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]
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.](image-url)
<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Methodology</th>
<th>Advantage</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guang Chen, Zida Song</td>
<td>2020</td>
<td>DL-based SOD</td>
<td>The network benefits from integrating multi-resolution context information from available modules rather than directly from a single layer, resulting in more efficient computing.</td>
<td>Accuracy could be significantly improved.</td>
</tr>
<tr>
<td>Yun Ren, Changren Zhu</td>
<td>2018</td>
<td>Modified faster R-CNN method</td>
<td>Information extended up to some extent. MER gather more region than bounding box method.</td>
<td>—</td>
</tr>
<tr>
<td>Chen Huang, Tingfa Xu</td>
<td>2020</td>
<td>Hyper spectral images in wireless network</td>
<td>Less noise.</td>
<td>Improves its accuracy and efficiency.</td>
</tr>
<tr>
<td>Xia Hua, Xinqing Wang</td>
<td>2020</td>
<td>A cascaded CNN</td>
<td>SSD offers two benefits: real-time processing and excellent precision.</td>
<td>Has high accuracy and adaptability.</td>
</tr>
<tr>
<td>Yundong Li, Han DONG</td>
<td>2019</td>
<td>Multi-block SSD</td>
<td>The advantage of this technique is that it makes use of point cloud data and the fusion style.</td>
<td>The SSD technique yields an overall accuracy of 96.6%.</td>
</tr>
<tr>
<td>Brais Bosquet, Manuel Mucientes</td>
<td>2020</td>
<td>STDnet</td>
<td>Enhancements to speed and memory optimization.</td>
<td>Small-item detection accuracy using CNNs falls behind other bigger things.</td>
</tr>
<tr>
<td>Kun Fu, Zhonghan Chang</td>
<td>2020</td>
<td>Rotation-aware and multi-scale CNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhuo Zheng, Yanfei Zhong</td>
<td>2020</td>
<td>HyNet framework</td>
<td>Improves the scale-invariant properties, making it ideal for object recognition in large-scale MSR remote sensing pictures.</td>
<td>Improves multi-scale item recognition accuracy in HSR remote sensing images.</td>
</tr>
<tr>
<td>Yonglin Tian, Kunfeng Wang</td>
<td>2020</td>
<td>Adaptive azimuth-aware fusion network</td>
<td>The advantage of this technique is that it makes use of point cloud data and the fusion style.</td>
<td>—</td>
</tr>
<tr>
<td>Chongyi Li, Runmin Cong</td>
<td>2020</td>
<td>A parallel down-up fusion network</td>
<td>In optical RSIs, the combination of various resolutions has a benefit in addressing size variation and seeing.</td>
<td>Increases the generalization capabilities and accuracy of networks.</td>
</tr>
<tr>
<td>Sondos Fadl, Qi Han</td>
<td>2020</td>
<td>CNN spatiotemporal features and fusion</td>
<td>Reduces the number of post-processing steps.</td>
<td>CNN system achieves high accuracy.</td>
</tr>
<tr>
<td>Laila Alhimal, Hussein Zedan</td>
<td>2014</td>
<td>video-based fall detection system using a NN</td>
<td>A trustworthy, accurate, user-friendly, and low-cost fall detection system that protects the privacy.</td>
<td>High sensitivity in detecting falls (greater than 90%).</td>
</tr>
<tr>
<td>Man-Hee Lee, Hun-Woo Yoo</td>
<td>2006</td>
<td>Improved ART2</td>
<td>The benefit of quick detection since no decompression time is required.</td>
<td>This method has a high degree of accuracy.</td>
</tr>
<tr>
<td>Nallyanna V, Kousik</td>
<td>2020</td>
<td>Hybrid CRNN</td>
<td>—</td>
<td>The proposed approach is successful in terms of accuracy as well as speed.</td>
</tr>
<tr>
<td>Gao Xu, Yongming Zhang</td>
<td>2019</td>
<td>Video smoke detection based on DSNN</td>
<td>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.</td>
<td>Achieves the best result in terms of existing forecast accuracy.</td>
</tr>
<tr>
<td>Aimin Yang, Huixiang Liu</td>
<td>2020</td>
<td>A digital video intrusion detection method</td>
<td>Consistent performance, minimal power usage, and dependable data.</td>
<td>Low accuracy.</td>
</tr>
<tr>
<td>Yuta Yamazaki, MD, Shingo Kanaji</td>
<td>2020</td>
<td>Automated surgical instrument detection</td>
<td>Complex image analysis for medical professionals.</td>
<td>—</td>
</tr>
<tr>
<td>Xuebin Yue</td>
<td>2022</td>
<td>YOLO-GD (GhostNet and depth wise convolution)</td>
<td>Function of detection dish will lead to further development.</td>
<td>Higher accuracy.</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Methodology</th>
<th>Advantage</th>
<th>Object type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xue Zhang, Zheng Wang</td>
<td>2020</td>
<td>Boundary-aware high-resolution network</td>
<td>The conspicuous object is successfully detected</td>
<td>Salient object</td>
<td>—</td>
</tr>
<tr>
<td>Fangfang Liang, Lijuan Duan</td>
<td>2020</td>
<td>RGB-D</td>
<td>The capacity of these structures to enhance feature maps is given an edge.</td>
<td>People, cars, building, and animals</td>
<td>—</td>
</tr>
<tr>
<td>Ashwani Kumar, Sonam Srivastava</td>
<td>2020</td>
<td>CNN</td>
<td>Multi-box detection at various layers gives different outcomes</td>
<td>Salient object</td>
<td>Achieves 65% accuracy only</td>
</tr>
<tr>
<td>Lin Jiao, Shengyu Zhang</td>
<td>2020</td>
<td>RFP-Net</td>
<td>—</td>
<td>Salient object</td>
<td>Largely improving the accuracy</td>
</tr>
<tr>
<td>Yanhua Liang, Guihe Qin</td>
<td>2020</td>
<td>MAFNet</td>
<td>To some extent efforts have been made to get worldwide information</td>
<td>People and cars</td>
<td>Improve accuracy while reducing network space redundancy</td>
</tr>
<tr>
<td>Chunsheng Liu, Shuang L</td>
<td>2019</td>
<td>Machine vision-based traffic sign detection</td>
<td>—</td>
<td>Salient object</td>
<td>Highest accuracy of 88%</td>
</tr>
<tr>
<td>Alexander Pollara, Dr. Alexander Sutin</td>
<td>2017</td>
<td>Small boat detection, tracking, and classification</td>
<td>—</td>
<td>People and cars</td>
<td>—</td>
</tr>
<tr>
<td>Jianhang Wang, Sitan Jiang</td>
<td>2019</td>
<td>Comparison of small objects</td>
<td>—</td>
<td>People and cars</td>
<td>High detection accuracy</td>
</tr>
<tr>
<td>Zhijun Xue, Wenjie Chen</td>
<td>2020</td>
<td>Enhancement and fusion of multi-scale feature maps</td>
<td>Increase the network’s size</td>
<td>People and cars</td>
<td>Enhance accuracy while minimizing network redundancy</td>
</tr>
<tr>
<td>Wang Zhiqiang, Liu Jun</td>
<td>2017</td>
<td>CNN</td>
<td>Reduce the loss of feature information</td>
<td>People, cars, buildings, and animals</td>
<td>Excellent outcomes in terms of accuracy and speed</td>
</tr>
<tr>
<td>Elakkiya R, Teja KS, Jegatha Deborah L, Bisogni C, Medaglia C</td>
<td>2021</td>
<td>RCNN</td>
<td>Reduced the complexity and improved the efficiency</td>
<td>Microscope-based object</td>
<td>Highest accuracy of 97.08%</td>
</tr>
<tr>
<td>Ji, Shu-Jun, Qing-Hua Ling, and Fei Han</td>
<td>2023</td>
<td>Yolov4</td>
<td>Multi-scale contextual information and Soft-CIOU</td>
<td>Small object</td>
<td>65.7 at AP50</td>
</tr>
<tr>
<td>Wei Sun, Liang Dai, Xiaorui Zhang, Pengshuai Chang, Xiaozheng He</td>
<td>2021</td>
<td>RSOD</td>
<td>Traffic monitoring</td>
<td>Small object</td>
<td>52.7 at mAP50</td>
</tr>
</tbody>
</table>
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...
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

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

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.

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.

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

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

Table 5: Comparative study of optimization technique

<table>
<thead>
<tr>
<th>Author</th>
<th>Methodology</th>
<th>Advantage</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delia Dumitru, Anca Andreica</td>
<td>Cellular automata rules optimized</td>
<td>Capable of discovering a specific rule for a picture or group of images</td>
<td>—</td>
<td>—</td>
<td>Good</td>
<td>Medium</td>
<td>Medium</td>
<td>Good</td>
</tr>
<tr>
<td>Chen Huang, Tingfa Xu, J. Sasikala, D. S. Jullet</td>
<td>CNN and saliency optimization</td>
<td>Less noise</td>
<td>Good</td>
<td>Good</td>
<td>—</td>
<td>—</td>
<td>Medium</td>
<td>Good</td>
</tr>
<tr>
<td>Delia Dumitru, Anca Andreica</td>
<td>The optimized hybrid adaptive cuckoo search algorithm</td>
<td>—</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Lokesh Jain, Rahul Katarya</td>
<td>Whale optimization algorithm</td>
<td>Better performance</td>
<td>Good</td>
<td>—</td>
<td>Medium</td>
<td>Good</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>D. Rammurthy, P. K. Mahesh</td>
<td>WHOA-DNN</td>
<td>—</td>
<td>Good</td>
<td>Good</td>
<td>—</td>
<td>—</td>
<td>Medium</td>
<td>Good</td>
</tr>
<tr>
<td>Yun Zhang, Yongguo Liu</td>
<td>WOCDA</td>
<td>—</td>
<td>Good</td>
<td>Good</td>
<td>—</td>
<td>—</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Delia Dumitru, Anca Andreica</td>
<td>R-CNN with NAS optimization</td>
<td>—</td>
<td>Good</td>
<td>Good</td>
<td>—</td>
<td>—</td>
<td>Good</td>
<td>Good</td>
</tr>
</tbody>
</table>
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 network 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 F1 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


A review of small object and movement detection based loss function


