Review

Ravi Prakash Chaturvedi* and Udayan Ghose

A review of small object and movement detection based loss function and optimized technique

https://doi.org/10.1515/jisys-2022-0324
received December 23, 2022; accepted March 02, 2023

Abstract: The objective of this study is to supply an overview of research work based on video-based networks and tiny object identification. The identification of tiny items and video objects, as well as research on current technologies, are discussed first. The detection, loss function, and optimization techniques are classified and described in the form of a comparison table. These comparison tables are designed to help you identify differences in research utility, accuracy, and calculations. Finally, it highlights some future trends in video and small object detection (people, cars, animals, etc.), loss functions, and optimization techniques for solving new problems.

Keywords: detection of small objects, detection of video objects, loss functions, optimization

1 Introduction

Object detection is a computer technology method that is connected to object recognition and networks. It can recognize instances of specific semantic item categories (for example, people, buildings, or cars) in computer-generated pictures and videos [1]. In-depth object recognition research may be shown in the recognition of faces and pedestrians. Object recognition is used in many image-processing applications, including image search and video surveillance [2]. Each object class contains distinguishing characteristics that help in categorizing the class. For example, all circles are circular. These specialized procedures are employed in the identification of object classes. When looking for a circle, for example, you are looking for anything that is at a certain distance away from the point (i.e., the center). Items that are upright at the corners and have the same side length are also necessary while looking for a square. A similar method is used for facial recognition, which can identify the eyes, nose, and mouth as well as skin color and the distance between the eyes [3,4]. The challenge of anticipating the types and locations of distinct things present in a picture is known as a realm of image processing, there is a difficulty with object recognition. In contrast to classification, each instance of an object is recognized in the object recognition task, so object recognition is a measurement task for each instance. In contrast to classification, each instance of an object is recognized in the object recognition task, so object recognition is a measurement task for each instance as the scale-invariant feature transform [5] and the histogram orientation gradient [6].

A remote sensing imaging study has gained a lot of interest as the remote sensing technology has advanced. Simultaneously, the identification of ships and airplanes using optical remote sensing images [7]

*** Corresponding author: Ravi Prakash Chaturvedi,** University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Delhi 110078, India, e-mail: rpchaturvedi51@gmail.com
Udayan Ghose: University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Delhi 110078, India, e-mail: udayan@ipu.ac.in

is significant in a wide range of applications [8]. Over the last 10 years, many regional convolutional neural network (R-CNN) techniques [9], particularly faster R-CNN [10], have been utilized in the high-resolution identification of high-resolution items in the PASCAL VOC dataset. They cannot detect extremely small things information in general because complex pictures are hard to evaluate due to their low construction and appearance [11]. Objects in optical distant sensing pictures, on the other hand, typically have smaller characteristics, which bring more problems than traditional object detection, and there are still few good solutions [12,13]. There have been some attempts to overcome the issue of small object detection (SOD). By simply raising the input picture resolution, it is simple to enhance the resolution of the object's fine details, which generally results in a substantial investment in training and testing [14,15]. Another method aimed to create a multi-scale representation that combined multiple functions at a lower level to expand the function at a higher level, thereby naively magnifying the size of the function [16,17].

Video motion detection is a function of IP cameras, recording software and NVR used to trigger an alarm by detecting physical movement in a designated area. In real time, the data from the current image is compared to the data from the previous image. Every major change triggers a camera warning [18]. This alarm can be used to trigger many operations, such as sending current real-time images via email, tilt or move the camera at a certain point, control external devices (such as turning on lights or beeping) etc. and use still images for regular object recognition. etc. Use still images for regular object recognition. There is increasing recognition of video objects (VOs), autonomous driving [19], and video surveillance [20,21]. In 2015, the use of video for object recognition became a new challenge for the surveillance [22]. The Image Net Visual Recognition Challenge is a large-scale visual recognition competition (ILSVRC2015). With the help of ILSVRC2015, the research on VO identification has progressed. Identifying things in each frame was one of the first attempts to recognize VOs.

This work studies and analyses the abovementioned network-based small target method, video detection, loss function, and optimization methods at different stages.

The main objective of this study is to give a summary of the studies on the identification of small objects and video-based networks. The first topic covered is the detection of small objects and VOs, as well as a study on current technology. The classification and description of the detection, loss function, and optimization strategies are presented as a comparison table.

2 Literature review

There are various approaches present for small object and movement detection. Some of the important literature that covers more important object detection is discussed below.

Chen et al. [23] proposed using deep learning to identify small objects. This study starts with a short overview of the four pillars of microscopic item identification: multi-scale rendering, contextual information, super-resolution, and range. Then, it offers a range of modern datasets for detecting small objects. Furthermore, current micro-object detection systems are being studied with an emphasis on modifications and tweaks to improve detection efficiency, in comparison to conventional object recognition technologies.

Ren et al. [24] studied how to tackle the challenge of employing remote sensing technology to identify tiny objects in optical imaging, and an enhanced faster R-CNN approach was developed. As a consequence of common characteristics, the studio built a comparable architecture that used downlink and avoided the use of connections to produce a single high-resolution, high-level feature map. This is critical so that we can view all the identified items.

Huang et al. [25] created a model for recognizing prominent objects in hyperspectral pictures on wireless networks, thereby using visibility optimization to CNN characteristics. The model first uses a two-channel CNN to extract the spatial and spectral properties of the same measurement and then employs functional combinations to produce the final bump map, which optimizes the bump value of the foreground and foreground signals. The CNN function is used to compute the background.

Hua et al. [26] proposed a real-time object recognition framework for cascaded convolutional networks using visual attention mechanisms, convolutional storage network inference methods, and semantic object relevance, combined with the fast and exact functions of deep learning algorithms, and performed ablation and comparative experiments. By testing the cascade network introduced in this study, different datasets can be used and more complex detection results can be obtained.

Yundong et al. [27] proposed a new method, that is, multi-block SSDs that add sub-layers to detect and extend local context information. The test results of multiple SSDs and conventional SSDs are compared. The algorithm shown increases the detection rate of small objects to 23.2%.

Bosquet et al. [28] proposed STDnet and ConvNet to identify tiny objects with a size of less than 16 × 16 pixels based on regional ideas. STDnet relies on an additional visual attention process called RCN, which chooses the most likely candidate area, consisting of one or more tiny items and their surrounding RCN feeds are more accurate and economical, improving accuracy while conserving memory and increasing the frame rate. This study also incorporates automated k-means anchoring, which improves on traditional heuristics.

Kunfu et al. [29] proposed a fully integrated framework for identifying objects in any orientation in remote sensing pictures. The web provides a functional aggregation architecture for obtaining functional representations for ROI discovery and ROI provision. The combination of quality recommendations and ROI-O is used to process recommendations for effective implementation.

Zheng et al. [30] introduced a new framework for large-scale target recognition, namely, HyNet for MSR remote sensing imaging, which opens up a new avenue for research of the depiction of scale-invariant functions. Display zoom functions are elements with pyramid-shaped detection areas, which are used to detect objects more accurately with multiple scales in MSR remote sensing images.

Tian et al. [31] provided a 3D recognition network that can provide a wide range of local functions from images, BEV maps, to point clouds. The adaptive merging network provides an effective method to merge multi-mode data functions. Whenever a vast number of objects appear, the adaptive weighting component restricts the intensity of each signal and chooses information for further evaluation, while the spatial fusion module includes azimuth and geometry info.

Li et al. [32] reported that PDF-Net is an optical RSI-specific SOD network that may employ mapping and cross-path data, as well as multi-resolution features, to efficiently and accurately identify outgoing objects of various sizes in optical RSIs. PDF-Net has always outperformed the modern SOD method in the ORSSD dataset in terms of visual comparison and quantification. Furthermore, ablation analysis verified the efficacy of the main components.

Fadl et al. [33] proposed a system that uses spatio–temporal information and fusion of two-dimensional convolutional neural networks (2D-CNNs) to detect inter-frame operations (delete frames, insert frames, and copy frames). RBF-Gaussian support vector machine (SVM) is utilized in the classification phase before automatically extracting depth characteristics.

Zhu et al. [34] outlined the approaches that have been discovered thus far for detecting VOs. This research examines the available datasets, scoring criteria, and provides an overview of the various classes of deep learning-based methods for identifying VOs. Depending on how time and space information is used, detection methods have been developed. These categories include flow-based technology, LSTM, nursing technology, and follow-up technology.

Alhimale et al. [35] researched and successfully developed a fall detection system that can fulfill the demands of the elderly (especially indoors). As a result, our video-based fall detection system decreases the likelihood that older individuals will be concerned about falling and will limit their activities at home or in solitude. Furthermore, fall detection systems have been created to preserve people's privacy, even when their everyday activities are dangerous, by tracking in real-time.

Lee et al. [36] proposed a new method using advanced neural network ART2 to detect scene changes. To capture the smooth interval, the suggested technique extracts the CC sequence from the video and then generates a gray-scale variance sequence. A typical progressively shifting local minimum sequence will develop during this procedure. It will be deleted from the softbox after being recovered by our local minimum detection method. Then, the resulting smooth intervals are combined to form a new sequence. From the new sequence, feature components such as pixel differences, histogram differences, and correlation coefficients can be extracted.

Kousik et al. [37] developed a deep learning problem-solving model that uses a new framework to combine CNNs with repetitive neural networks to discover the value of videos. By using recursive convolutional neural network (CRNN) to record time, space, and local restricted features to complete the task of finding obvious objects in the dynamic reference video dataset. Compared with conventional video recognition methods, the evaluation based on the reference dataset has advantages in accuracy, *F*-measure, mean absolute error, and calculation amount.

Xu et al. [38] presented a unique video smoke detection system based on a deep distribution network. The goal of bump detection is to emphasize the most important parts of things in a photograph. To generate realistic smoke highlights, outbound CNNs at the pixel and object levels are merged. For use in video smoke detection, an end-to-end architecture for recording departing smoke and predicting the existence of smoke is given.

Yang et al. [39] described a narrowband Internet of Things (NB-IoT) based digital video intrusion detection method, and an NB network-based digital video intrusion detection system was constructed. Intelligent categorization is accomplished through the usage of IoT and the SVM algorithm. The classification time, accuracy, and false alarm rate of the model were examined. The classification time is 40.80 s, the shortest is 27 s, the recognition rate is 87.60%, and the worst is 83.70%. The false detection rate may reach 15%, but it is always less than 20%, demonstrating that the classification system is reliable and accurate.

Yamazaki et al. [40] proposed a method for autonomously identifying surgical tools from video footage during laparoscopic gastrectomy. Validation has been performed on a unique automated approach based on the open-source neural network framework YOLOv3 for detecting surgical instrument operation in laparoscopic gastrostomy videotapes.

Yue et al. [41] used YOLO-GD (Ghost Net and Depth wise convolution) to detect the images of cups, chopsticks, bowls etc., and capture the different types of dishes (Table 1).

The above comparison table represents some small objects as well as movement detection techniques. Compared to the above techniques the Multi-block SSD approach achieves 96.6% percent overall accuracy, while CNN spatiotemporal features and fusion for surveillance video forgery detection yields excellent accuracy.

3 Studies related to SOD

The task of detecting little items is to detect small objects. Small object identification [42] is an intriguing issue in computer vision. In particular, we run models with different backbones in different datasets with multi-scale objects to find the object types and frameworks suitable for each model [43]. In this section, we will go through various techniques for enhancing tiny object detectors, such as

- Increasing picture capture resolution.
- Increasing the input resolution of the model.
- Using tiling on the pictures.
- Increasing data generation through augmentation.
- Model anchoring for self-learning.
- Eliminating superfluous classifications.

Figure 1 specifies the simplest way of detecting small objects.

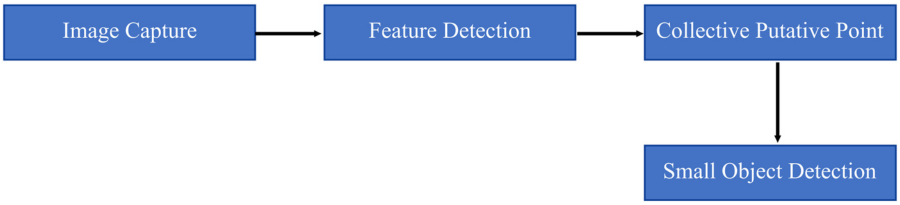

Figure 1: Structure of SOD.

Table 1: Comparative study of SOD as well as movement detection technique

Author	Year	Methodology	Advantage	Accuracy
Guang Chen, Zida Song	2020	DL-based SOD	The network benefits from integrating multi-resolution context information from available modules rather than directly from a single layer, resulting in more efficient computing.	Accuracy could be significantly improved.
Yun Ren, Changren Zhu	2018	Modified faster R-CNN method	Information extended up to some extent. MER gather more region than bounding box method.	—
Chen Huang, Tingfa Xu	2020	Hyper spectral images in wireless network	Less noise.	Improves its accuracy and efficiency.
Xia Hua, Xinqing Wang	2020	A cascaded CNN	—	Has high accuracy and adaptability.
Yundong LI, Han DONG	2019	Multi-block SSD	SSD offers two benefits: real-time processing and excellent precision.	The SSD technique yields an overall accuracy of 96.6%.
Brais Bosquet, Manuel Mucientes	2020	STDnet	Enhancements to speed and memory optimization.	Small-item detection accuracy using CNNs falls behind other bigger things.
Kun Fu, Zhonghan Chang	2020	Rotation-aware and multi-scale CNN	—	—
Zhuo Zheng, Yanfei Zhong	2020	HyNet framework	Improves the scale-invariant properties, making it ideal for object recognition in large-scale MSR remote sensing pictures.	Improves multi-scale item recognition accuracy in HSR remote sensing images.
Yonglin Tian, Kunfeng Wang	2020	Adaptive azimuth-aware fusion network	The advantage of this technique is that it makes use of point cloud data and the fusion style.	—
Chongyi Li, Runmin Cong	2020	A parallel down-up fusion network	In optical RSIs, the combination of various resolutions has a benefit in addressing size variation and seeing.	Increases the generalization capabilities and accuracy of networks.
Sondos Fadl, Qi Han	2020	CNN spatiotemporal features and fusion	Reduces the number of post-processing steps.	CNN system achieves high accuracy.
Laila Alhimale, Hussein Zedan	2014	video-based fall detection system using a NN	A trustworthy, accurate, user-friendly, and low-cost fall detection system that protects the privacy.	High sensitivity in detecting falls (greater than 90%).
Man-Hee Lee, Hun-Woo Yoo	2006	Improved ART2	The benefit of quick detection since no decompression time is required.	This method has a high degree of accuracy.
Nalliyanna V, Kousik	2020	Hybrid CRNN	—	The proposed approach is successful in terms of accuracy as well as speed.
Gao Xu, Yongming Zhang	2019	Video smoke detection based on DSNN	CNN was used to automatically learn the interaction between several low-level saliency signals and to capitalize on the information gained from these classical saliency detections.	Achieves the best result in terms of existing forecast accuracy.
Aimin Yang, Huixiang Liu	2020	A digital video intrusion detection method	Consistent performance, minimal power usage, and dependable data.	Low accuracy.
Yuta Yamazaki, MD, Shingo Kanaji	2020	Automated surgical instrument detection	Complex image analysis for medical professionals.	—
Xuebin Yue	2022	YOLO-GD (GhostNet and depth wise convolution)	Function of detection dish will lead to further development.	Higher accuracy.

Zhang et al. [44] proposed the boundary-aware high-resolution network (BHNet), which is a novel protruding item-detecting technique. BHNet is intended to be a parallel architecture. It allows for high-resolution information extraction from low-level functions, which is reinforced by various semantics, using a parallel architecture with a low resolution. There are also several multipath channel estimators and region extenders that capture more precise context-sensitive layer functionalities. To track the borders of visible objects, a loss function is given, which can assist us in determining precise detection bounds. BHNet is a specialist at locating exceptional items with powerful functions for extracting numerous characteristics.

Liang et al. [45] provided a context-sensitive network for identifying outgoing RGB-D objects. The suggested approach is divided into three components: feature extraction, multi-mode context fusion, and context-sensitive expansion. The first component is in charge of determining hierarchical functions based on color and depth. CNN was used in each photograph. The second component employs an LSTM version to include additional characteristics to represent multimodal spatial correlation in context. Experiment findings with two publicly accessible reference datasets demonstrate that the suggested technique is capable of providing the most recent performance for recognizing significant stereo RGB-D objects.

Kumar and Srivastava [46] developed an object identification method that recognizes things in pictures using deep learning neural networks. To obtain high target detection accuracy in real-time, this study integrates the Single Shot Multi-Block Detection method with faster CNN. This method is appropriate for both still pictures and videos. The proposed model's accuracy is greater than 75%. This model takes around 5–6 h to train. To extract information from visual characteristics, this model employs a CNN. The class names are then classified using function mapping. This technique, by default, employs distinct filters with various frames to remove aspect ratio discrepancies, as well as multi-scale feature maps for object recognition.

Jiao et al. [47] developed a new network for object identification, RFP-Net. RFP-Net was the first to apply the RF and eRF concepts to generate bids based on regions. The RF from each sliding window is used as a reference frame in this technique, and the eRF range is used to filter out low-quality phrases. In addition, we developed an eRF-based matching technique to identify positive and negative samples trained by RFP-Net, therefore addressing the imbalance between positive and negative samples as well as the scaling problem in object recognition.

Liang et al. [48] proposed a multi-style attention fusion network (MAFNet). MAFNet, in particular, is made up of a dual signal spatial attention (DSA) module, an attention middle presentation module, and a dual service module (DAIR). He used a multi-level service function merging module and advanced channel attention module (HCA and MLFF). DSA seeks to increase low-level performance while filtering out background noise. DAIR utilizes two branches to adaptively integrate spatial and semantic information from intermediate layer functions. HCA reserves the block's high-level semantic characteristics via two distinct channel operations. The abovementioned multi-level functions are successfully integrated in a trainable manner by MLFF.

Liu et al. [49] presented image processing-based integrated traffic sign recognition. Color-based techniques, shape-based methods, color and shape-based methods, LIDAR, and machine learning are the five primary inspection methods studied in this study. To comprehend and summarize the mechanics of different techniques, the methods in each category are also split into distinct sub-categories. Some of the comparison techniques have been implemented in some updated methods that are not compared in public records.

Pollara et al. [50] described different ways of detecting and monitoring low-cost, low-power devices using certain hydrophones. The ship's acoustic properties were thoroughly examined to establish its physical specifications. These variables can be used to categorize ships. The Stevens Acoustic Library is a collection of acoustic instruments.

Wang et al's. [51] study is broken into two sections: A data collection based on the drone's point of view is developed and a variety of approaches are utilized to detect tiny objects. Through a series of comparative experiments, a machine learning technique based on SVM and a deep learning method based on the YOLO network were effectively constructed. We can see that the SVM-based machine learning method uses less computer resources and saves time. However, due to the selection of the region of interest, it is impossible to enhance accuracy and dependability in some particular scenarios. Deep learning based on neural networks, on the other hand, can give more accuracy.

Xue et al. [52] presented an improved approach for identifying small things, which improves the performance of different scales and integrates contextual semantic information across them. The results of tests on the large MS COCO dataset show that this method can improve the accuracy of small object identification while staying reasonably quick.

Zhiqiang and Jun [53] introduced CNN-based object recognition, CNN structure, features of CNN-based object recognition structure, and methods to improve recognition efficiency. CNN has a powerful feature extraction function, which can make up for the inconvenience caused by using it. Compared with traditional real-time methods, CNN also has more advantages, accuracy, and adaptability, but there is still room for improvement. This can reduce the loss of functional information, make full use of object relationships, and context and fuzzy inference can help computers deal better with issues such as occlusion and low resolution.

Elakkiya et al. [54] gave an idea of how the cervical lesions can be found and categorized. The proposed method used the tiny object identification mechanism to identify the cervical closure from the colposcopy pictures because the cervical cells are much smaller than the uterine cells. The proposed strategy also used Bayesian optimization to optimize the SOD-GAN's hyper parameters, which reduced time complexity and improved performance in terms of efficient classification. The proposed improved SOD-GAN uses eight alternative colposcopy images as inputs and eight randomly generated noise images as outputs to produce the right colposcopy image.

Ji et al. [55] combined the YOLOv4 with two other approaches which are multi-scale contextual information and Soft-CIOU, and called it as MCS-YOLOv4. Extra scales were added to the approach to gain definite data. The authors also encompassed the perception block within the structure of the model.

Sun et al. [56] talked about real time detection of small objects especially for the moving vehicles. The approach was to gain better results from less deeper networks and by assigning the weights to the feature gained in a such a way so as to have better quantifying results (Table 2).

The table above compares several approaches for tiny item identification. In comparison to the preceding approaches, RFP-Net, the object detection technique, employs a receptive field-based proposal generation network, which results in significantly improved accuracy.

4 Studies related to moving object detection

VO detection [57] is the task of detecting VOs instead of images. VO are free-format video clips with semantic meaning. A two-dimensional snapshot of a VO at a certain point in time is called the video object plane (VOP). VOP is determined by its texture (luminance and chroma values) and shape.

4.1 Methods for detecting objects in videos

As seen in Figure 2, VO detectors may be categorized as streaming based on how they use temporal dependencies and aggregate attributes generated from video clips, LSTM [58], due diligence [59], and subsequent detectors. These methods of VO detection are shown schematically in Figure 2 [60].

4.2 Video forgery detection

Video forgery detection (video forensic technique) is a scientific study used to check for alterations in video information. Depending on the degree of change, these changes can be classified within or between tables [61].

Table 2: Comparative study of SOD

Author	Year	Methodology	Advantage	Object type	Accuracy
Xue Zhang, Zheng Wang	2020	Boundary-aware high-resolution network	The conspicuous object is successfully detected	Salient object	—
Fangfang Liang, Lijuan Duan	2020	RGB-D	The capacity of these structures to enhance feature maps is given an edge.	People, cars, building, and animals	—
Ashwani Kumar, Sonam Srivastava	2020	CNN	Multi-box detection at various layers gives different outcomes	Salient object	Achieves 65% accuracy only
Lin Jiao, Shengyu Zhang	2020	RFP-Net	—	Salient object	Largely improving the accuracy
Yanhua Liang, Guihe Qin	2020	MAFNet	To some extent efforts have been made to get worldwide information	People and cars	Improve accuracy while reducing network space redundancy
Chunsheng Liu, Shuang L	2019	Machine vision-based traffic sign detection	—	Salient object	Highest accuracy of 88%
Alexander Pollara, Dr. Alexander Sutin	2017	Small boat detection, tracking, and classification	—	People and cars	—
Jianhang Wang, Sitan Jiang	2019	Comparison of small objects	—		High detection accuracy
Zhijun Xue, Wenjie Chen	2020	Enhancement and fusion of multi-scale feature maps	Increase the network's size	People and cars	Enhance accuracy while minimizing network redundancy
Wang Zhiqiang, Liu Jun	2017	CNN	Reduce the loss of feature information	People, cars, buildings, and animals	Excellent outcomes in terms of accuracy and speed
Elakkiya R, Teja KS, Jegatha Deborah L, Bisogni C, Medaglia C	2021	RCNN	Reduced the complexity and improved the efficiency	Microscope-based object	Highest accuracy of 97.08%
Ji, Shu-Jun, Qing-Hua Ling, and Fei Han	2023	Yolov4	Multi-scale contextual information and Soft-CIOU	Small object	65.7 at AP50
Wei Sun, Liang Dai, Xiaorui Zhang, Pengshuai Chang, Xiaozheng He	2021	RSOD	Traffic monitoring	Small object	52.7 at mAP50

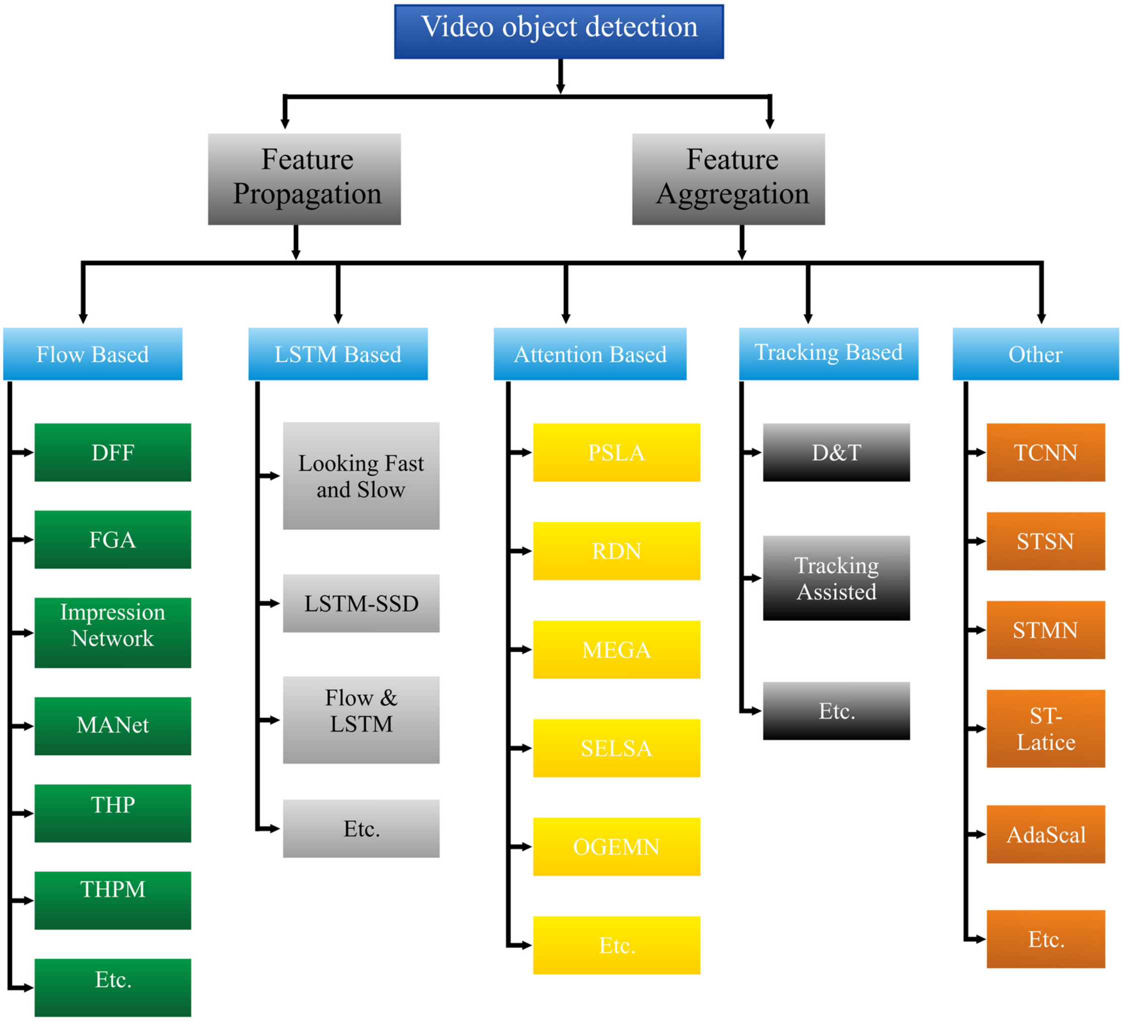

Figure 2: Categories of VO detection.

Depending on the degree of change, these changes can be classified within or between tables based on Spatio-temporal domain [61] (for example, partial change frame). The changes between frames occur in the time domain (that is, the entire frame has undergone the forgery process). The changes between frames occur in the time domain (that is, the entire frame has undergone the forgery process) of the videos, because they are easy and nearly unnoticeable duties. As seen in Figure 3, the purpose of forgeries between frames in surveillance video may be separated into three categories.

- Activity removal: removing the frames in question using frame deletion.
- Activity addition: to introduce a foreign video from some other video, frame insertion is used.
- Activity replication: the process of repeating an event by using frame duplication.

Salvadori et al. [62] reduced the transmission capacity of uncompressed video streams and thereby boosted frame rate using a low-complexity approach based on background removal and error recovery technologies. JPEG is a modern solution. The findings of this study will be taken into account while designing next-generation smart cameras for 6LoWPAN.

Amosov et al. [63] proposed to employ a set of deep neural networks (DNNs) to develop an intelligent context classifier that can recognize and discriminate between regular and critical occurrences in the security service system's continuous video feed. Their artworks are examined by utilizing cutting-edge

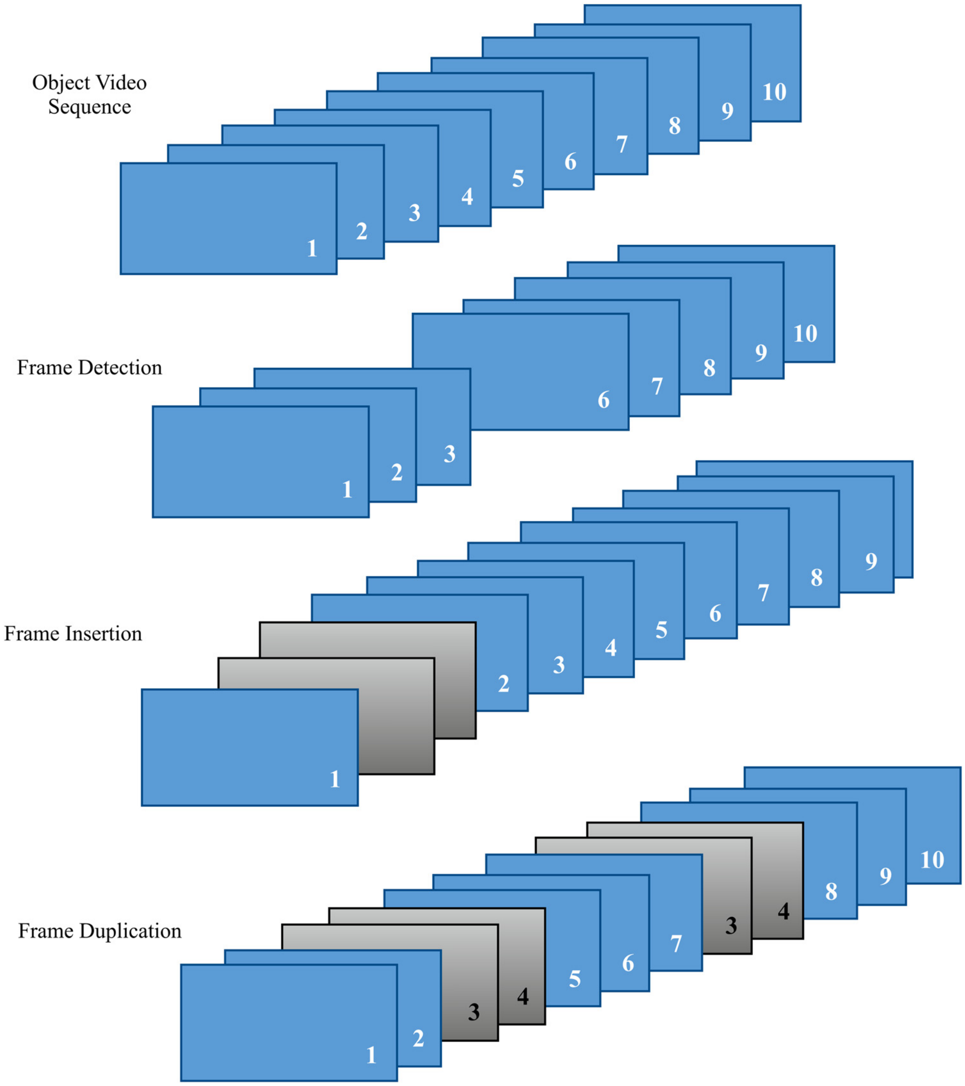

Figure 3: Inter-frame forgeries.

technologies. A probability score for each video segment is the outcome of computer vision and software technologies. To identify and detect normal and abnormal situations, a Python software module was built.

El Kaid et al. [64] proposed a CNN model, which can be used to minimize the false alarm rate, because we can delete 98% of images of someone in a wheelchair, and can more or less reduce false alarms by 17%. However, there are numerous false positives in the blank space image, and none of the evaluated CNN models can identify them owing to the image's complexity. As a result, another concept should be considered in this study to increase the accuracy of the fall detection system.

Najva and Bijoy [65] presented a unique method for detecting and categorizing objects in movies, which uses a tensor function and SIFT to categorize items detected by a DNN. DNN, like the human brain, is capable of analyzing massive quantities of high-dimensional data with billions of variables. The results of this study show that the proposed classifier and most of the existing techniques for feature extraction and classification combine SIFT and tensor features.

Yan and Xu [66] proposed a straight-through pipeline for video caption detection. To recognize video subtitles, the Connected Text Proposal Network (CTPN) is utilized, while the residual network (ResNet), gated recurrent unit (GRU), and connected time classification (CTC) are used to detect Chinese and English subtitles in video pictures. First, use the CTPN technique to determine the subtitle region in the video picture. The identified subtitle range should then be pasted into ResNet to extract the function sequence. Then, add a bidirectional GRU layer to represent the feature sequence.

Wu et al. [67] proposed a straight-through pipeline to detect video captions. To recognize video subtitles, the CTPN is utilized, while the ResNet, GRU, and CTC are used to detect Chinese and English subtitles in video pictures. To begin, identify the subtitle region in the video picture using the CTPN technique. After determining the subtitle range, use ResNet to extract the function sequence. After that, add a bidirectional GRU layer to represent the feature sequence.

Fang et al. [68] introduced a Deep Video Saliency Network (DevsNet), a new deep learning platform with which the meaning of video streams can be determined. DevsNet is primarily made up of two parts: 3D convolutional network (3D-ConvNet) and bidirectional long-term and short-term memory convolutional networks. (BConvLSTM). 3D-ConvNet aims to examine short-term spatio–temporal information, while B-ConvLSTM examines long-term spatio–temporal attributes.

Wang et al. [69] proposed a completely scalable network with a communication structure for high-precision VO recognition and cost-effective computation. The scale recognition module, in particular, is added to acquire characteristics with bigger alterations. The ROI structure module retrieves and combines RoI's location and context functions. Feature aggregation is also used to improve the performance of the reference frame by deforming the flow. SCNet's efficacy has been demonstrated through several trials. In our RoI module, you may add another auxiliary branch with a paired structure for invoking RoI functions, similar to the local function block in BConvLSTM. In addition, SCNet now mainly controls accuracy, so there is still a lot of room for speed improvement.

Zhu and Yan [70] proposed traffic sign recognition using YOLOv5 and compared with SSD with some extended features (Table 3).

The above comparison table represents some moving object detection techniques.

5 Studies related to loss function

In object recognition tasks, the loss function is the most important element in determining identification accuracy. First, the connection between location and classification is established by multiplying the factor based on IoU by the classification loss function's typical cross-entropy loss [71]. The square mistake represented by the root (MSE) [72] is the main force of the basic loss function. It is simple to comprehend and apply, and it works effectively in most cases. Take the difference between the forecast and the ground truth, blockage, and the average of the whole dataset to compute the MSE. In statistics, the loss function is frequently used to estimate parameters, and the event in question is a function of the difference between the estimated and true values of the data instance. Abraham Wald reintroduced statistics in the middle of the 20th century, reintroducing this concept is as old as Laplace [73,74]. For example, in an economic context, this is usually economic loss or regret. In classification, this is the penalty for misclassifying the example. In actuarial science, especially after Harald Kramer's work in the 1920s, it is used in the insurance industry to model premium payment models. The model manages the Loss which is the price of not meeting expectations, in the best way. Loss is the price of not meeting expectations. In financial risk management, this function is allocated to monetary loss [75–77]. Some important studies covering the more important objective-based loss function research are discussed below.

Fang et al. [78] proposed a hostile network based on conditional patches, which uses a generator network based on sampled data patches and a conditional discriminator network with additional loss functions to check fine blood vessels and coarse data. Experiments will be conducted on the public STARE and DRIVE datasets, showing that the proposed model is superior to more advanced methods.

Fan and Liu [79] investigated GAN training with various combination techniques and discovered that synchronization of the discriminator and generator between clients offers the best outcomes for two distinct challenges. The study also discovered empirical results indicating that federated learning is typically resilient for the number of consumers having IID learning data and modest non-IID learning data. However,

Table 3: Comparative study of moving object detection

Author	Year	Methodology	Advantage	Accuracy
Claudio Salvadori, Matteo Petracca	2013	Video streaming 6LoWPAN	Reduce the number of discarded packets	—
O.S Amosov, S.G. Amosova	2019	Ensemble of DNN stream of the security system	—	Accuracy 96% for 80 epochs
Amal EL KAID, Karim BA¨ INA	2019	Video detection algorithm by CNN	—	Improve its accuracy
Najva N, Edet Bijoy K	2016	SIFT and DNN	The major benefit of SIFT is that they are resistant to distortion, noise addition, and changes in light	It requires high accuracy
Hongyu Yan, Xin Xu	2020	DRNN	Its advantage is more flexible	89.2% recognition accuracy
Peng Wu, Jing Liu	2020	Fast Sparse Coding Networks	Encoding-decoding neural networks have several advantages, including efficient inference	Higher accuracy
Yuming Fang, Chi Zhang,	2020	DevsNet	—	—
Fengchao Wang, Zhewei Xu	2020	SCNet	—	High-precision and low-cost calculation
Yanzhao Zhu	2022	Traffic sign recognition by CNN	—	Improve its accuracy

if the data distribution is significantly skewed, the existing compound learning scheme (such as FedAvg) would be anomalous owing to the weight difference.

Liu et al. [80] proposed a model based on a two-layer backbone architecture, it provides end-to-end pose estimation at the 6D category level to detect bounding boxes. In this scenario, the 6D posture is created straight from the network and ensures that no further steps or post-processing are needed, such as Perspective-n-point. Our loss function and CNN's two-layer architecture make collaborative multi-task learning quick and effective. This study increases posture estimation accuracy by substituting completely linked layers with fully folded layers. Transform your pose estimation challenge into a classification and regression problem with the aid of our network, which are termed as Pose-cls and Pose-reg.

Sharma and Mir [81] developed a unique technique for segmenting VOs using unsupervised learning. The process is divided into two stages, each of which considers the basic frame and the current frame for segmentation. We build dense region clauses, bounding boxes, and scores in the first step. Following that, we develop a feature extraction technique that utilizes the attention network for feature encoding. Finally, using the Softmax technique, these functions are scaled and combined to generate object segmentation.

Liu et al. [82] proposed a continuous deep network based on mixed sampling and mixed loss computation to detect salient items. Not only the hybrid sampling may integrate original and sample features but it can also acquire a wider receiving field using horrible convolution. The hybrid loss function, which combines cross-entropy loss and area loss, can further minimize the gap between the salient map and the terrain's realism. A fully linked CRF model might be used to increase spatial coherence and contour placement even further.

Steno et al. [83] attempted to enhance the accuracy of threat localization and minimize detection time by employing a quicker and better R-CNN (with a suggested network divided by region). The planned network by area has been modified to make it simpler to discover things using the new docking box design. Improved RPN can give a more comprehensive summary of characteristics. Furthermore, by including sample weights into the classification loss function, an enhanced cross-entropy function is created, which improves the classification deficit and the multi-task loss function's performance. In MATLAB, the average accuracy is improved to 0.27, the average processing time is lowered, and the average processing time is increased by 0.27.

Gu et al. [84] proposed better lightweight detection using Context Aware Dense Feature Distillation. And use rich contextual feature for SOD (Table 4).

The above comparison table represents some loss functions and their calculation techniques. Compared to the above techniques federated generative adversarial learning produces a higher accuracy and has the advantage of accurate trajectory prediction with few attempts.

6 Studies related to optimization technique

In the network, optimization methods are employed to minimize a function known as the loss function or error function. The optimization approach may generate the smallest difference between the actual output and the predicted output by minimizing the loss function, allowing our model to accomplish the task more correctly.

Dumitru et al. [85] suggested an edge detector, which was compared against one of the most sophisticated techniques, the "Tricky Edge" detector. Our edge detection methodology combines particle swarm optimization with monitored optimization of cellular machine rules. We developed transferable rules that may be used for a variety of pictures with comparable features. On average, the recommended approach outperforms Canny in our advanced dataset.

Huang et al. [25] proposed a model for detecting prominent items in hyperspectral pictures on wireless networks, which employs visibility optimization to the characteristics of CNN. To define the ultimate melting behavior, to extract spatial and spectral characteristics of the same size, we first use a CNN with two channels. By maximizing the bump values of the foreground and background signals from the CNN

Table 4: Comparative study of loss function

Author	Year	Methodology	Advantage	Accuracy
Yunchun Fang, Yilu Cao, Wei Zhang, Quilong Yuan	2019	Dual networks for attribute prediction	—	Improve its accuracy
Chenyou Fan, Ping Liu	2020	Federated generative adversarial learning	The capacity to forecast a trajectory with high accuracy in a limited number of trials	90% accuracy
Fuchang Liu, Pengfei Fang	2018	Recovering 6D object pose from RGB	—	Achieve increased accuracy
Vipul Sharma, Roohie Naaz Mir	2020	SSFNET-VOS	—	Accuracy is improved while computing complexity is kept to a minimum
Zhengui Liu, Jiting Tang, Peng Zhao	2019	Hybrid upsampling and hybrid loss computing	—	Less accuracy
Priscilla Steno, Abeer Alsadoon	2020	DNN	—	Improve the accuracy
Lingyun Gu	2023	Context-aware Dense Feature Distillation (CDFD)	—	—

Table 5: Comparative study of optimization technique

Author	Methodology	Advantage	Sensitivity	Specificity	Precision	Recall	*F*-measure	Accuracy
Delia Dumitru, Anca Andreica	Cellular automata rules optimized	Capable of discovering a specific rule for a picture or group of images	—	—	Good	Medium	Medium	Good
Chen Huang, Tingfa Xu	CNN and saliency optimization	Less noise	Good	Good	—	—	Medium	Good
J. Sasikala, D. S. Juliet	The optimized hybrid adaptive cuckoo search algorithm	—	—	Good	Good	Good	Good	Good
Lokesh Jain, Rahul Katarya	Whale optimization algorithm	Better performance	Good	—	Medium	Good	Medium	Medium
D. Rammurthy, P. K. Mahesh	WHOA-DNN	—	Good	Good	—	—	Medium	Good
Yun Zhang, Yongguo Liu	WOCDA	—	Good	Good	—	—	Good	Good
Delia Dumitru, Anca Andreica	R-CNN with NAS optimization	—	Good	Good	—	—	Good	Good

characteristics, the final bump map is generated. The findings of this study show that the approach is effective and performs well in the creation of hyper spectral pictures.

Sasikala et al. [86] used a classifier in conjunction with an optimal model. Even with hundreds of blood vessel pictures, this experimental model outperforms previous detection techniques. This hybrid and adaptive optimization approach based on rhododendron search produces the greatest results in dynamic regions affected by the ocean, and the findings indicate a reduction in the false alarm rate of ports and other coastal surveillance locations.

Jain et al. [87] presented a novel social media-based whale optimization algorithm for identifying N thought leaders by analyzing user reputation using various popular Internet optimization functions. The approach is effective for identifying opinion leaders since it is based on humpback whale hunting behavior with bubble nets. As the number of users on the network grew, the algorithm determined the optimal option. As a consequence, the method's total complexity remains constant. We also offered a novel community classification method based on the similarity index, which contains the clustering coefficient and the similarity of neighbors as important components. Local and worldwide opinion leaders were identified by using priorities and recommended methods and optimization features. We applied the suggested method to real-world and large-scale datasets and compared the outcomes in terms of precision, accuracy, recall, and *F*1 score.

Rammurthy and Mahesh [88] recommended the Whale Harris Hawks Optimization (WHHO) technique to identify brain cancers using magnetic resonance imaging. For segmentation, we employed cellular automata and approximation set theory. Furthermore, characteristics such as tumor size, local optical orientation pattern, mean, variance, and kurtosis are retrieved from sections. Furthermore, brain tumor identification is performed using a deep CNN, while training is performed utilizing the suggested WHHO. The Whale Optimization Algorithm and the Harris Hawks Optimization Algorithm were combined (HHO). According to WHHO, deep CNN recommends utilizing alternative techniques with a maximum accuracy of 0.816, a maximum specificity of 0.791, and a maximum sensitivity of 0.974.

Zhang et al. [89] proposed the community detection based on whale optimization (WOCDA) method as a novel community discovery technique. WOCDA's initialization strategy and three optimization operations simulate humpback whale hunting behavior and determine the community in experiments of synthetic and real networks, demonstrating that the community ratio algorithm identified by WOCDA can be detected in modern meta-heuristics in most cases. WOCDA's efficacy, however, declines as the number of nodes in the network grows, because the random search process takes a long time until a big search space is reached.

Luo et al. [90] suggested a unique multi-scale and target vehicle recognition approach for identifying complex vehicles in natural situations. We improve the image of the dataset by utilizing the Retinex-based adaptive image correction approach to reduce the influence of shadows and highlights. This study describes a multi-layer feature extraction approach that explores the neural architecture for the best connection between layers, increasing the representation of the fundamental properties of the quicker R-CNN model and aims to analyze performance of multi-scale vehicles. We provide a target feature enhancement approach that integrates multi-layer feature information and context information from the final layer after the layers are connected to enrich the target information and improve the model's reliability in recognizing big and small targets (Table 5).

The above comparison table represents some optimization techniques. Compared to the above optimization techniques, salient object identification on hyperspectral pictures in wireless networks utilizing CNN and saliency optimization results in improved accuracy and efficiency, as well as the benefit of fewer noise.

7 Conclusion

This study reviewed different small object and movement detection, loss functions, and optimization techniques. This approach is used to increase the small object in addition to movement detection with new ideas. In this study, there are 84 research articles with the same background as this article. Articles were selected from various journals. Through the overview and reference section of the previous research articles, individual articles were selected to study the previous literature. The selected research supports the detection of smaller moving objects through performance analysis, loss functions, and optimization techniques. After careful analysis of the previous work, some landmark articles were selected for research, which may be useful for this research.

8 Future scope

Over the past few years, the communities of computer vision and pattern recognition have paid a lot of attention to object detection in images and videos. Although we have created numerous ways for detecting objects, deep learning applications promise greater accuracy for a wider range of object types. In future, we would like to implement and compare models for aerial images and video frames. Also, there is a need for certain methods which would not only detect the objects but also analyze them for further investigations. It will be crucial to use this remarkable computer technology, which is related to computer vision and image processing that recognizes and characterizes items from digital images and videos, such as people, cars, and animals.

Author contributions: Ravi Prakash Chaturvedi collected, filtered, organized, compared and worked upon the data. Udayan Ghose validated and analyzed the results. He also audited the approach and results.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: Data was collected from various research papers that are already mentioned as references in paper.

References

[1] Dasiopoulou S, Mezaris V, Kompatsiaris I, Papastathis VK, Strintzis MG. Knowledge-assisted semantic video object detection. IEEE Trans Circuits Syst Video Technol. 2005;15(10):1210–24.

[2] Meng Q, Song H, Li G, Zhang Y, Zhang X. A block object detection method based on feature fusion networks for autonomous vehicles. Complexity. Feb. 2019;2019:1–14.

[3] Guan L, He Y, Kung SY. Multimedia image and video processing. CRC Press; 1 March 2012. p. 331. ISBN 978-1-4398-3087-1.

[4] Wu J, Osuntogun A, Choudhury T, Philipose M, Rehg JM. A scalable approach to activity recognition based on object use. 2007 IEEE 11th International Conference on computer Vision. IEEE; 2007.

[5] Hassan T, Akram MU, Hassan B, Nasim A, Bazaz SA. Review of OCT and fundus images for detection of Macular Edema. In 2015 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE; 2015, September. p. 1–4.

[6] Bagci AM, Ansari R, Shahidi M. A method for detection of retinal layers by optical coherence tomography image segmentation. In 2007 IEEE/NIH Life Science Systems and Applications Workshop. IEEE; 2007, November. p. 144–7.

[7] Dong C, Liu J, Xu F. Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor. Remote Sens. 2018;10:400.

[8] Yang X, Sun H, Fu K, Yang J, Sun X, Yan M, et al. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. Remote Sens. 2018;10:132.

[9] Girshick R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: 7–13 December 2015. p. 1440–8.

[10] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada. Cambridge, MA, USA: MIT Press; 7–12 December 2015. p. 91–9.

[11] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In European Conference on Computer Vision, Proceedings of the 13th European Conference, Zurich, Switzerland. Cham, Switzerland: Springer; 6–12 September 2014. p. 346–61.

[12] Xu F, Liu J, Sun M, Zeng D, Wang X. A hierarchical maritime target detection method for optical remote sensing imagery. Remote Sens. 2017;9:280.

[13] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: 23–28 June 2014. p. 580–7.

[14] Chen X, Kundu K, Zhu Y, Berneshawi AG, Ma H, Fidler S, et al. 3D object proposals for accurate object class detection. Lect Notes Bus Inf Process. 2015;122:34–45.

[15] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016. Cham, Switzerland: Springer; 2016. p. 21–37.

[16] Li H, Lin Z, Shen X, Brandt J, Hua G. A convolutional neural network cascade for face detection. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: 7–12 June 2015. p. 5325–34.

[17] Yang F, Choi W, Lin Y. Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: 27–30 June 2016. p. 2129–37.

[18] Bateni S, Wang Z, Zhu Y, Hu Y, Liu C. Co-optimizing performance and memory footprint via integrated CPU/GPU memory management, an implementation on autonomous driving platform. In Proceedings of the 2020 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS). Sydney, Australia: 21–24 April 2020.

[19] Lu J, Tang S, Wang J, Zhu H, Wang Y. A review on object detection based on deep convolutional neural networks for autonomous driving. In Proceedings of the 2019 Chinese Control and Decision Conference (CCDC). Nanchang, China: 3–5 June 2019.

[20] Wei H, Laszewski M, Kehtarnavaz N. Deep learning-based person detection and classification for far field video surveillance. In Proceedings of the 2018 IEEE 13th Dallas Circuits and Systems Conference. Dallas, TX, USA: November 2018.

[21] Guillermo M, Tobias RR, De Jesus LC, Billones RK, Sybingco E, Dadios EP, et al. Detection and classification of public security threats in the philippines using neural networks. In Proceedings of the 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech). Kyoto, Japan: 10–12 March 2020. p. 1–4.

[22] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211–52.

[23] Chen G, Wang H, Chen K, Li Z, Song Z, Liu Y, et al. A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal. IEEE Transactions on Systems, Man, and Cybernetics: Systems; 2020 Jul 17.

[24] Ren Y, Zhu C, Xiao S. Small object detection in optical remote sensing images via modified faster R-CNN. Appl Sci. 2018 May;8(5):813.

[25] Huang C, Xu T, Zhang Y, Pan C, Hao J, Li X. Salient object detection on hyperspectral images in wireless network using CNN and saliency optimization. Ad Hoc Netw. 2021 Mar 1;112:102369.

[26] Hua X, Wang X, Rui T, Zhang H, Wang D. A fast self-attention cascaded network for object detection in large scene remote sensing images. Appl Soft Comput. 2020 Sep 1;94:106495.

[27] Yundong LI, Han DO, Hongguang LI, Zhang X, Zhang B, Zhifeng XI. Multi-block SSD based on small object detection for UAV railway scene surveillance. Chin J Aeronautics. 2020 Jun 1;33(6):1747–55.

[28] Bosquet B, Mucientes M, Brea VM. STDnet: Exploiting high resolution feature maps for small object detection. Eng Appl Artif Intell. 2020 May 1;91:103615.

[29] Kunfu K, Chang Z, Zhang Y, Xu G, Zhang K, Sun X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS J Photogramm Remote Sens. 2020 Mar 1;161:294–308.

[30] Zheng Z, Zhong Y, Ma A, Han X, Zhao J, Liu Y, et al. HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. ISPRS J Photogramm Remote Sens. 2020 Aug 1;166:1–4.

[31] Tian Y, Wang K, Wang Y, Tian Y, Wang Z, Wang FY. Adaptive and azimuth-aware fusion network of multimodal local features for 3D object detection. Neurocomputing. 2020 Oct 21;411:32–44.

[32] Li C, Cong R, Guo C, Li H, Zhang C, Zheng F, et al. A parallel down-up fusion network for salient object detection in optical remote sensing images. Neurocomputing. 2020 Nov 20;415:411–20.

[33] Fadl S, Han Q, Li Q. CNN spatiotemporal features and fusion for surveillance video forgery detection. Signal Processing Image Commun. 2021 Jan;90:116066.

[34] Zhu H, Wei H, Li B, Yuan X, Kehtarnavaz N. A review of video object detection: datasets, metrics and methods. Appl Sci. 2020 Jan;10(21):7834.

[35] Alhimale L, Zedan H, Al-Bayatti A. The implementation of an intelligent and video-based fall detection system using a neural network. Appl Soft Comput. 2014 May 1;18:59–69.

[36] Lee MH, Yoo HW, Jang DS. Video scene change detection using neural network: Improved ART2. Expert Syst Appl. 2006 Jul 1;31(1):13–25.

[37] Kousik N, Natarajan Y, Raja RA, Kallam S, Patan R, Gandomi AH. Improved salient object detection using hybrid convolution recurrent neural network. Expert Syst Appl. 2021;166:114064.

[38] Xu G, Zhang Y, Zhang Q, Lin G, Wang Z, Jia Y, et al. Video smoke detection based on deep saliency network. Fire Saf J. 2019 Apr 1;105:277–85.

[39] Yang A, Liu H, Chen Y, Zhang C, Yang K. Digital video intrusion intelligent detection method based on narrowband Internet of Things and its application. Image Vis Comput. 2020 May 1;97:103914.

[40] Yamazaki Y, Kanaji S, Matsuda T, Oshikiri T, Nakamura T, Suzuki S, et al. Automated surgical instrument detection from laparoscopic gastrectomy video images using an open source convolutional neural network platform. J Am Coll Surg. 2020 May 1;230(5):725–32.

[41] Yue X, Li H, Shimizu M, Kawamura S, Meng L. YOLO-GD: a deep learning-based object detection algorithm for empty-dish recycling robots. Machines. 2022;10(5):294.

[42] Hu GX, Yang Z, Hu L, Huang L, Han JM. Small object detection with multiscale features. Int J Digital Multimed Broadcasting. 2018 Sep 30;2018.

[43] Ren Y, Zhu C, Xiao S. Small object detection in optical remote sensing images via modifed faster R-CNN. Appl Sci. 2018;8(5):813.

[44] Zhang X, Wang Z, Hu Q, Ren J, Sun M. Boundary-aware High-resolution Network with region enhancement for salient object detection. Neurocomputing. 2020 Dec 22;418:91–101.

[45] Liang F, Duan L, Ma W, Qiao Y, Miao J, Ye Q. Context-aware network for RGB-D salient object detection. Pattern Recognit. 2021;111:107630.

[46] Kumar A, Srivastava S. Object detection system based on convolution neural networks using single shot multi-box detector. Procedia Comput Sci. 2020 Jan 1;171:2610–7.

[47] Jiao L, Zhang S, Dong S, Wang H. RFP-Net: Receptive field-based proposal generation network for object detection. Neurocomputing. 2020 Sep 10;405:138–48.

[48] Liang Y, Qin G, Sun M, Yan J, Jiang H. MAFNet: Multi-style attention fusion network for salient object detection. Neurocomputing. 2021 Jan;422:22–33.

[49] Liu C, Li S, Chang F, Wang Y. Machine vision based traffic sign detection methods: Review, analyses and perspectives. IEEE Access. 2019 Jun 26;7:86578–96.

[50] Pollara A, Sutin A, Salloum H. Passive acoustic methods of small boat detection, tracking and classification. In 2017 IEEE International Symposium on Technologies for Homeland Security (HST). IEEE; 2017 Apr 25. p. 1–6.

[51] Wang J, Jiang S, Song W, Yang Y. A comparative study of small object detection algorithms. In 2019 Chinese Control Conference (CCC). IEEE; 2019 Jul 27. p. 8507–12.

[52] Xue Z, Chen W, Li J. Enhancement and fusion of multi-scale feature maps for small object detection. In 2020 39th Chinese Control Conference (CCC). IEEE; 2020 Jul 27. p. 7212–7.

[53] Zhiqiang W, Jun L. A review of object detection based on convolutional neural network. In 2017 36th Chinese Control Conference (CCC). IEEE; 2017 Jul 26. p. 11104–9.

[54] Elakkiya R, Teja KS, Jegatha Deborah L, Bisogni C, Medaglia C. Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks. Multimed Tools Appl. 2022;81:191–207.

[55] Ji SJ, Ling QH, Han F. An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information. Comput Electr Eng. 2023;105:108490.

[56] Sun W, Dai L, Zhang X, Chang P, He X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. Appl Intell. 2021;1–16.

[57] Wu H, Chen Y, Wang N, Zhang Z. Sequence level semantics aggregation for video object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Gangnam-gu, Seoul, Korea: 27 October–2 November 2019.

[58] Lu Y, Lu C, Tang C-K. Online video object detection using association LSTM. In Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: 22–29 October 2017. p. 2363–71.

[59] Chen Y, Cao Y, Hu H, Wang L. Memory enhanced global-local aggregation for video object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: June 2020. p. 16–8.

[60] IEEE International Conference on Multimedia and Expo. Shanghai, China: 8–12 July 2019. p. 1750–5.

[61] Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE; 2018 Dec 11. p. 1–7.

[62] Salvadori C, Petracca M, Madeo S, Bocchino S, Pagano P. Video streaming applications in wireless camera networks: A change detection based approach targeted to 6LoWPAN. J Syst Architecture. 2013 Nov 1;59(10):859–69.

[63] Amosov OS, Amosova SG, Ivanov YS, Zhiganov SV. Using the ensemble of deep neural networks for normal and abnormal situations detection and recognition in the continuous video stream of the security system. Procedia Comput Sci. 2019 Jan 1;150:532–9.

[64] El Kaid A, Baïna K, Baïna J. Reduce false positive alerts for elderly person fall video-detection algorithm by convolutional neural network model. Procedia Comput Sci. 2019 Jan 1;148:2–11.

[65] Najva N, Bijoy KE. SIFT and tensor based object detection and classification in videos using deep neural networks. Procedia Comput Sci. 2016 Jan 1;93:351–8.

[66] Yan H, Xu X. End-to-end video subtitle recognition via a deep residual neural network. Pattern Recognit Lett. 2020 Mar 1;131:368–75.

[67] Wu P, Liu J, Li M, Sun Y, Shen F. Fast sparse coding networks for anomaly detection in videos. Pattern Recognit. 2020 Nov 1;107:107515.

[68] Fang Y, Zhang C, Min X, Huang H, Yi Y, Zhai G, et al. DevsNet: Deep video saliency network using short-term and long-term cues. Pattern Recognit. 2020 Jul 1;103:107294.

[69] Wang F, Xu Z, Gan Y, Vong CM, Liu Q. SCNet: Scale-aware coupling-structure network for efficient video object detection. Neuro Comput. 2020 Sep 3;404:283–93.

[70] Zhu Y, Yan WQ. Traffic sign recognition based on deep learning. Multimed Tools Appl. 2022;81(13):17779–91.

[71] Hou S, Wang C, Quan W, Jiang J, Yan DM. Text-aware single image specular highlight removal. In Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4. Springer International Publishing; 2021. p. 115–27.

[72] Temlioglu E, Erer I, Kumlu D. A least mean square approach to buried object detection in ground penetrating radar. In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2017 Jul 23. p. 4833–6.

[73] Wald A. Statistical decision functions. Wiley; 1950.

[74] Cramér CH. On the mathematical theory of risk. Centraltryckeriet, Stockholm: Forsakringsaktiebolaget Skandias Festskrift; 1930.

[75] Hermans A, Beyer L, Leibe B In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737; 2017.

[76] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: Leibe B, Matas J, Sebe N, Welling M, (eds.). ECCV 2016. LNCS. 9911, Cham: Springer; 2016. p. 499–515. doi: 10.1007/978-3-319-46478-731.

[77] Chaturvedi RP, Ghose U. Small object detection using retinanet with hybrid anchor box hyper tuning using interface of Bayesian mathematics. J Inf Optim Sci. 2022;43(8):2099–110.

[78] Fang Y, Cao Y, Zhang W, Yuan Q. Enhance feature representation of dual networks for attribute prediction. In International Conference on Neural Information Processing. Cham: Springer; 2019 Dec 12. p. 13–20.

[79] Fan C, Liu P. Federated generative adversarial learning. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer; 2020 Oct 16. p. 3–15.

[80] Liu F, Fang P, Yao Z, Fan R, Pan Z, Sheng W, et al. Recovering 6D object pose from RGB indoor image based on two-stage detection network with multi-task loss. Neurocomputing. 2019 Apr 14;337:15–23.

[81] Sharma V, Mir RN. SSFNET-VOS: Semantic segmentation and fusion network for video object segmentation. Pattern Recognit Lett. 2020 Dec 1;140:49–58.

[82] Liu Z, Tang J, Zhao P. Salient object detection via hybrid upsampling and hybrid loss computing. Vis Computer. 2019;36(4):843–53.

[83] Steno P, Alsadoon A, Prasad PW, Al-Dala'in T, Alsadoon OH. A novel enhanced region proposal network and modified loss function: threat object detection in secure screening using deep learning. J Supercomputing. 2021 Apr;77(4):3840–69.

[84] Gu L, Fang Q, Wang Z, Popov E, Dong G. Learning lightweight and superior detectors with feature distillation for onboard remote sensing object detection. Remote Sens. 2023;15(2):370.

[85] Dumitru D, Andreica A, Dioşan L, Bálint Z. Robustness analysis of transferable cellular automata rules optimized for edge detection. Procedia Comput Sci. 2020 Jan 1;176:713–22.

[86] Sasikala J, Juliet DS. Optimized vessel detection in marine environment using hybrid adaptive cuckoo search algorithm. Comput Electr Eng. 2019 Sep 1;78:482–92.

[87] Jain L, Katarya R, Sachdeva S. Opinion leader detection using whale optimization algorithm in online social network. Expert Syst Appl. 2020 Mar 15;142:113016.

[88] Rammurthy D, Mahesh PK. Whale Harris hawks Optimization based deep learning classifier for brain tumor detection using MRI images. J King Saud University-Comput Inf Sci. 2022;34(6):3259–72.

[89] Zhang Y, Liu Y, Li J, Zhu J, Yang C, Yang W, et al. WOCDA: A whale optimization based community detection algorithm. Phys A: Stat Mech Appl. 2020 Feb 1;539:122937.

[90] Luo JQ, Fang HS, Shao FM, Zhong Y, Hua X. Multi-scale traffic vehicle detection based on faster R-CNN with NAS optimization and feature enrichment. Def Technol. 2022;17(4):1542–54.