Review

Not peer-reviewed version

A Survey of Computer Vision Algorithm Applied in Driverless Robotic Vehicles with Sensing Capability

Lu Chen , Gun Li , Weisi Xie , Jie Tan , Yang Li , Junfeng Pu , Lizhu Chen , Decheng Gan , Weimin Shi *

Posted Date: 15 May 2024

doi: 10.20944/preprints202405.1051.v1

Keywords: computer vision; driverless robotic vehicles; visual SLAM; deep learning; sensing capability

Review

A Survey of Computer Vision Algorithm Applied in Driverless Robotic Vehicles with Sensing Capability

Lu Chen [1], Gun Li [1], Weisi Xie [1], Jie Tan [1], Yang Li [1], Junfeng Pu [1], Lizhu Chen [1], Decheng Gan [3] and Weimin Shi [2,*]

[1] School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 611731, China;lchen@std.uestc.edu.cn (L.C.)

[2] School of Microelectronics and Communication Engineering, Chongqing University,Chongqing 400044, China

[3] School of Electronic Information Engineering, Yangtze Normal University, Chongqing,400044, China

* Correspondence: wmshi@cqu.edu.cn

Abstract: Within environmental perception, automatic navigation and object detection, computer vision is a crucial and demanding field with many applications in modern industries, such as multi-target long-term visual tracking in automated production, defect detection, and driverless robotic vehicles with sensing ability. The performance of computer vision has greatly improved recently thanks to developments in deep learning algorithms and hardware computing capabilities, which has spawned the creation of a large number of related applications. This paper presents the results of a detailed review of over 50 papers published over the course of two decades (1999–2024), with a primary focus on the technical advancement of computer vision. To elucidate the foundational principles, an examination of typical visual algorithms based on traditional correlation filtering was initially conducted. Subsequently, a comprehensive overview of the most recent advancements in deep learning-based computer vision techniques was compiled. Furthermore, a comparative analysis between conventional and novel algorithms was undertaken to discuss the future trends and directions of computer vision. Lastly, the feasibility of employing visual SLAM (Simultaneous Localization and Mapping) algorithms in the context of autonomous vehicles with sensing capability was explored. Furthermore, we explored a thorough and efficient architecture for utilizing these techniques in autonomous robotic vehicles into unmanned green and low-carbon smart factories, underlining novel advances as well as potential prospects for future research.

Keywords: computer vision; driverless robotic vehicles; visual SLAM;deep learning; sensing capability

1. Introduction

In the contemporary digital age, the deployment of computer vision systems as a substitute for traditional, expensive sensors, serving as the primary data source for a multitude of application scenarios, not only reduces costs but also significantly elevates performance ceilings. Furthermore, various systems' intelligence can be enhanced by combining AI algorithms with cognitive thinking patterns.

Visual object detection, a cornerstone task of computer vision, entails the identification of specific visual objects (such as humans, animals, roads, manufactured products, or vehicles) within digital images or video streams. By analyzing the motion patterns, characteristics, and behavioral states of the target, the automated intelligent system can determine the specific actions or inferential outcomes. Similar to this, Visual SLAM (Simultaneous Localization and Mapping) navigation operates by employing image feature collection for indoor mapping and navigation. The purpose of object detection is to advance the development of computer vision models and methods, which can be applied to other domains, such as long- and short-term visual tracking.

In recent years, the integration of vision detection technology in robotic vehicles has positioned them at the forefront of innovation in autonomous systems. These vehicles have been extensively applied across a broad spectrum of technical applications, including localization and navigation, path planning, multitask collaboration, target detection, three-dimensional pose estimation, obstacle detection and avoidance strategies, robotic grasping, automated robotic welding, and security surveillance. This widespread application underscores the pivotal role of robotic vehicles in advancing the capabilities and functionalities of autonomous systems. These applications can be realized through various methods, including vision detection, localization, and tracking. The convergence of computer vision and green energy technologies in multi-driverless systems presents a unique set of challenges and opportunities, encompassing aspects from perception and navigation to energy management and environmental impact. This survey aims to elucidate the potential of computer vision algorithms to revolutionize the interaction of multi-driverless robotic vehicles with their surroundings and contribute to a more sustainable future by examining the latest developments and applications in the field.

On the other hand, the evolution of deep convolutional neural networks coupled with the augmentation of GPU computing capabilities has been instrumental in the accelerated progression of computer vision technology in recent years, primarily due to their synergistic contributions. Presently, the majority of state-of-the-art object detection systems employ deep learning networks as the foundational framework for feature extraction and the classification of images or video streams. Therefore, this paper offers a concentrated review of deep learning algorithms, highlighting their pivotal role in the advancement of computer vision technology.

Computer vision technology has found extensive applications in a myriad of real-world domains, as illustrated in Figure 1, including multi-driverless vehicles, robotic vision, video surveillance, SLAM (Simultaneous Localization and Mapping) navigation, human behavior detection, and automated low-carbon, environmentally friendly production in unmanned factories. These applications have permeated various sectors of modern life, encompassing security, automation, the military, transportation, and medicine, demonstrating the pervasive influence and utility of computer vision technology in contemporary society.

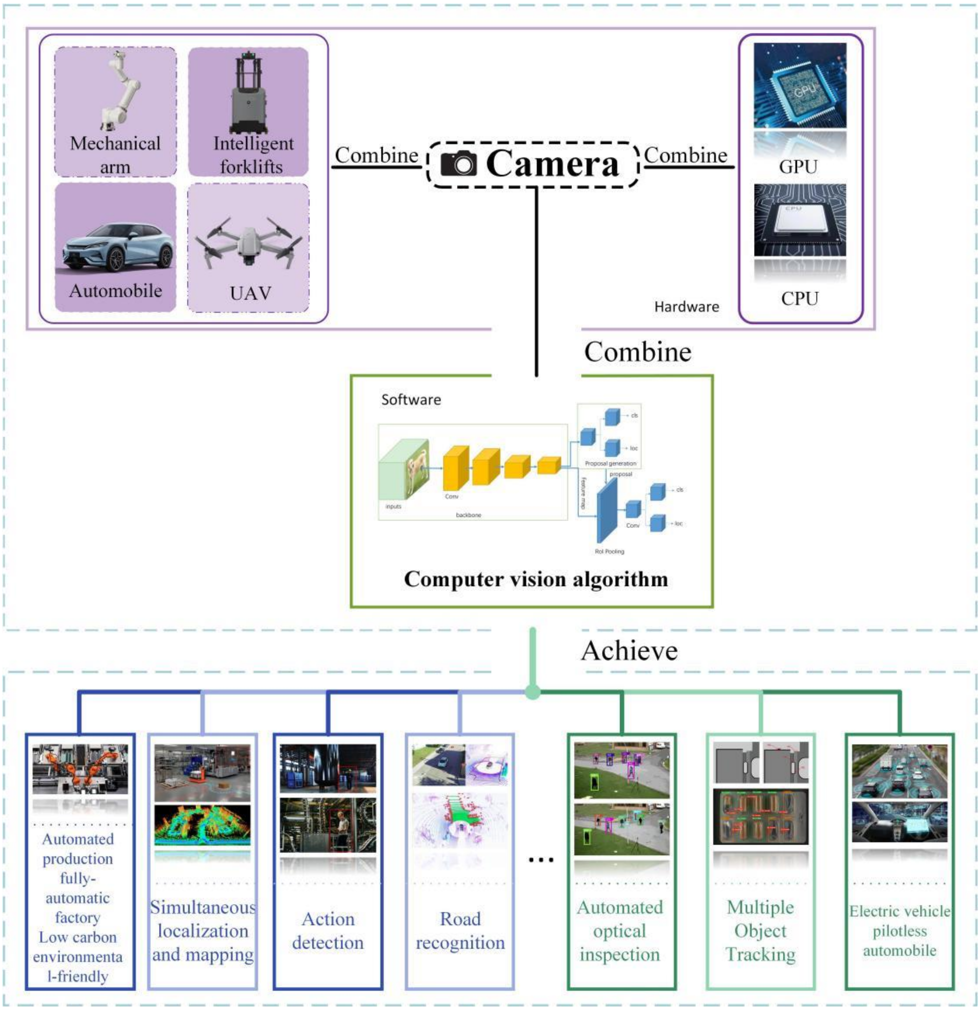


Figure 1. The Applying Scenarios of Computer Vision.

The field of object detection has experienced significant advancements, attributable in part to the establishment of numerous benchmarks such as Caltech [1], KITTI [2], ImageNet [3], PASCAL VOC [4], MS COCO [5], and Open Images V5 [6]. Furthermore, the organizers of the ECCV VisDrone 2018 competition introduced an innovative drone platform-based dataset [7], comprising a comprehensive collection of images and videos. Figure 2 illustrates the ascending trend in the number of papers tagged with "computer vision" over the past decade, highlighting the field's growing prominence and impact within the scientific and academic communities.

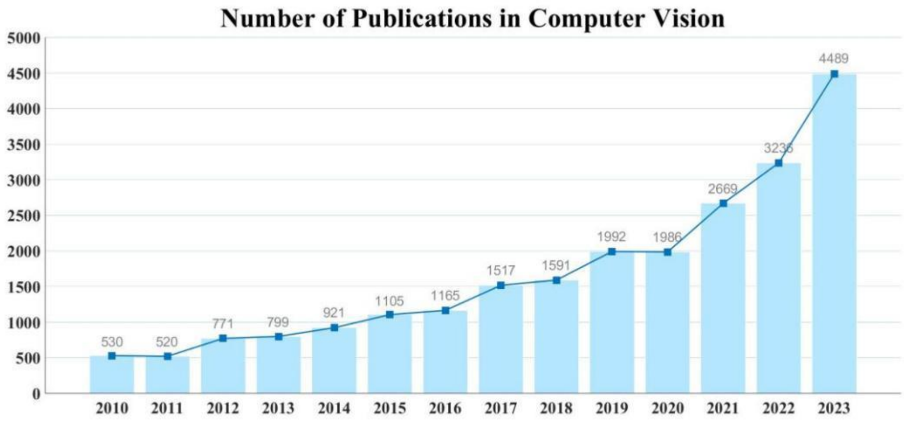


Figure 2. The increasing number of publications in object detection from 2010 to 2023. (Data from IEEE Xplore advanced search: allintitle: "computer vision".).

In 2001, P. Viola and M. Jones introduced the Viola-Jones (VJ) detector [8], which effectively achieved face detection by employing techniques such as feature extraction selection, integral image processing, and multi-stage detection to enhance the performance and speed of the VJ detector. Subsequently, in 2005, N. Dalal and B. Triggs proposed the HOG (Histogram of Oriented Gradients) Detector [9], which implemented scale-invariant feature transformation, enabling effective recognition of objects of different sizes, particularly for human detection. This method has become foundational for many computer vision technologies, providing a robust framework for object detection and recognition.

The DPM (Deformable Part-based Model), developed by P. Felzenszwalb [10] based on HOG, represents a significant milestone in traditional vision detection. It introduced concepts such as mixture models, hard negative mining, and bounding box regression, which continue to exert a profound influence on contemporary computer vision algorithms. DPM primarily employs multiple related filters to replace the process of manually specified screening, functioning as a weakly supervised self-learning method. This detection algorithm achieved championships in the VOC-07, -08, and -09 detection challenges, demonstrating its efficacy and robustness in object detection tasks.

Building on the foundation of computer vision detection, Bolme et al. constructed a novel visual object tracking framework in 2010 using correlation filtering and proposed the MOSSE (Minimum Output Sum of Squared Error) algorithm [11]. MOSSE resolved an optimization problem using the target's grayscale features and a Gaussian expected output function, training a discriminative correlation filter. In 2014, Henriques introduced the KCF (Kernelized Correlation Filters) tracking algorithm [12], providing a complete theoretical derivation of the cyclic nature and offering a method to integrate multi-channel features into the correlation filter framework. This method incorporated multidimensional HOG features into the algorithm, further enhancing its performance. In 2019, Xu et al. proposed the LADCF (Learning Adaptive Discriminative Correlation Filters) algorithm [13], addressing common issues of boundary effects and temporal filter degradation in correlation filter-based algorithms. This tracking algorithm combined adaptive spatial feature selection and temporal consistency constraints, enabling effective representation in low-dimensional manifolds and improving tracking accuracy. In the same year, Huang et al. introduced the ARCF (Aberrance Repressed Correlation Filters) tracking algorithm [14], a long-term target tracking algorithm based on the BACF(Background-Aware Correlation Filters) algorithm. The ARCF incorporated aberration repressed correlation filters in the detection module, restraining the variation rate of response maps generated during the detection phases to suppress aberrations, thereby achieving more stable and accurate target tracking.

The introduction of the CNN (Convolutional Neural Network) model by R. Girshick in 2014 [15] marked a significant milestone in the application of convolutional neural networks for object detection and tracking. Subsequently, in 2015, S. Ren et al. proposed Faster R-CNN [16], which introduced the RPN (Region Proposal Network), further advancing the field. In the same year, R. Joseph et al. presented the YOLO model based on One-stage Detectors [17], significantly enhancing computational speed. R. Joseph subsequently made a series of improvements to the YOLO model, leading to the development of its v2 and v3 editions [18]. Currently, Alexey Bochkovskiy and others have updated the algorithm to YOLOv8, which has been applied in various scenarios [19].

With the maturation of deep learning models, in 2017, Chen B. X. et al. proposed a pedestrian tracking system for unmanned vehicles based on binocular stereo cameras [20], addressing the issue of RGBD cameras being unsuitable for outdoor use. This type of camera calculates depth information through the disparity between two cameras. In 2021, Liu et al. [21] designed a deep learning-based robotic system closed-loop tracking framework, DeepSORT, capable of effectively implementing automatic detection and tracking of new vehicles with sensing capability in complex environments. In the same year, TEED Z. et al. proposed DROID-SLAM [22], which had a significant impact on visual SLAM. This method liberated unmanned vehicles from the constraints of expensive LiDAR by utilizing visual cameras, enabling not only object detection but also localization and navigation. The

DROID-SLAM model, based on deep learning, can effectively extract features from the front end and compute optical flow fields to iteratively update the pose and depth of the current keyframe, while performing BA(Bundle Adjustment) optimization in the backend, greatly enhancing the system's robustness, generalization ability, and performance.A plethora of deep learning-based computer vision technologies applied to multi-vehicle unmanned driving and automated intelligent unmanned factories have effectively improved the efficiency and intelligence of various systems [23].

This article also presents a comprehensive survey of research on computer vision and a roadmap of milestone vision detectors, as illustrated in Figure 3. The authors have summarized and organized the key nodes proposed by the computer vision detection network, categorizing them into traditional methods and deep learning methods based on their chronological development. Post-2014, the classification was further refined into one-stage detectors, two-stage detectors, and visual SLAM, providing an intuitive reflection of the technological evolution. These highlights distinguish the current research landscape from the numerous general object detection reviews that have been published in recent years [24-31].

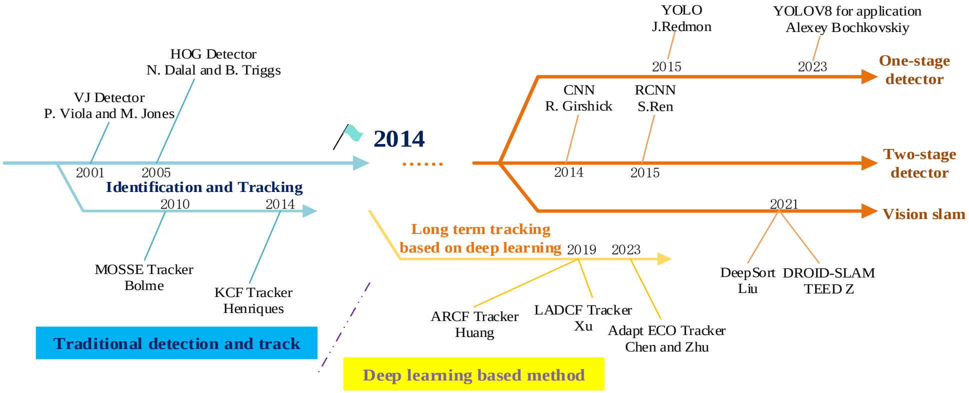


Figure 3. A road map of computer vision with milestone detectors and trackers.

Section one serves as an introduction, encompassing the background, application, research status, and classification of the field. Section two delineates the computer vision system and elucidates the basic principles of traditional correlation filtering algorithms. Section three explores vision algorithms based on deep learning developed in recent years and compares them with traditional algorithms. Additionally, this part offers a summary of detection datasets that we have created and are widely used, along with statistics for each. Section four delves into visual SLAM (Simultaneous Localization and Mapping) algorithms that support computer vision, encompassing long-term visual tracking and the collaboration between multi-driverless robotic vehicles; current challenges and future research directions are also addressed. Section five concludes the article.

2. Review of Traditional Computer Vision Algorithms

This chapter offers an introduction to the Viola-Jones (VJ) detector model, aimed at facilitating an understanding of the basic principles and modular composition of traditional computer vision algorithms. Subsequently, it introduces the composition of the ECO (Efficient Convolution Operators) detection and tracking algorithm and proposes a framework for an adaptive detection and tracking model that integrates ECO [32]. This framework is applied to the visual detection and tracking of targets such as roads, traffic lights, and pedestrians in the context of sensing vehicles, thereby preparing for the execution of various unmanned driving actions.

2.1. The Basic Theory of Computer Vision

In the domain of computer vision-based target location detection, the efficiency of employing effective image feature extraction methods is significantly higher than that of pixel-based calculation approaches. Initially, the image undergoes an integration process as a preliminary step in feature extraction as shown in equation (1):

$$ii(x,y)=\sum_{x'\le x,y'\le y} i(x',y'), \tag{1}$$

Where x,y represent the position information, ii represents the integral image, and i represents the original image. The following recursive formula completes the integral processing of the initial image.

$$\begin{aligned} s(x,y)&=s(x,y-1)+i(x,y)\\ ii(x,y)&=ii(x-1,y)+s(x,y) \end{aligned} \tag{2}$$

The use of rectangular features combined with integral images improves extraction efficiency and flexibility. A weak learning algorithm is employed to design single rectangular features that separate positive and negative examples, identifying the final threshold classification function to reduce the occurrence of misclassification. The weak classifier h(x) is designed as follows:
Where represents the feature, and represents the threshold.

$$h_j(x)=\begin{cases}1 & \text{if } p_j f_j(x)<p_j\theta_j\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

Among them, f_i represents the characteristics， θ is threshold.

Upon completion of the image integration process, a cascading classifier is assembled. The visual detection mechanism adopts the structure of a degenerate decision tree, commonly known as cascading. In this sequence, each subsequent classifier is activated only upon the successful passage of the previous stage; a 'pass' from the first-level classifier prompts the activation of the second-level classifier, and this progression continues accordingly. Any negative classification at a given level halts further evaluation down the cascade. This operational principle is elucidated in Figure 4.

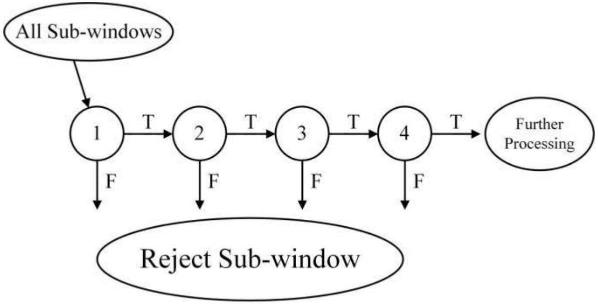


Figure 4. Cascade classifier structure.

The threshold for each stage need to be defined according to the situations of different detection targets, the number of stages in the classifier and the number of features corresponding to each stage. The aim is to balance the number of features and the number of stage in classifier, thus minimizing the expected number of evaluated features.

2.2. VisionVisual Framework Based on ECO for Target Detection and Long-term Tracking

The Efficient Convolution Operators (ECO) algorithm, serving as a foundation for target detection and tracking, has exhibited outstanding performance on tracking datasets including the Visual Object Tracking (VOT) and Object Tracking Benchmark (OTB). By executing convolution operations within the continuous domain, the algorithm facilitates the natural integration of multi-resolution feature mappings. This approach permits the flexible selection of features across various resolutions without the necessity for resampling, thereby enhancing the adaptability and efficiency of the tracking process. The result of the target detection by the algorithm is a continuous function, enabling sub-pixel level positioning and improving tracking accuracy. The ECO algorithm learns the continuous convolutional filter f. The mapping matrix P uses a set of sample feature maps $\{x_j\}_1^M = \{x_1, ..., x_M\}$ and corresponds desired response values $\{y_j\}_1^M = \{y_1, ..., y_M\}$. The position of the target is predicted by the following convolution operation:

$$S_{f,P}\{x\} = \sum_{c=1}^{C} f^c * (\sum_{d=1}^{D} P_{dc} J_d\{x^d\}) \tag{4}$$

In the proposed method, the interpolation operator J_d performs interpolation calculations on the d-th feature channel x^d of the extracted sample image x, transforming the feature map into a continuous domain. Subsequently, a projection matrix P of size $D \times C$ is utilized to map the feature map to a lower-dimensional space. The mapped samples are then convolved with a continuous convolutional filter f to obtain the response map.

The continuous convolutional filter f and the projection matrix P are jointly learned by minimizing an optimization loss function (5),

$$E(f) = \sum_{j=1}^{M} \alpha_j \| S_{f,P}\{x_j\} - y_j \|^2 + \sum_{c=1}^{C} \| w f^c \|^2 + \lambda \| P \|_F^2 \tag{5}$$

where α_j denotes the weight for each training sample. The label function y_j is a two-dimensional Gaussian distribution function centered at the target position in the corresponding sample x_j.The second term is a spatial regularization term, introduced to mitigate the boundary effect issues caused by circular convolution. The last term is the F-norm of the matrix P, serving as a regularization term with the regularization parameter λ.

By employing the Parseval's theorem, the optimization problem (5) can be transformed into a minimization problem in the Fourier domain:

$$E(f) = \sum_{j=1}^{M} \alpha_k \left\| \widehat{S_{f,P}\{x_j\}} - \widehat{y_j} \right\|^2 + \sum_{c=1}^{C} \left\| \hat{w} * \hat{f}^c \right\|^2 + \lambda \| P \|_F^2 \tag{6}$$

where "$\hat{\cdot}$" denotes the Fourier coefficients of the corresponding values. The resulting system of equations can be efficiently solved using this conjugate.

2.3. Adaptive Long Term Tracking Framework Based on Computer Vision

In 2023, Chen et al. [33] proposed a computer vision-based framework for long-term tracking that incorporated an innovative uncertainty estimation method. This method determined the necessity for model re-detection and updates in a timely fashion and assessed each frame to enable adaptive model updating. Such a method has been shown to enhance the algorithm's performance and robustness, as evidenced by its superior results on public datasets. When applied to unmanned vehicles, this framework facilitated the effective identification and tracking of pedestrians, thereby empowering robotic or autonomous vehicular systems to execute critical functions, including obstacle avoidance and target tracking.

Within the framework of the ECO tracking algorithm, the conjugate gradient method is employed to solve for the filter parameters, which necessitates the prior determination of the sample energy as a known parameter. Assuming X_t is the discrete Fourier transform of the training sample in the t-th frame, the sample energy E_s is calculated by equation （7）:

$$E = \| X_t \|^2 \tag{7}$$

The sample energy for each frame is updated using the learning rate (8):

$$E_t = \eta \| X_t \|^2 + (1-\eta) E_{t-1} \tag{8}$$

The framework determines whether the model needs updating based on the proposed tracking confidence and peak-to-sidelobe ratio, as shown in equation:

$$\eta = \begin{cases} P & M_{PSR} \geqslant \tau_{PSR},\ t \geqslant \tau_t \\ P/2 & M_{PSR} < \tau_{PSR},\ t \geqslant \tau_t \ or \ \ M_{PSR} \geqslant \tau_{PSR},\ t < \tau_t \\ 0 & M_{PSR} \leqslant \tau_{PSR},\ t \leqslant \tau \end{cases} \tag{9}$$

Here, P is the value of the sample learning rate η set during the initialization of the tracking algorithm, which is a constant. τ_{PSR} and τ_t are the thresholds for tracking uncertainty detection. As seen in the equation, when both the peak-to-sidelobe ratio and tracking confidence are above the thresholds, the target is considered clearly visible, and the model is updated normally. If either falls below the threshold, it indicates significant deformation or partial occlusion of the target, which is still visible. Setting the learning rate to zero in this case might lead to overfitting, so the sample learning rate is reduced. When both values are below the thresholds, it implies severe occlusion or disappearance of the target, and the sample learning rate τ_t is set to zero to prevent erroneous updates. The model structure, as shown in figure x, includes three distinct modules: detection, redetection, and tracking.

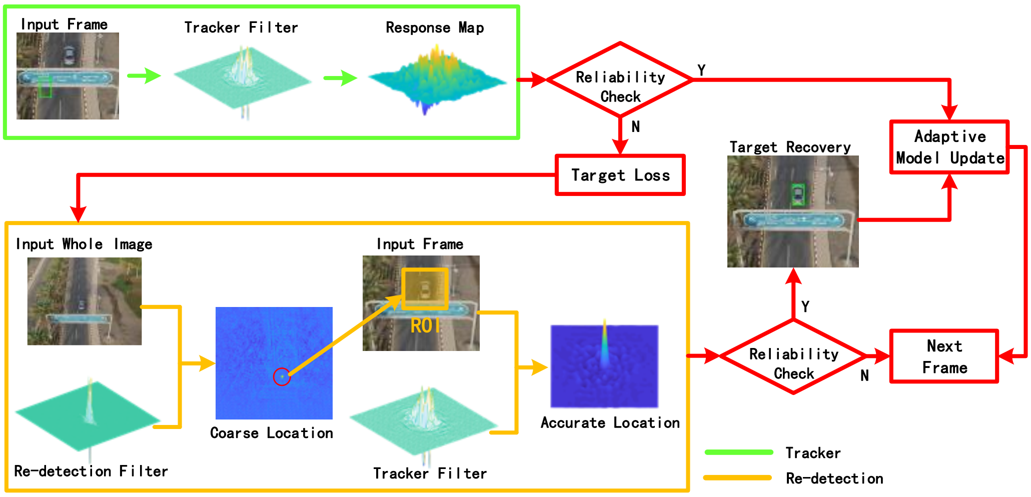
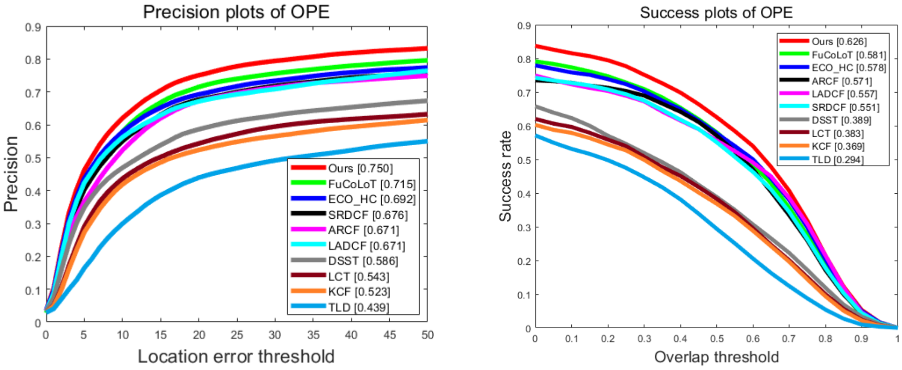


Figure 5. Adaptive long-term tracking framework based on visual detection.

The framework (red) demonstrates its test accuracy and coverage rate on public datasets UAV20 compared to other visual detection and tracking algorithms, as shown in the following figure. The localization accuracy reached 75%, an improvement of 8.4%, and the coverage rate was 62.6%, an increase of 8.3%.



(a) (b)

Figure 6. Positioning accuracy map and coverage of system frame.

3. Review of deep learning-based computer vision

3.1. Computer Vision Datasets and Metrics

The construction of larger datasets with reduced bias is a fundamental aspect of advancing computer vision algorithms. Over the past decade, a multitude of well-known datasets and benchmarks have been released, such as those from the PASCAL VOC Challenges [34,35](e.g., VOC2007, VOC2012) and visual recognition challenges [36,37](e.g., UAV20L, UAV123). In addition to general object detection, the past two decades have witnessed a proliferation of detection applications in specific domains, including pedestrian detection, face detection, text detection, traffic sign/light detection, and remote sensing target detection. Table 1 provides a compilation of some of the popular datasets for these specific detection tasks.

Table 1. An overview of some popular and challenged computer vision datasets.

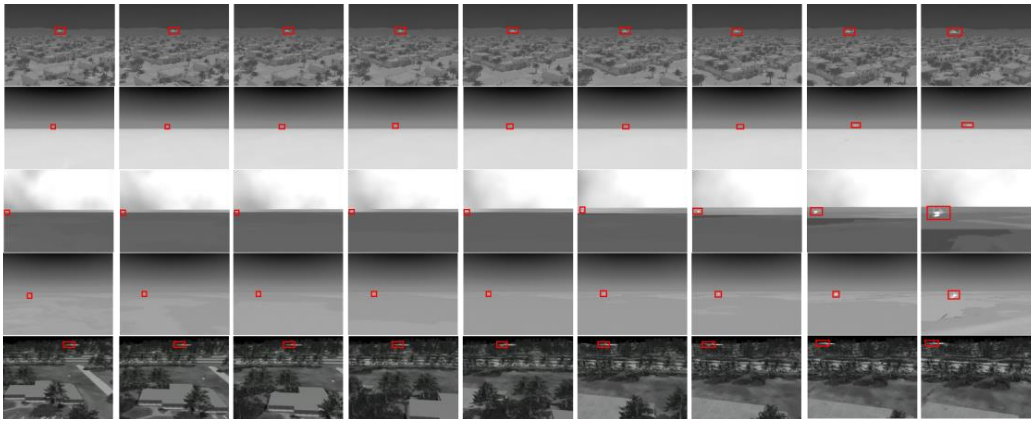
Dataset/Year	Cites	Description	Frames	URL
TLR[38] 2009	164	Traffic scenes in Paris	20,200	http://www.lara.prd.fr/benchmarks/trafficlightsrecognition
KITTI[39] 2012	2620	The traffic scene analysis in Germany.	16,000	http://www.cvlibs.net/dataset-s/kitti/index.php
BelgianTSD[40] 2012	224	The traffic sign annotations of 269 types. With the 3D location	138,300	https://btsd.ethz.ch/shareddata/
GTSDB[41] 2013	259	Traffic scenes in different climates	2,100	http://benchmark.ini.rub.de/?section=gtsdb&subsecti-on=news
IJB[42] 2015	279	IJB scenes for recognition and detection tasks.	50,000	https://www.nist.gov/programs-projects/face-challenges
WiderFace[43] 2016	193	Face detection scene	32,000	http://mmlab.ie.cuhk.edu.hk/pr-ojects/WIDERFace/
NWPU-VHR10[44] 2016	204	Remote sensing detection scenario	4,600	http://jiong.tea.ac.cn/people/JunweiHan/NWPUVHR10dataset.html

doi:10.20944/preprints202405.1051.v1

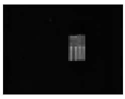
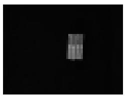
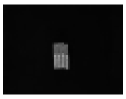
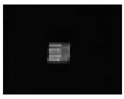
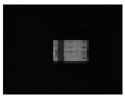
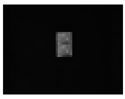
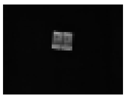
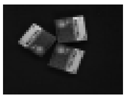
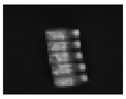
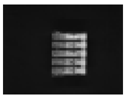
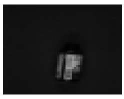
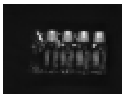
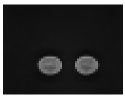
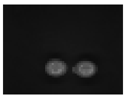
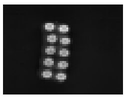
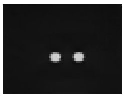
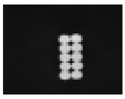
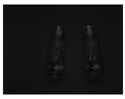
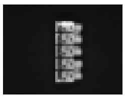
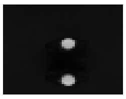
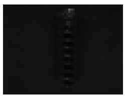
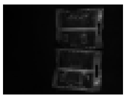
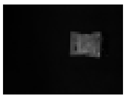
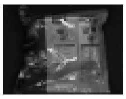
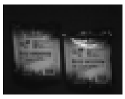
Concurrently, we have established our own datasets to meet the demands of in-depth research and practical applications. These datasets encompass various domains such as weak military targets in infrared scenes (WMTIS), the appearance of medical industry medicine boxes (MB), express packages (EP), personnel in fully automated unmanned factories, and product detection and tracking (FPP). Table 2 outlines the parameters and descriptions of these diverse datasets, while Figure 2 provides illustrative examples from each dataset.

Table 2. Our computer vision datasets for different challenges

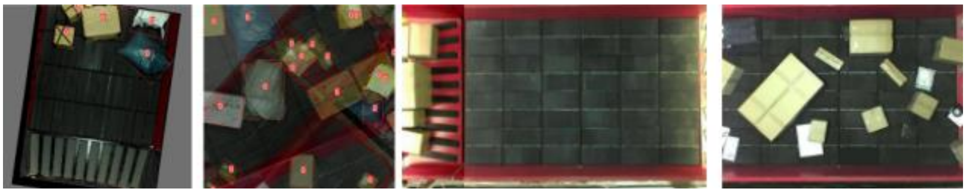
Our own Datasets (Year)	**Train**		**Description**
	images	**objects**	
WMTIS（2019）	1632	1808	The infrared simulation weak target dataset constructed based on the infrared characteristics of military targets includes a series of challenging samples with scale changes, such as fighter jets, tanks, and warships, in desert, coastal, inland, and urban backgrounds
MB（2023）	3345	9612	Various types of medicine boxes made of various materials, covering all mainstream types of pharmacies, including challenges such as reflection caused by waterproof plastic film
EP（2022）	25127	60393	A comprehensive sample covering all types of packages in the logistics and express delivery industry, with sizes ranging from 5 centimeters to 3 meters and heights ranging from 0.5 millimeters to 1.2 meters in various shapes
FPP（2022~2024）	9716	17435	Multi target samples in complex industrial scenes face many challenges such as easy occlusion, uneven illumination, inconsistent imaging quality, and open scenes. This includes production personnel and samples of various types of products, collected through various methods such as ground robotic vehicles and UAV



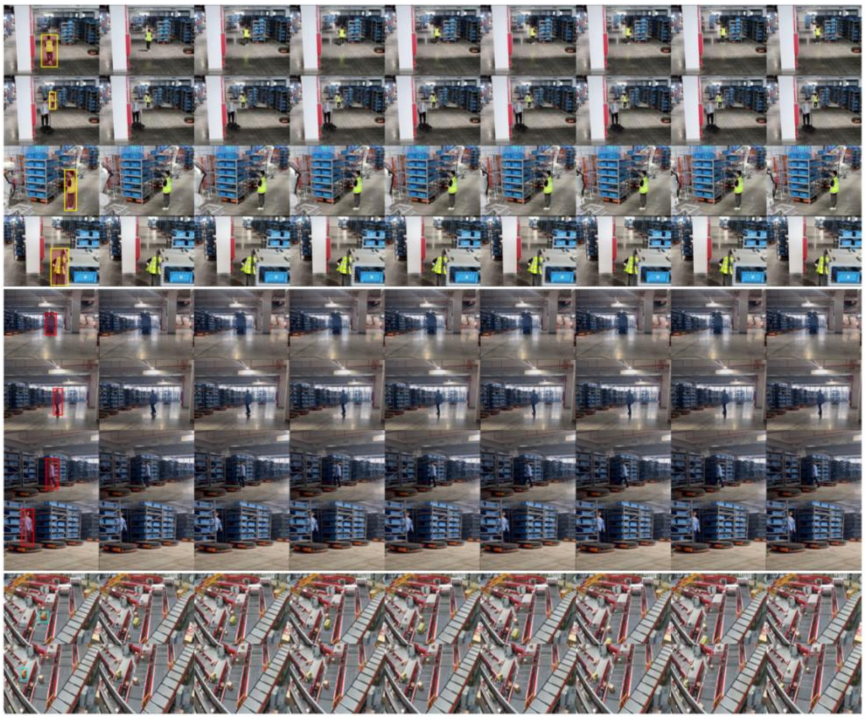
(a)WMTIS

Box-side B0					
Box-front B1					
Bottle-side B2					
Bottle-top B3					
Pot-side B4					
Pot-top B5					
bagB6					

(b)MB



(c)EP



(d)FPP

Figure 7. The datasets we have built ourselves in recent years.

3.2. Review of Deep Learning Computer Vision Based on Convolutional Neural Networks

Machines can now interpret and understand visual data in ways that were before impractical due to deep learning, which has completely changed the field of computer vision. Deep neural networks, especially convolutional neural networks (CNNs), which are built to automatically and adaptively learn spatial hierarchies of features from vast volumes of image data, are at the core of this transformation. These networks are made up of several neuronal layers, each of which is trained to detect increasingly complex features, such as edges and textures in the initial layers and higher-level objects and patterns in the deeper layers. Training these networks involves using large datasets of labeled images and a backpropagation algorithm to adjust the weights of the neurons so that the network can accurately classify or recognize objects. Deep learning in computer vision encompasses principles that apply to tasks like object detection, semantic segmentation, and image generation. These tasks utilize deep neural networks to extract meaningful information from pixels and convert it into realistic images or actionable insights. Applications ranging from surveillance and augmented reality to medical imaging and autonomous cars with sensing capability have significantly advanced as a result.

As was mentioned in the preceding section, there are now two main types of core visual detectors. The first is the one-stage detector network, which is appropriate for relatively basic application situations requiring high real-time performance because of its rapid inference speed. YOLO, SSD[45], and other instances are among the most prevalent ones. Figure 8 shows the fundamental structure of this type of detector.

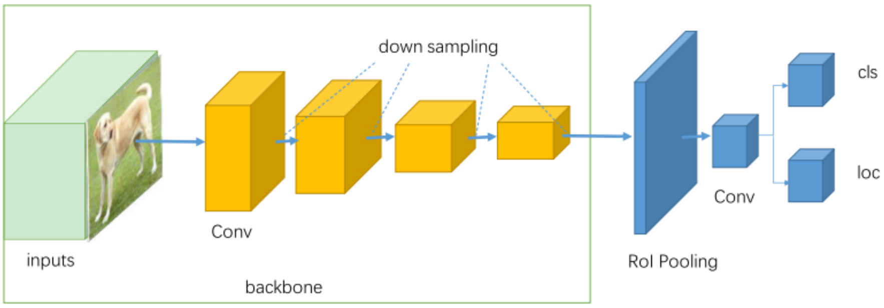


Figure 8. Network structure of one-stage detector.

In the structure, the yellow part represents a series of convolutional layers in the backbone network with the same resolution, while the blue part is the RoI pooling layer, which generates feature maps for objects of the same size.

The second type is the two-stage detector network, exemplified by CNN, RCNN, and Fast RCNN. This type complements the one-stage detectors with its higher localization precision and recognition accuracy. However, it has longer inference times and demands greater computational power, making it suitable for complex applications requiring high accuracy. The two stages of two-stage detectors are separated by the RoI (Region of Interest) pooling layer. The first stage, known as the Region Proposal Network (RPN), proposes candidate object bounding boxes. In the second stage, the RoI Pooling operation extracts features from each candidate box for subsequent classification and bounding box regression tasks. Taking Fast-RCNN as an example, its basic structure is illustrated in Figure 9.

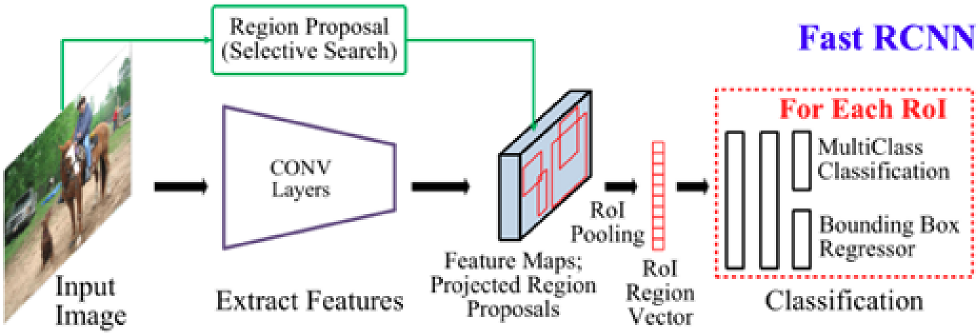


Figure 9. Network structure of two-stage detector.

The RCNN network is composed of four integral modules. The first module is dedicated to region proposal. The second module extracts fixed-length feature vectors from these proposed regions. The third module comprises a suite of linear support vector machines for specific categories. The final module is tasked with bounding box regression. The Fast R-CNN, as an advancement over its predecessor, commences by extracting features from the entire input image. Subsequently, it achieves scale-invariant feature representation through the RoI (Region of Interest) pooling layer. These processed features serve as the input for the subsequent fully connected layers responsible for classification and bounding box regression. This integrated approach to one-time classification and localization significantly enhances the inference speed, thereby improving the overall efficiency of the model.

In recent years, a plethora of enhanced model algorithms have been applied to diverse public datasets. Through experimental evaluation, it has been noted that, despite not being tested under entirely identical sub-datasets, their performance generally exhibits a upward trend over time. Table 3 presents the results of various algorithms on the MS COCO dataset, with Average Precision (AP) serving as the primary evaluation metric.

Table 3. Detection results on the MS COCO

Algorithmic network	Data	Backbone	AP
Fast R-CNN[46]	train	VGG-16	19.7
Faster R-CNN[47]	trainval	VGG-16	21.9
SSD321[48]	Trainval35K	ResNet-101	28.0
YOLOv3[49]	Trainval35K	DarkNet-53	33.0
RefineDet512+[50]	Trainval35K	ResNet-101	41.8
NAS-FPN[51]	Trainval35K	AmoebaNet	48.0

Figure 10 illustrates the performance of several mainstream algorithms on the VOC dataset[29], showing a steady improvement in their performance.

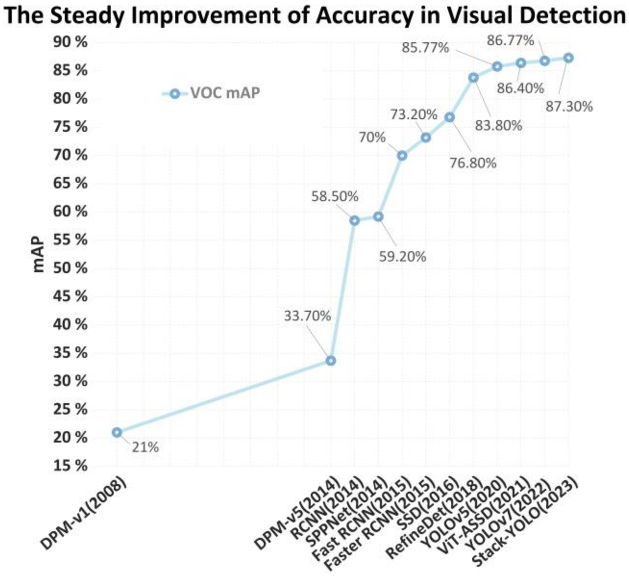


Figure 10. The steady improvement of accuracy in visual detection algorithms on VOC dataset.

3.3. ACDet:A Vector Detection Model for Drug Packaging Based on Convolutional Neural Network

ACDet is an algorithmic solution proposed by the authors for computer vision detection in the medical industry, integrating the algorithms mentioned previously within the YOLOv8 framework. The medical sector, with its rapid automation advancements, exhibits a substantial demand for visual detection technologies. The medical industry encounters numerous challenges in drug detection, attributable to the complexity of pharmaceuticals, the diversity of packaging materials, and the variety of formats, as exemplified by the EP dataset introduced in Section 3.1. These challenges encompass issues such as uneven lighting and the necessity for high response speeds. To address these problems, we devised a universal lightweight vector detection model. By optimizing the multi-computation module C2F-A, the model amplifies attention across multiple dimensions of gradient flow outputs, thereby enabling efficient and rapid classification of various drugs by improving the sensing ability of the network . The architecture of the model is illustrated in Figure 11.

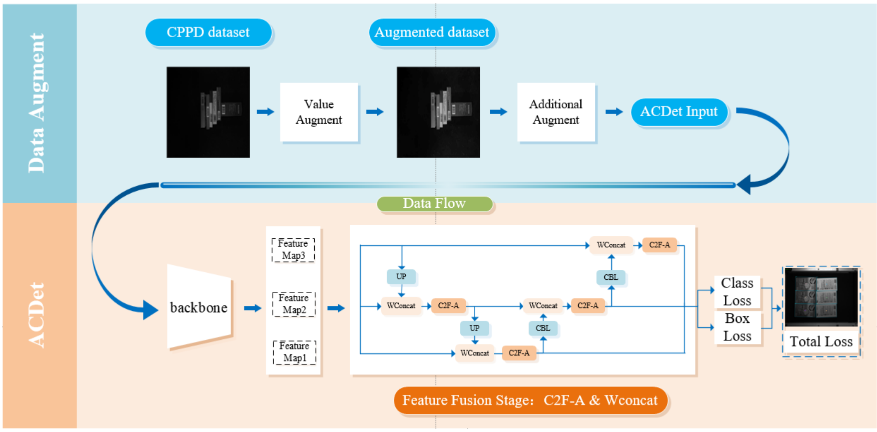
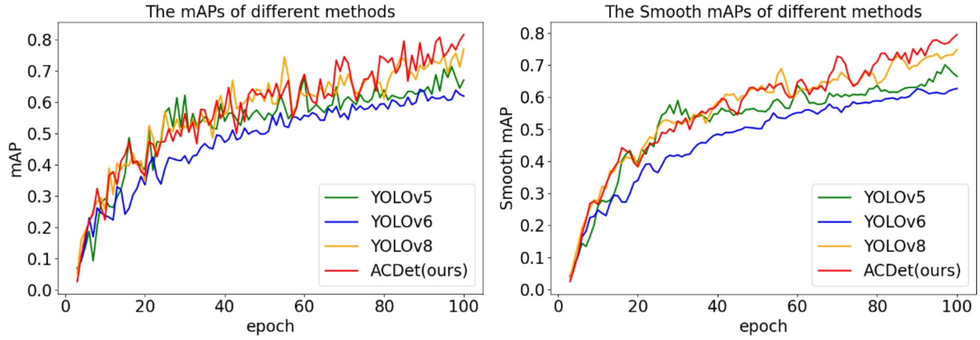


Figure 11. ACdet Model System Architecture.

Upon testing, the model achieved an mAP of over 81% on the EP dataset. Under identical testing conditions, its performance surpasses that of YOLOV5 to V8 versions by 6.3% to 19.4%, as demonstrated in Figure 12. This vision system has been extensively deployed, generating substantial market value, as depicted in Figure 12. In practical applications, the system's accuracy can exceed 99.9%.



(a) (b)

Figure 12. Test results of ACdet model on EP dataset.

Figure 13. A Computer Vision Detection System Based on Deep Learning Applied in the Medical Industry.

3.4. Exploration and Future Trends

The development of deep learning algorithms is intrinsically linked to the computational capabilities of hardware. In recent years, the rapid advancement of high-performance graphics processing units (GPUs) has accelerated this process, propelling the revolution in artificial

doi:10.20944/preprints202405.1051.v1

intelligence. Notably, the recent introduction of the Blackwell architecture GPU, featuring 208 billion transistors and employing a custom, dual-reticle TSMC 4NP process, has been a significant milestone. The interconnect speed between two smaller chips reaches up to 10 TBps, which not only elevates the computational power to 20 petaflops (FP4 precision) but also reduces energy consumption to one-twenty-fifth of its previous level. We believe that this technological breakthrough will have a profound impact on several aspects of computer vision, shaping its future direction and trends.

1.Enhanced Performance and Real-time Processing

High-performance hardware, such as the Blackwell architecture, equipped with parallel processing capabilities and high memory bandwidth, has dramatically accelerated the performance of computer vision algorithms. Its ability to handle complex matrix operations and parallel computations efficiently has enabled real-time processing of high-resolution images and videos. This is crucial for applications like autonomous vehicles, where split-second decisions based on visual data can be lifesaving. Furthermore, the enhanced performance of these latest cards has facilitated the development of more sophisticated and accurate computer vision models, pushing the boundaries of what is achievable in object detection, image recognition, and 3D reconstruction..

2.Energy Efficiency and Sustainability

The energy efficiency of high-performance hardware represents another critical factor influencing the future trajectory of computer vision. As AI models grow in complexity, the energy consumption associated with their training and inference processes has emerged as a pressing concern. Blackwell cards, designed with energy-efficient architectures, effectively reduce power consumption without compromising performance. This advancement not only facilitates the sustainable deployment of computer vision models but also extends their applicability to mobile and embedded systems, where power availability is constrained.

3. Democratization of AI Research

High-performance hardware has contributed to the democratization of AI research by making high-performance computing more accessible to a broader range of researchers and developers. The affordability and availability of these cards have lowered the entry barrier for individuals and small organizations to experiment with and develop computer vision models. This democratization fosters innovation and diversity in the field, as more people from different backgrounds can contribute to the advancement of computer vision technology.

4.Future Prospects

The future development of computer vision is intricately linked to advancements in GPU technology, with high-performance hardware playing a pivotal role. As these cards continue to evolve, we can anticipate further enhancements in processing speed, energy efficiency, and accessibility. This progress will facilitate the development of more sophisticated computer vision applications, such as augmented reality experiences, advanced surveillance systems, and more intelligent robotics. Moreover, the integration of AI-specific hardware features in Blackwell cards, such as tensor cores for deep learning, will further augment the capabilities of computer vision models, enabling more complex and efficient computations.

Simultaneously, we posit that an infinite increase in data volume results in the exponential growth of computing nodes, and addressing this issue solely through hardware advancements is not a sustainable solution. Current deep learning recognition algorithms exhibit an excessive dependency on training datasets. Consequently, exploring environmentally friendly, low-carbon models equipped with forgetting screening mechanisms constitutes one of the critical development directions in the field.

4. Review of visual Simultaneous Localization and Mapping (SLAM)algorithms

Visual SLAM (Simultaneous Localization and Mapping) plays a pivotal role in the realm of sensing driverless robotic vehicles. By leveraging computer vision techniques, these autonomous vehicles can dynamically construct a map of their surroundings while simultaneously determining their position within that map. This capability is crucial for navigating complex environments, avoiding obstacles, and ensuring efficient route planning. In the context of sensing driverless vehicles, visual SLAM contributes to sustainable operation by optimizing energy consumption through intelligent path planning and reducing the reliance on energy-intensive sensors. Furthermore, the integration of visual SLAM in these vehicles supports the development of advanced driver-assistance systems (ADAS) and autonomous driving technologies, thereby promoting safety and enhancing the overall efficiency of transportation systems.

4.1. The Basic Principles of SLAM

Visual SLAM digitizes real-world scenes by projecting 3D spatial points onto 2D pixel coordinates in the camera coordinate system, primarily utilizing the pinhole camera model. After the 3D space is projected onto the normalized image plane, distortion correction becomes necessary. Once processed, this data is fed into the visual front-end for VO (Visual Odometry) processing. The primary function of VO is to estimate the camera's motion roughly based on a series of adjacent image information and then provide this coarse information to the back end. Traditional visual odometry methods are mainly categorized into feature-based and direct methods, with feature-based visual odometry being the most widely utilized and developed. This process involves extracting feature points from each image, finding descriptors for feature matching, and then estimating different camera poses to obtain the corresponding visual odometry, as illustrated in Figure 14:

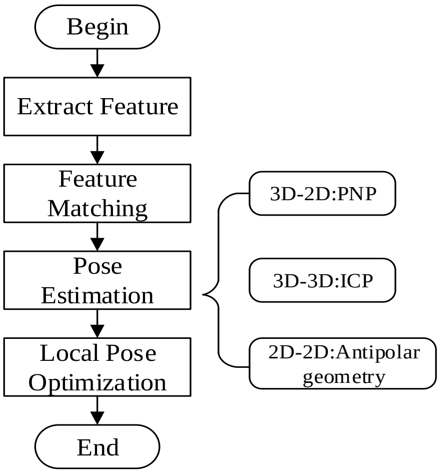


Figure 14. Principles of visual odometry.

The image information captured by different cameras varies, and so do the methods used for their pose estimation. For a monocular camera that obtains 2D pixel coordinates, the epipolar geometry method is employed. For stereo or depth cameras that can acquire 3D pixel coordinates, the Iterative Closest Point (ICP) method is used. If both 3D pixel coordinates and 2D camera projection coordinates are available, a combination of the two pieces of information can be utilized, employing the Perspective-n-Point (PnP) method.
After the front-end obtains the camera's motion and observation equation, as shown in Equation (10):

$$\begin{cases} x_k = f\left(x_{k-1}, u_k\right) + w_k \\ z_{k,j} = h\left(y_j, x_k\right) + v_{k,j} \end{cases} \quad k = 1, \ldots, N,\ j = 1, \ldots, M \tag{10}$$

The Backend optimization is conducted to eliminate the accumulated errors and uncertainties caused by noise. This is primarily achieved through graph optimization, where the objective function can be solved using methods such as Gauss-Newton or Levenberg-Marquardt to obtain an optimized solution, resulting in better camera poses. Finally, loop closure detection is performed to determine if the detected trajectory forms a loop by comparing the similarity between the previous and current frames.Integrating visual SLAM and tracking algorithms with sensing multi-driverless vehicles can result in a system with autonomous intelligence, as illustrated in Figure 15.

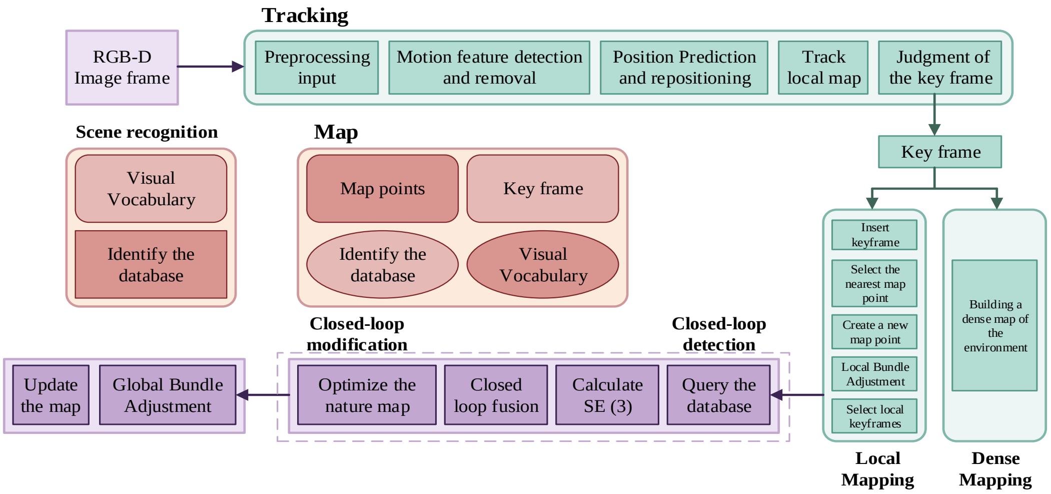


Figure 15. Unmanned driving system based on visual SLAM and tracking algorithms.

Utilizing an RGB-D camera as the sensor, the surrounding environmental information is perceived, while encoders and gyroscopes sense the internal operating information of the mobile robot. In the front end, based on the RGB and depth images obtained by the RGB-D camera, optical flow values are calculated and predicted. A dynamic target mask is generated by fusing the difference in depth values, thereby eliminating feature points on dynamic targets. Subsequently, based on the remaining static feature points, collinear and coplanar relationships are identified to extend static lines and planes. In the backend optimization phase, the pose is optimized by minimizing the residuals of static point features and static line features, thereby improving localization accuracy and enhancing navigation efficiency.

4.2. The basic Principles of SLAM

In recent years, the combination of deep learning and computer vision has achieved remarkable results in many tasks. Vision-based SLAM systems, grounded in computer vision, provide a broad development space for the application of neural networks in this field. Similar to object detection algorithms based on deep learning, the first step for SLAM systems based on deep learning is also the establishment of datasets. Table 4 shows some of the common public datasets in recent years.

Table 4. Common visual SLAM datasets.

Dataset	Release date	Collection situation
KITTI	2011	Camera/Radar/IMU/GNSS
TUM RGBD	2012	RGB Camera/Depth camera
EUROC	2016	Binocular Camera/RGB Camera/UAV
TUM VI	2018	RGB Camera/IMU
Openloris	2019	RGB Camera/IMU/Radar
Urbanloco	2020	Multiple Cameras/IMU/Radar/GNSS
TartanAir	2020	Outdoor simulation dataset
Brno Urban	2020	RGB Camera/IMU/Radar/Infrared
TUM-VIE	2021	Binocular Camera/IMU
M21XGR	2022	Multiple Cameras/laserGNSS

In recent years, a significant amount of research has been dedicated to optimizing the odometry of visual SLAM using deep learning methods, which are mainly categorized into supervised and unsupervised learning-based visual odometry approaches. Wang et al. [52] introduced the first end-to-end monocular visual odometry method based on deep recurrent convolutional neural networks, which directly learns from image sequences to achieve more accurate and stable visual odometry estimation. Xiao et al. [53] utilized a convolutional neural network to construct an SSD object detector combined with prior knowledge, enabling semantic-level detection of dynamic objects in a new detection thread. By employing a selective tracking algorithm to handle the feature points of dynamic objects, the pose estimation error caused by incorrect matching is significantly reduced. Duan et al. [54] proposed a keyframe retrieval method based on deep feature matching, which treats the local navigation map as an image and the keyframes as keypoints of the map image. Convolutional neural networks are used to extract descriptors of keyframes for loop closure detection. DROID-SLAM, proposed in 2021, is a novel SLAM system based on deep learning. Its front-end part performs feature extraction and calculates the optical flow field, calculates three keyframes with the highest degree of co-visibility based on the optical flow field, and then iteratively updates the pose and depth of the current keyframe based on the co-visibility relationship. The back-end part optimizes all keyframes using Bundle Adjustment (BA). Each time an update iteration is performed, a frame-graph is reconstructed for all keyframes. Compared to previous approaches, the robustness, generalization ability, and success rate have been greatly improved, achieving end-to-end visual SLAM using deep learning methods.

4.3. Discuss the Current Challenges and Future Research Directions of Visual SLAM

Visual Simultaneous Localization and Mapping (SLAM) faces several challenges, including robustness to varying lighting conditions, dynamic environments, feature extraction and matching, computational efficiency, and scalability. Future research trends include the integration of deep learning techniques for improved feature extraction and object recognition, the development of more efficient algorithms for real-time processing, the fusion of multiple sensors for enhanced accuracy and robustness, and the exploration of semantic SLAM for a deeper understanding of the environment. However, current deep SLAM methods still have the following shortcomings:

1.Data volume and labeling: Deep learning necessitates large-scale data and accurate labeling, yet acquiring large-scale SLAM datasets poses a significant challenge.

2.Low real-time performance: Visual SLAM often operates under real-time constraints, and even input from low-frame-rate, low-resolution cameras can generate a substantial amount of data, requiring efficient processing and inference algorithms.

3.Generalization ability: A critical consideration is whether the model can accurately locate and construct maps in new environments or unseen scenes. Future advancements in deep SLAM methods are expected to increasingly emulate human perception and cognitive patterns, making strides in high-level map construction, human-like perception and localization, active SLAM methods, integration with task requirements, and storage and retrieval of memory. These developments will aid robots in achieving diverse tasks and self-navigation capabilities. The end-to-end training mode and information processing approach, which align with the human cognitive process, hold significant potential.

4.4. Visual Framework for Unmanned Factory Applications with Multi-Driverless Robotic Vehicles and UAVs

Based on the methods summarized in this article, we have discussed and proposed a complete vision system for complex intelligent factory environments. This system can be deployed on multiple green energy robots, drones, or vehicles, and can perform a series of tasks such as path planning, automatic navigation, intelligent obstacle avoidance, cargo grabbing, human following, emergency rescue, and more. It includes a computer vision detection module based on deep learning, comprising a composite detection and recognition module (including SVM classifiers and EPDet), a vision tracking module, a vision fusion laser SLAM positioning module, and a vehicle chassis drive control module, as shown in Figure 16. This can provide a reference for the application of computer vision technology in the industrial field.

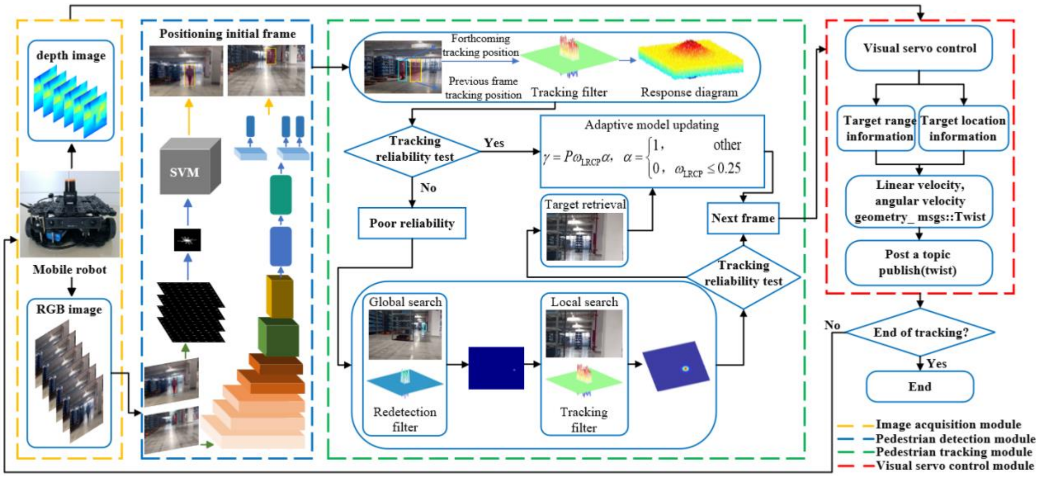

depth image
Mobile robot
RGB image
Positioning initial frame
SVM
Forthcoming tracking position
Previous frame tracking position
Tracking filter
Response diagram
Tracking reliability test
Yes
No
Adaptive model updating
$\gamma = P\omega_{LRCP}\alpha, \quad \alpha = \begin{cases} 1, & \text{other} \\ 0, & \omega_{LRCP} \le 0.25 \end{cases}$
Poor reliability
Target retrieval
Next frame
Tracking reliability test
Global search
Redetection filter
Local search
Tracking filter
Visual servo control
Target range information
Target location information
Linear velocity, angular velocity geometry_ msgs::Twist
Post a topic publish(twist)
No
End of tracking?
Yes
End
Image acquisition module
Pedestrian detection module
Pedestrian tracking module
Visual servo control module

Figure 16. Visual framework for unmanned factory applications with Multi-Driverless Robotic Vehicles.

5. Conclusions

With the rapid advancement of computer hardware, the potential for significant development in the field of computer vision is evident, as demonstrated in this paper. Computer vision serves as a foundational element in the realm of intelligent systems and AI technologies. It can effectively replace sensors as the primary source of data input, acting as a pair of intelligent eyes that can more adeptly process data sources to drive a range of unmanned actions, such as autonomous driving in sensing driverless vehicles, obstacle avoidance, and detection tracking. The potential ceiling for this field is exceptionally high.

Furthermore, this article provides a systematic and comprehensive overview of the concepts and application directions of computer vision algorithms. It encompasses a range from traditional detection algorithms to deep learning detection algorithms based on convolutional neural networks, and further extends to visual SLAM, deep visual SLAM, and long-term tracking algorithms, offering a generalized description of their principles. It also delineates the principles, characteristics, and test conclusions of typical detection and tracking models. Additionally, the article outlines a system framework proposed based on existing computer vision algorithms, which has achieved notable results.

The future development direction of this field is also explored in depth in this paper. The current mainstream deep learning algorithms exhibit a high dependency on datasets and hardware. Concurrently, there is an increasing demand for high-precision real-time performance in current systems. Developing models that are "healthy" and "green" , incorporating a forgetting mechanism to promote low-carbon environmental protection, represents one of the significant challenges for the future..

Author Contributions: Writing—original draft preparation, L.C ,J.T.,Y.L and J.P; writing—review and editing, W.X,D.G and Z.C; supervision, W.S.and G.L.All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research was funded by Science and Technology Research Project of Chongqing Municipal Education Commission under grant number 010725111.

Data Availability Statement: Data are contained within the article.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. P. Dollár, C. Wojek, B. Schiele, and P. Perona, ''Pedestrian detection: An evaluation of the state of the art,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 743–761, Apr. 2012.
2. A. Geiger, P. Lenz, and R. Urtasun, ''Are we ready for autonomous driving? The KITTI vision benchmark suite,'' in Proc. Int. Conf. Pattern Recognit., Jun. 2012, pp. 3354–3361.
3. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ''ImageNet large scale visual recognition ch
4. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ''Microsoft COCO: Common objects in context,'' in Computer Vision—ECCV, D. Fleet, T. Pajdla, B. Schiele, and
5. T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 740–755.
6. A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, ''The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale,'' 2018, arXiv:1811.00982. [Online]. Available: https://arxiv.org/abs/1811.00982
7. P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, ''Vision meets drones: A challenge,'' 2018, arXiv:1804.07437. [Online]. Available: https://arxiv.org/abs/1804.07437
8. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1. IEEE, 2001, pp. I–I.N.
9. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.
10. P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.
11. Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C].2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2010: 2544-2550.
12. Henriques J F, Caseiro R, Martins P, et al.High-Speed Tracking with Kernelized Correlation Filters[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2015, 37 (3): 583-596.
13. Xu T, Feng Z H, Wu X J, et al.Learning Adaptive Discriminative Correlation Filters via Temporal Consistency Preserving Spatial Feature Selection for Robust Visual Object Tracking[J].IEEE Transactions on Image Processing,2019, 28 (11): 5596-5609.
14. Huang Z, Fu C, Li Y, et al. Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking[C].2019 IEEE/CVF International Conference on Computer Vision (ICCV),2019: 2891-2900.
15. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on compputer vision and pattern recognition, 2014, pp. 580–587.
16. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.
17. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
18. J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," arXiv preprint, 2017.
19. A. Aboah, B. Wang, U. Bagci and Y. Adu-Gyamfi, "Real-time Multi-Class Helmet Violation Detection Using Few-Shot Data Sampling Technique and YOLOv8," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023, pp. 5350-5358
20. Chen B X, Sahdev R, Tsotsos J K. Person Following Robot Using Selected Online Ada-Boosting with Stereo Camera[C].2017 14th Conference on Computer and Robot Vision (CRV),2017: 48-55.
21. Liu, X., & Zhang, Z. (2021). A vision-based target detection, tracking, and positioning algorithm for unmanned aerial vehicle. Wireless Communications and Mobile Computing, 2021, 1-12.
22. TEED Z,DENG J. DROID SLAM: Dccp Visual SLAM for Monocular, Stereo, and RGB-D Cameras[J]. arXiy.:2108.10869.2021.
23. Evjemo, L. D., Gjerstad, T., Grøtli, E. I., & Sziebig, G. (2020). Trends in smart manufacturing: Role of humans and industrial robots in smart factories. Current Robotics Reports, 1, 35-41.
24. L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikainen, "Deep learning for generic object detection: A survey," arXiv preprint arXiv:1809.02165, 2018.
25. S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," arXiv preprint arXiv:1809.03193, 2018.
26. A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," Computer vision and image understanding, vol. 117, no. 8, pp. 827–891, 2013.
27. J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in IEEE CVPR, vol. 4, 2017.

28. K. Grauman and B. Leibe, “Visual object recognition(synthesis lectures on artificial intelligence and machine learning),” Morgan & Claypool, 2011.
29. Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.
30. Jiao L, Zhang F, Liu F, et al. A survey of deep learning-based object detection[J]. IEEE access, 2019, 7: 128837-128868.
31. Liu Y, Dai Q. A survey of computer vision applied in aerial robotic vehicles[C]//2010 International Conference on Optics, Photonics and Energy Engineering (OPEE). IEEE, 2010, 1: 277-280.
32. Danelljan M, Bhat G, Khan F S, et al. ECO: Efficient Convolution Operators for Tracking[C].2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017: 6931-6939.
33. Chen L, Li G, Zhao K, et al. A Perceptually Adaptive Long-Term Tracking Method for the Complete Occlusion and Disappearance of a Target[J]. Cognitive Computation, 2023, 15(6): 2120-2131.
34. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” International journal of computer vision, vol. 88, no. 2, pp. 303–338, 2010.
35. M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International journal of computer vision, vol. 111, no. 1, pp. 98–136, 2015.
36. Mueller M, Smith N, Ghanem B. A benchmark and simulator for uav tracking. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing; 2016. p. 445–61.
37. Wu Y, Lim J, Yang MH. Online object tracking: a benchmark. Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 2411–8.
38. R. De Charette and F. Nashashibi,"Real time visual traffic lights recognition based on spot light detection and adaptive traffic lights templates ,” in Intelligent Vehicles Symposium, 2009 IEEE. IEEE,2009, pp. 358-363.
39. A. Geiger, P Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite,"in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, Pp.3354-3361.
40. R. Timofte, k. Zimmermann, and L. Van Gool, "Multiview traffic sign detection, recognition, and 3d localisation," Machine vision and applications, vol. 25, no. 3, pp.633 647,2014
41. S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, andC. Igel, "Detection of traffic signs in real-world images:The german traffic sign detection benchmark," in Neural Networks (I/CNN), The 2013 International Joint Conference on. IEEE,2013,PP.1-8.
42. B. F. Klare, B. Klein, E. Taborsky, A. Blanton,J . Cheney,K. Allen, P.Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition:Iarpa janus benchmark a," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015Pp.1931-1939
43. S. Yang, P. Luo, C.-C. Loy and X. Tang, "Wider face:A face detection benchmark," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016,PP.5525 5533.
44. G. Cheng and J. Han, "A survey on object detection in optical remote sensing images,"ISPRS Journal of Photogrammetry and Remote Sensing, vol. 117, pp. 11-28, 2016
45. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ‘‘SSD: Single shot multibox detector,’’ in Computer Vision— ECCV, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.
46. C. P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in Computer vision, 1998. sixth international conference on. IEEE, 1998, pp. 555–562.
47. K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang et al., “T-cnn: Tubelets with convolutional neural networks for object detection from videos,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2018.
48. ——, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.
49. Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” Journal-Japanese Society For Artificial Intelligence, vol. 14, no. 771-780, p. 1612, 1999.
50. T. Malisiewicz, A. Gupta, and A. A. Efros, “Ensemble of exemplar-svms for object detection and beyond,” in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 89–96.
51. R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
52. Wang S, Clark R, Wen H, et al. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]//2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017: 2043-2050.
53. Xiao L, Wang J, Qiu X, et al. Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment[J]. Robotics and Autonomous Systems, 2019, 117: 1-16.

54. Duan R, Feng Y, Wen C Y. Deep pose graph-matching-based loop closure detection for semantic visual SLAM[J]. Sustainability, 2022, 14(19): 11864.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.