection, the constraints posed by temporal dynamics and background noise, and the increasing significance of AI-driven digital twinning. The primary objective is to enhance vehicle interactivity, safety, and personalization through the integration of DT frameworks with CNN-based gesture control systems.
**Keywords:** CNN; Digital twinning; automotive application; hand gestures

---

## Introduction

In recent years, the automotive industry has experienced a significant transition towards advanced and safer human–machine interface (HMI) technology, with gesture recognition emerging as a pivotal approach for intuitive, touchless engagement. Traditional buttons and touchscreens are progressively being substituted or augmented with gesture-based controls that improve usability and minimize driver distraction. Gesture recognition enables drivers to effortlessly control in-car features, including infotainment, navigation, and temperature systems, without diverting their focus from the road or relinquishing their grip on the steering wheel. Achieving reliable gesture detection in a vehicle's dynamic environment remains challenging due to factors such as variable lighting conditions, occlusions, limited cabin space, and individual driver variations.

Convolutional Neural Networks (CNNs) have transformed computer vision and are extensively utilized in gesture detection due to their capacity to autonomously extract spatial information from unprocessed input data. In contrast to previous methods that depended on manually designed features, contemporary CNN-based systems are capable of learning hierarchical representations from optical, depth, radar, or thermal data. The transition from conventional 2D-CNNs to advanced designs like 3D-CNNs, CNN-LSTM hybrids, and CNN-TCN models has facilitated the effective capture of both spatial and temporal dimensions of gestures, rendering them especially appropriate for in-cabin applications. In addition to vision-based techniques, radar-based gesture recognition has gained prominence for its dependability in low-light or privacy-sensitive environments, with performance enhanced through multimodal sensor fusion approaches.

In conjunction with these improvements, digital twinning has arisen as a revolutionary notion for the real-time simulation and optimization of physical systems. A digital twin functions as a virtual representation of a physical environment—such as a vehicle's cabin—constantly refreshed with real-time data. In automobile gesture recognition, digital twins provide a secure, adaptable, and economical framework for the development, testing, and enhancement of AI models across various scenarios, including infrequent or perilous conditions that are difficult to reproduce in real life. The integration of AI-driven CNN models with digital twins facilitates adaptive learning, predictive behavior analysis, and the extensive implementation of gesture recognition technologies in autonomous and semi-autonomous cars. This research offers an extensive analysis of CNN-based gesture detection systems included within digital twin frameworks for automotive applications. It examines recent advancements in deep learning models, multimodal sensing techniques, embedded system deployment, and simulation-based assessment. This paper analyzes cutting-edge technologies from 2020 to 2025 to elucidate the possibilities and limitations of AI-driven gesture recognition in influencing the future of human-vehicle interaction.

## Literature Review

Recent advancements in gesture recognition systems have been significantly enhanced by the incorporation of deep learning, especially through the utilization of convolutional neural networks (CNNs) tailored for real-time applications in intelligent transportation and human-machine interface (HMI). Convolutional Neural Networks (CNNs) offer a robust framework for the automatic extraction of spatial and temporal characteristics from many input modalities, such as video, radar, and thermal sensors. Initial implementations of 2D-CNNs, primarily designed for static gesture detection, have progressed to 3D-CNNs and hybrid architectures adept at accurately modeling spatiotemporal dynamics essential for deciphering intricate movements in confined settings like automobile cabins. Molchanov et al. pioneered 3D-CNN architectures for gesture detection, influencing later research such as DriverMHG by Köpüklü et al. (2020), which offered a multimodal dataset for micro hand gesture classification with MobileNetV2. Zhang et al. (2021) introduced a dual-stream CNN-LSTM framework for precise radar-based gesture detection utilizing FMCW radar data, attaining great accuracy in identifying modest in-cabin movements under diverse lighting situations.

As practical deployment gains prominence, research has transitioned towards lightweight CNN architectures and hardware-efficient models. Scherer et al. (2020) presented TinyRadarNN, a framework that combines spatial and temporal CNNs for short-range radar gesture detection, markedly decreasing processing requirements while maintaining accuracy. Fekry et al. (2024) created a quantized CNN hardware design that enables real-time gesture classification on low-power devices. These developments meet the automotive industry's requirement for energy-efficient, low-latency, real-time gesture control systems in automobiles. Furthermore, Gomaa et al. (2023) investigated personalized in-cabin gesture recognition employing time-of-flight cameras in conjunction with CNN models, improving adaptability by customizing detection systems to unique user behaviors.

The growing implementation of digital twin technology signifies a notable and supplementary progression in this domain. Digital twins serve as real-time virtual representations of physical systems, facilitating simulation, monitoring, and the training of machine learning models in various synthetic settings. Wang et al. (2022) and Chen et al. (2020) utilized digital twin frameworks for HMI systems, employing synthetic settings to train CNNs such 3D ResNet and VGG, which improved data diversity and strengthened the robustness of

gesture recognition models. This methodology facilitates fault injection testing, rare-event simulation, and iterative learning procedures that are difficult to accomplish with physical prototypes. Liu et al. (2021) and Hazra et al. (2022) integrated radar point clouds and graph-based learning into digital twin frameworks for long-range gesture recognition, thereby enabling cross-modal learning and enhancing domain adaption.

The research emphasizes the essential function of multimodal fusion and spatiotemporal modeling in enhancing recognition robustness. Holzbock et al. (2022, 2023) presented spatiotemporal multilayer perceptrons and keypoint-CNN fusion techniques to amalgamate radar keypoints with video inputs. Zhu et al. (2023) established a dual-stream fusion approach that utilizes deformable convolutional features from radar data within a CNN-TCN architecture, facilitating dependable hand motion identification in noisy environments. Ma et al. (2022) introduced RGTNet, which integrates transformer networks with CNNs to effectively capture temporal attention patterns in radar sequences, enhancing recognition accuracy. Hybrid models frequently demonstrate performance over 97% on automotive datasets, highlighting the efficacy of deep feature fusion in sensor-dense in-vehicle contexts. A primary emphasis in contemporary research is on attaining real-time performance and deployment efficacy. In addition to TinyRadarNN, many models—such as TRANS-CNN by Kehelella et al. (2022) and the edge-based real-time framework by Sun et al. (2020)—illustrate that CNN architectures can be efficiently utilized in limited edge environments without dependence on cloud resources. Network designs are progressively using hardware-aware methodologies such as quantization, pruning, and binarization. These advancements have resulted in systems that provide swift inference with minimum processing burden, an essential necessity for safety-critical automotive applications.

There is an increasing interest in the interpretation of contextual gestures, especially for traffic management and infotainment interactions. Wiederer et al. (2023) investigated graph-based gesture recognition frameworks for autonomous vehicle responses to traffic officer signals, whereas Eshi and Jilani (2023) utilized CNN-KNN models for the management of entertainment systems. Furthermore, new sensing modalities, including thermal imaging and wireless communications, have been explored to augment vision and radar. Research

undertaken in 2024 on thermal video-based gesture recognition and wireless sensing in digital twin settings enhances the usability and versatility of CNN-driven gesture recognition systems.

**Gesture Recognition in Automotive Systems**

Gesture recognition is a type of human-computer interaction (HCI) that interprets movements of the hands, arms, or body as input commands. In automobile environments, it facilitates touchless operation of services including entertainment, navigation, climate control, and safety systems. In contrast to voice commands or touch-based interfaces, gestures provide a less distracting, nonverbal, and minimally intrusive way of engagement. Implementing gesture recognition in vehicles is exceedingly complex, as the system must maintain reliability despite varying illumination conditions, different driver behaviors, occlusions, and motion blur.

**Convolutional Neural Networks (CNNs) for Gesture Recognition**

Convolutional Neural Networks (CNNs) are deep learning architectures that extract spatial feature hierarchies from input data using convolutional layers. They are exceptionally proficient at processing optical and radar frames, facilitating precise gesture classification. Early methods utilized 2D-CNNs for the recognition of static, image-based movements, whereas contemporary developments utilize 3D-CNNs to concurrently record spatial and temporal information from video sequences. Hybrid models, including CNN-LSTM and CNN-Transformer networks, augment sequential modeling skills. Convolutional Neural Networks (CNNs) are especially adept in automobile gesture detection as they can directly acquire resilient and invariant characteristics from unprocessed sensor data.

**Digital Twin Technology**

A digital twin is a dynamic, real-time virtual representation of a physical system that reflects its status and behavior by continuous integration of sensor data. In gesture recognition, a digital copy of the car interior can replicate hand gestures, lighting conditions, seating arrangements, and sensor locations. This enables AI models to be taught and assessed in various virtual environments without need on physical prototypes. The digital twin structure facilitates synthetic data production, expedited experimentation, system validation, and

adaptive learning, while additionally functioning as a secure testbed for infrequent or perilous driving conditions.

Vision-based gesture recognition has emerged as an essential element of in-vehicle human–machine interfaces (HMI), providing an intuitive, hands-free mode of engagement. Convolutional Neural Networks (CNNs) underpin vision-based gesture analysis due to its capacity to autonomously extract significant information from camera inputs. This section provides a classified analysis of notable current studies (2020–2025), highlighting the function of vision sensors in CNN-based vehicle gesture recognition and their incorporation with digital twin technologies.

**Vision-Based Gesture Recognition Using CNN Architectures**

Conventional 2D-CNNs, originally designed for image classification, have been effectively modified for static gesture recognition utilizing RGB image frames. Initial research depended on extensive gesture datasets, including NVIDIA Dynamic Hand Gesture and EgoGesture. Köpüklü et al. (2020) presented DriverMHG, a multimodal dataset comprising RGB, depth, and infrared data acquired in authentic driving environments. Their research indicated that CNN models such as MobileNetV2 and ResNet may attain over 94% accuracy when trained on spatiotemporal input sequences. To enhance the representation of motion, researchers transitioned to 3D-CNNs, which expand convolution beyond spatial dimensions (width and height) to encompass the temporal domain (frames). Holzbock et al. (2023) introduced a hybrid 3D-CNN architecture that amalgamates spatial and temporal characteristics from RGB video data. Their approach markedly boosted gesture detection precision in low-motion applications, such as volume adjustment and menu navigation, by training with keypoint-enriched video streams.

Another innovation is the use of hierarchical CNN architectures, wherein an initial CNN detects the presence of a gesture, followed by a subsequent CNN that classifies the exact gesture. This stratified methodology enhances efficiency and diminishes false positives, particularly in congested in-vehicle settings. Hierarchical CNNs have been effectively utilized

in real-time applications employing lightweight yet robust architectures such as ResNeXt-101 and MobileNetV3.

**Use of Depth and IR Vision Sensors**

While RGB cameras exhibit commendable performance in optimal lighting, their precision diminishes markedly in low-light or obstructed environments. To mitigate this issue, depth cameras and infrared sensors have been incorporated into car gesture recognition systems. Time-of-flight (ToF) and structured light sensors produce depth maps that quantify pixel-specific distances, hence improving spatial awareness. Gomaa et al. (2023) introduced a personalized in-cabin gesture recognition framework utilizing a ToF camera and a CNN model trained on user-specific gesture variations, resulting in significant enhancements in accuracy and adaptability among diverse users and seating positions. Likewise, stereo depth data has been utilized in alternative systems to improve hand segmentation and mitigate background noise, allowing CNNs to focus on essential gesture areas even in difficult lighting conditions. Infrared (IR) vision, especially thermal imaging, has been investigated as an alternative modality. Although generally possessing poorer resolution, thermal pictures provide privacy-preserving gesture recognition and robustness in low-light conditions. Numerous lightweight CNN variations, such as quantized and binary networks, have been modified to identify specified control motions for infotainment systems utilizing thermal video inputs.

**CNN + LSTM and Transformer Networks for Spatio-Temporal Modeling**

Although CNNs excel at capturing spatial characteristics, gestures are fundamentally temporal. Hybrid architectures have been created that integrate CNNs with sequence modeling units, such LSTMs (Long Short-Term Memory) and Transformers. In these models, CNNs extract features at the frame level, which are further processed by temporal layers to capture motion relationships. Zhang et al. (2021) employed a CNN-LSTM architecture to identify dynamic hand motions from depth video, attaining over 96% classification accuracy and exhibiting significant robustness across diverse contexts. Similarly, Ma et al. (2022) presented RGTNet, which amalgamates CNNs with a Transformer module to represent attention patterns on a frame-by-frame basis. Their methodology, initially developed for radar data, was then modified for RGB-D inputs, demonstrating improved efficacy in dynamic

gesture identification. These hybrid models are especially proficient in differentiating motions with nuanced temporal variations such as “swipe left” vs “scroll up” that may appear similar in static images but diverge over time.

**Integration of Digital Twin for Vision-Based Systems**

Digital twinning facilitates the simulation of vision-based gesture recognition systems across many situations, including differing lighting, hand positions, occlusions, and sensor locations. This technique enables the training and validation of CNNs in controlled virtual environments prior to their use in actual cars. Wang et al. (2022) created a digital twin for a smart car human-machine interface with 3D-ResNet and VGG models. The virtual environment generated synthetic RGB-D sequences that emulated driving motions, enhancing training datasets and augmenting model generalization across various users and environmental situations. Chen et al. (2020) utilized a digital twin framework to train convolutional neural networks in multimodal fusion systems, integrating RGB and depth inputs to replicate various driver actions and environmental changes. A primary benefit of digital twins is their capacity to simulate rare or atypical scenarios—such as abrupt hand motions, unconventional gesture velocities, or actions in low-light environments—that are challenging to get using conventional data gathering methods. This feature improves the resilience and security of CNN-based gesture recognition systems in automotive contexts.

**Limitations and Research Gaps**

Notwithstanding their prevalent application, vision-based CNN gesture recognition algorithms encounter numerous obstacles. A significant disadvantage is their dependence on RGB inputs, rendering them exceedingly sensitive to lighting conditions. Although depth and infrared sensors can alleviate this problem, fusion-based models frequently entail significant computing expenses. Moreover, in-cabin factors such as seating arrangement, attire, or hand obstructions can markedly diminish identification accuracy, and only a limited number of solutions exhibit consistent performance across diverse users and vehicle models without necessitating retraining. A significant impediment is the restricted accessibility of detailed in-cabin gesture records. While datasets such as DriverMHG and NVIDIA Dynamic provide useful benchmarks, numerous systems continue to rely on synthetic augmentation or digital twin

simulations for scalability. Real-time inference continues to be a limitation, especially for deep CNN models operating on embedded platforms with limited CPU resources.

*Table 1.Comparison of Vision-Based CNN Gesture Recognition Models*

| Author(s) | Year | Architecture | Dataset | Sensor Type | Accuracy | Remarks |
|---|---|---|---|---|---|---|
| Köpüklü et al. | 2020 | MobileNetV2 | DriverMHG | RGB, Depth, IR | 94.1% | Multi-modal micro-gesture dataset in vehicle environment |
| Gomaa et al. | 2023 | CNN (customized) | Custom (ToF) | Time-of-Flight (RGB-D) | 93.2% | Personalized in-vehicle gestures with user variations |
| Holzbock et al. | 2023 | 3D-CNN + MLP | Custom | RGB + keypoints | 95.8% | Keypoint fusion improves detection in low-motion cases |
| Zhang et al. | 2021 | CNN + LSTM | Custom RGB-D | Depth (video) | 96.0% | Dynamic gesture modeling with spatial-temporal features |
| Ma et al. (RGTNet) | 2022 | CNN + Transformer | Adapted to RGB | RGB-D (adapted) | 97.6% | Frame-level temporal attention fusion |
| Wang et al. | 2022 | 3D-ResNet + DT sim | Synthetic/Real mix | RGB-D (Digital Twin) | 94.8% | Synthetic data via digital twin simulation |
| Chen et al. | 2020 | VGG + DT Fusion | Digital Twin | RGB-D + IR (Simulated) | 92.5% | Virtual training and validation with synthetic data |
| Zhu et al. | 2023 | CNN-TCN (Deformable Conv) | Custom | RGB-D | 98.6% | Strong performance in occluded and noisy frames |

| Author(s) | Year | Architecture | Dataset | Sensor Type | Accuracy | Remarks |
|---|---|---|---|---|---|---|
| Fekry et al. | 2024 | Quantized CNN | Custom | RGB | 91.5% | Optimized for edge inference in embedded devices |

The comparison analysis Table 1 underscores several significant trends and observations. Hybrid architectures that combine CNNs with sequential learning modules like LSTMs or Transformers surpass independent CNNs in dynamic gesture detection due to their superior temporal modeling abilities. Prominent instances comprise Ma et al.'s RGTNet and Zhang et al.'s CNN-LSTM framework, each attaining accuracies of 96%. Secondly, the selection of sensors is pivotal—multimodal fusion systems (e.g., RGB + Depth or RGB + IR) typically exhibit enhanced performance relative to single-sensor models, particularly in adverse settings such as fluctuating lighting and occlusions. Research, including DriverMHG (Köpüklü et al., 2020) and Zhu et al. (2023), underscores the efficacy of integrating depth maps and heat data into CNN frameworks. Third, digital twins exhibit significant potential for synthetic data generation and system validation, as evidenced by Wang et al. (2022) and Chen et al. (2020), who utilized them to simulate rare events and varied driver behaviors, thereby improving model robustness without the need for extensive physical data collection. Fourth, customization is becoming a significant focus, as Gomaa et al. (2023) illustrate how gesture systems may adjust to individual driving behaviors, hand placements, and reach. Real-time deployment issues are being addressed by techniques like as quantization, with Fekry et al. (2024) demonstrating that CNNs may be optimized for embedded vehicle systems while preserving high accuracy.

## Conclusion

This study examined advancements in vision-based gesture recognition for automotive human-machine interaction, highlighting the significance of Convolutional Neural Networks (CNNs) and their incorporation with hybrid and digital twin methodologies. Convolutional Neural Networks, when integrated with temporal models like Long Short-Term Memory networks and Transformers, exhibit enhanced efficacy in capturing spatiotemporal

dependencies, facilitating precise identification of intricate dynamic gestures. Multimodal sensor fusion utilizing RGB, depth, or infrared significantly improves resilience, especially in the presence of illumination changes and occlusion. Digital twins introduce an additional dimension by producing synthetic data, mimicking various settings, and enabling the validation of gesture systems without solely depending on expensive real-world experiments. Notwithstanding these advancements, considerable obstacles persist. Recognition performance is frequently influenced by driver variability, seating arrangements, and environmental factors. The elevated computing complexity also limits implementation on embedded automotive systems. Moreover, the lack of extensive, consistent in-cabin gesture datasets persists in constraining scalability and benchmarking.

Future research should concentrate on enhancing models via quantization, pruning, and knowledge distillation to facilitate real-time use in vehicles. Customized acknowledgment that adjusts to certain driver behaviors signifies another viable avenue. Ultimately, the expansion of digital twin environments for rare-event modeling and multimodal fusion will be essential for developing reliable, generalizable, and safe in-vehicle gesture recognition systems.

**References**


1. H. S. Sacks, J. N. Fain, "Human epicardial adipose tissue: a review," American Heart Journal, vol. 153, no. 6, pp. 907-917, 2007.
2. J. M. Massaro, U. Hoffman, G. A. Rosito et al., "Pericardial fat, visceral abdominal fat, cardiovascular disease risck factors, and cardiovascular calcification in a community-based sample: the framingham heart study," Circulation, vol. 117, no. 605-613, 2008.
3. R. Sicari, Lombardi, E. Picano, A. Gastaldelli, "Pericardial Rather Than Epicardial Fat is a Cardiometabolic Risk Marker: An MRI vs Echo Study," Journal of the American Society of Echocardiography, vol. 24, no. 10, pp. 1156-1162, 2011.
4. U. Hoffmann, C.L. Schlett, M. Ferencik, M.F. Kriegel, F. Bamberg, B.B. Ghoshhajra, S.B. Joshi, J.T. Nagurney, C.S. Fox, Q.A. Truong, Association of pericardial fat and coronary high-risk lesions as determined by cardiac CT, Atherosclerosis 222 (1) (2012) 129–134.

5. M.O. Al Chekakie, C.C. Welles, R. Metoyer, A. Ibrahim, A.R. Shapira, J. Cytron, P. Santucci, D.J. Wilber, J.G. Akar, Pericardial fat is independently associated with human atrial fibrillation, J. Am. Coll. Cardiol. 56 (10) (2010) 784–788.
6. M. Granér, R. Siren, K. Nyman, J. Lundbom, A. Hakkarainen, M.O. Pentikainen, K. Lauerma, N. Lundbom, M. Adiels, MS Nieminen, M. Taskinen, Cardiac steatosis associates with visceral obesity in nondiabetic obese men, J. Clin. Endocrinol. Metab. 98 (3) (2013) 1189–1197.
7. J. Ouyang, S. Y. Chun, Y. Petibon, A. A. Bonab, N. Alpert and G. El Fakhri, "Bias Atlases for Segmentation-Based PET Attenuation Correction Using PET-CT and MR," in IEEE Transactions on Nuclear Science, vol. 60, no. 5, pp. 3373-3382, Oct. 2013.
8. F. Commandeur et al., "Deep Learning for Quantification of Epicardial and Thoracic Adipose Tissue From Non-Contrast CT," in IEEE Transactions on Medical Imaging, vol. 37, no. 8, pp. 1835-1846, Aug. 2018.
9. V. Zlokolica et al., "3D epicardial fat registration optimization based on structural prior knowledge and subjective-objective correspondence," 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), Belgrade, Serbia, 2015, pp. 1-6.
10. Pednekar, A. N. Bandekar, I. A. Kakadiaris and M. Naghavi, "Automatic Segmentation of Abdominal Fat from CT Data," 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1, Breckenridge, CO, USA, 2005, pp. 308-315.
11. V. Zlokolica et al., "Epicardial fat registration by local adaptive morphology-thresholding based 2D segmentation," Proceedings ELMAR-2014, Zadar, Croatia, 2014, pp. 1-4.
12. A.Kazemi, A. Keshtkar, S. Rashidi, N. Aslanabadi, B. Khodadad and M. Esmaeili, "Automated Segmentation of Cardiac Fats Based on Extraction of Textural Features from Non-Contrast CT Images," 2020 25th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 2020, pp. 1-7.
13. Q. Zhang, J. Zhou, B. Zhang, W. Jia, and E. Wu, "Automatic Epicardial Fat Segmentation and Quantification of CT Scans Using Dual U-Nets With a Morphological Processing Layer," in IEEE Access, vol. 8, pp. 128032-128041, 2020.

14. G. Yu and B. Shao, "Garbage Classification and Detection Based on Improved YOLOv7 Network," 2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA), Beihai, China, 2023, pp. 103-107.
15. S. Harini, M. Suguna, A. T. V. Subramani, and G. H. Krishna, "The Traffic Violation Detection System using YoloV7," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-7.
16. H. Zhao, H. Zhang, and Y. Zhao, "YOLOv7-sea: Object Detection of Maritime UAV Images based on Improved YOLOv7," 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2023, pp. 233-238.
17. Y. Zhang, J. Miao, and C. Liu, "Detection of bolts and nuts of automobile sheet metal parts based on YOLOV7," 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2023, pp. 1648-1651.
18. F. Jia and C. Xu, "Chest X-ray Lesion Detection Based on Improved YOLOv7," 2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI), Taiyuan, China, 2023, pp. 239-243.
19. Y. Liu et al., "Surface Defect Detection Algorithm for Air Filters Based on Improved YOLOv7," 2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Dali, China, 2023, pp. 978-982.
20. X. Qu, Q. Wang, Z. Liu, Y. Gu, Z. Tao and T. Shen, "Improved YOLOv7 Based on Small Target Information Extraction for Road Crack Detection," 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), Jiuzhaigou, China, 2023, pp. 425-430.
21. S. Liu, Y. Wang, Q. Yu, H. Liu, and Z. Peng, "CEAM-YOLOv7: Improved YOLOv7 Based on Channel Expansion and Attention Mechanism for Driver Distraction Behavior Detection," in IEEE Access, vol. 10, pp. 129116-129124, 2022.