e-Prime - Advances in Electrical Engineering, Electronics and Energy 8 (2024) 100535

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

journal homepage: www.elsevier.com/locate/prime

Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms in Surveillance Videos

T. Mohandoss [a,*], J. Rangaraj [b]

[a] Department of ECE, Annamalai University, Chidambaram, Tamilnadu, India
[b] Department of ECE, (Deputed to GCT Coimbatore), Annamalai University, Chidambaram, Tamilnadu, India

ARTICLE INFO

Keywords:
Multi-object detection
LuNet
Deep learning
You Only Look Once version 2(YOLOv2)

ABSTRACT

Multiple object tracking (MOT) in videos benefits multiple applications, including robot navigation, video surveillance, video analytics, and intelligent transportation systems. Although significant progress has been made since early studies, visual tracking of many objects remains challenging because of frequent occlusions in measurements, environmental noise, changeable number of objects, and appearance similarity across objects. The proposed work focused on three significant processes, feature extraction, object detection, and classification, to identify moving objects before sharing information. This work proposes a multi-object video detection method using LuNet and deep reinforcement learning. The enhanced "you only look once" version 2 (YOLOv2) initially detects numerous objects. In this work, a base network of the YOLOv2 changed by lowering the metrics and substituting it with LuNet. In the enhanced model, the LuNet network is used for feature extraction to extract the most expected characteristics from the image. Furthermore, the proposed model is compact because of the underlying network's LuNet architecture. To demonstrate the proposed technique's performance, this method compares it to numerous state-of-the-art algorithms on the MOT20 vehicle benchmark dataset. The proposed method achieves a higher % classification accuracy of 94% for moving object settings. The experiments demonstrate that the proposed method outperforms existing models in terms of performance and accuracy.

1. Introduction

Many people are interested in object detection and tracking nowadays since it incorporates various applications in recent research advances and its equal importance in real-time and academic technologies [1], such as security surveillance and transportation wait—robot vision, as well as surveillance and autonomous driving. Object tracking and detection can be accomplished using different sensing modalities, including light detection and ranging (LIDAR) and computer vision (CV) radar.

MOT is more complex than tracking a single distinct object [2]. MOTs in a single category should use the recognition results to produce new tracked objects and remove objects when they leave the camera's field of view [3].

Furthermore, the pose varies; the occlusion and backdrop clutter problem is more complex than tracking a single object. Deep learning (DL) techniques have been developed to solve these challenges. Traditional features, for example, can be replaced by deep neural network features to correlate detection outcomes, even though these features are learned via recognition or classification tasks [4]. Furthermore, exploring MOT features such as temporal ordering or spatial attention maps improves performance. In addition, specific end-to-end DL structures are designed to extract features from motion data and appearance descriptors. Although DL techniques have the potential to be used for the MOT problem [5], there is still much opportunity to harness the power of DL to increase tracking efficiency due to DL's tremendous developments in the field of image identification and classification [6].

The scanning and searching for specific categories of objects (such as people, cars, and buildings) in image/video frames is called object detection. Image processing methods or DL networks can be utilized to detect objects. Techniques for image processing are frequently unsupervised, requiring no prior knowledge for training. These methods, however, are limited by various reasons, including complicated scenes, lighting effects, occlusion effects, and clutter effects. All of these issues are best addressed by deep learning-based object detection. The supervised nature of DL networks' operation limits them to massive amounts of training data and GPU computational capacity. DL models can detect and track objects in general and specialized domains. A deep

* Corresponding author.
E-mail addresses: tmohandossece@gmail.com (T. Mohandoss), ranga.jsrd@gmail.com (J. Rangaraj).

https://doi.org/10.1016/j.prime.2024.100535
Received 23 February 2024; Received in revised form 22 March 2024; Accepted 1 April 2024
Available online 6 April 2024

convolutional neural network (DCNN) is the detection network's backbone, extracting crucial information from the input video/image frame, and the retrieved characteristics are employed to locate and classify objects inside a similar frame. These recognized objects are followed via object tracking based on the proximity of features from one frame to the next.

MOT seeks to maintain object recognition in motion and appearance across time and create trajectory data for objects in a scene [7]. Furthermore, pedestrians play a significant role in tracking objects during research. MOT algorithms that can reliably evaluate pedestrian motions are helpful in various applications like autonomous driving, smart cities, and video analysis. Moreover, the tracking efficiency still needs to be enhanced because the tracked objects are frequently obscured by obstacles or other people, as well as their comparable visual features.

In recent times, there have been tremendous achievements in research on MOT for autonomous driving. However, challenges such as the different shapes of cars and pedestrians in traffic situations require more work to fully utilize existing MOT technologies for autonomous driving, motion blur, and background interference. Existing multi-object visual tracking technology still has several limitations. Multi-object visual tracking should begin with progressively tough challenges, including an unexpected number of objects, challenging object identification, frequent object occlusion, etc. Object continually entering and leaving the field of view is a common and expected phenomenon.

Because of the variability in the different kinds of objects MOT algorithms face in autonomous driving applications, real-time detection of MOT techniques is required [8]. To separate comparable components, the algorithm must be able to extract features from them. Finally, the difficulties that autonomous vehicles confront when tracking multiple objects can be split into two categories: object-tracking variables and contextual considerations. Object-tracking factors can cause problems such as shape changes, scale changes, motion blur, etc. Background factors, particularly softening of background interference, disappearance and occlusion of objects, comparable backdrop interference, weather changes, and so on, have a significant impact [2].

This work employs the most recent and enhanced version of "You Only Look Once version 2" (YOLOv2), with the LuNet deep learning framework, a well-known sliding window-based deep learning model in computer vision, as the foundation for the implementation in this work. Object detection and parameter tuning for near-real-time, high-precision performance. The enhanced algorithm's goal is to enhance the accuracy of vehicle recognition by altering the network model many times and comparing the proposed algorithm's accuracy to other models.

This work proposes a novel approach for detecting vehicle objects in real-time datasets that combines Enhanced YOLOv2 and LuNet. Combining Enhanced YOLOv2 with LuNet entails incorporating LuNet features into the YOLOv2 framework. This could imply using LuNet as a pre-processing step to improve the input image data and incorporating LuNet's features into the feature extraction layers. Because of its high efficiency and accuracy, the LuNet is the backbone of this feature extraction method. Instead of using fixed feature maps in the original YOLOv2 framework, the determination of feature maps for vehicle detection in this work is based on the coincidence of object scale and receptive field. The new feature map selection significantly improves the performance of the proposed object detection model. LuNet feature maps are reselected based on their receptive fields for enhanced object detection instead of the fixed-chosen strategy in the original YOLOv2 framework.

The significant contribution of this work is as follows:

- Collect and convert various sets of video data into frames in real time.
- To remove high-frequency noise elements from video frames and identify various objects in a single frame, YOLOv2-LuNet uses the Kalman filter.
- This work proposes a novel lightweight multi-object detection deep learning model that can efficiently extract and detect multiple objects in real-time scenarios.
- This work first offers a typical generalized enhanced YOLOv2 architecture with LuNet-based multi-object tracking.
- To propose a LuNet classifier-based feature extraction method to enhance feature representation and capture more discriminative information about vehicles in the MOT20 dataset.
- To combine LuNet features with the YOLOv2 object detection framework to achieve higher accuracy in detecting vehicles in challenging scenarios such as crowded scenes, occlusions, and varying lighting conditions.
- Using LuNet features could enhance the generalization capability of the YOLOv2 model, allowing it to perform well on unseen data beyond the MOT20 dataset.
- The experiments conducted using the MOT20 dataset could serve as a benchmark for evaluating the effectiveness of the proposed approach against existing methods, providing insights into its strengths and limitations.

The remaining part of this work is as follows: Existing video object detection works are discussed in Section 2. Section 3 describes our suggestedtechnique in detail. The simulation data and analyses are presented in Section 4. Section 5 concludes the proposed method.

2. Literature survey

RamachandranAlagasamy et al. (2023) introduced a reptile search optimization method based on DL-based multi-object detection and tracking technology (RSOADL-MODT). The suggested technique is intended to detect and track existent objects using location analysis, tracking, and motion identification. It consists of three approaches: object detecting, classifying, and tracking. Initially, the proposed RSOADL-MODT technology employs an object identification module based on path-augmented RetinaNet (PA-RetinaNet), which enhances feature extraction. Subsequently, a quasi-RNN (QRNN) classifier is used for classification [9].

S. Prabhu et al. (2023) suggested a modified ResNet model (M-Resnet) to improve images influenced by limited light. Experiment outcomes comparing the output of existing approaches and a modified architecture of the ResNet model demonstrate significant gains in object detection in surveillance videos. The proposed method achieves superior outcomes in metrics such as recall, precision, and pixel accuracy, as well as reasonable improvements in object detection [10].

Malik JavedAkhtar et al. (2022) introduced an enhanced YOLOv2 algorithm to detect objects in surveillance videos, i.e., vehicle detection and identification. This article updates the YOLOv2 primary network, reducing the parameters and substituting it with DenseNet. Because of the underlying network's dense construction, the proposed model is more compact. DenseNet-201 is used as the base network in this work because there are direct connections between all levels, which aids in retrieving relevant data from the initial layer and sending it to the last layer. The suggested model was trained using Kaggle and KITTI datasets, and its performance was cross-validated using Pascal VOC and MS COCO datasets [11].

MagedFaihanAlotaibi et al. (2022) presented a unique computational intelligence-based harmony search algorithm (CIHSA-RTODT) for real-time object recognition and tracking technology in a video surveillance model. The proposed technology includes an upgraded object detection module based on RefineDet to recognize multiple objects in video frames. Additionally, the Adagrad optimizer is used to tweak the hyperparameter outcomes of the upgraded RefineDet method. HSA with the twin support vector machine (TWSVM) algorithm was also used for object categorization. Extensive experimental performances are demonstrated using open-available datasets, and the outcomes are scrutinized in various methods [12].

Wang Xiyang et al. (2022) suggested a robust and quick MOT approach using camera-LiDAR fusion. Using the properties of LiDAR cameras and sensors, an efficient depth correlation mechanism is created and implemented into the suggested MOT approach. This correlation technique monitors objects in the 2D domain while they are far away and only identified by the camera. This will update the 2D motion with 3D information collected when the object is in the LiDAR field of view [13].

PalashYuvrajInger et al. (2022) suggested that the detection classifier is a multi-class subclass detection CNN for distinguishing object frames into multiple subclasses like abnormal and normal. The average accuracy of cutting-edge gun and knife detection frameworks in a single camera view is 84.21% or 90.20%. After thorough testing, this method's excellent precision for recognizing various types of knives and guns on the ImageNet and IMFDB data sets was 97.50%, on the Open Images data set was 90.50%, on the Olmos data set was 93%, and on multi-view cameras was 90.7% [14].

Chen Zhang et al. (2021) suggested a ConvLSTM-based video object detection model with event recognition and object relation systems. The suggested event-aware system can identify places where these complex events occur. Compared to classic ConvLSTM, the suggested technique uses temporal context information more efficiently to assist video-based object detectors in demanding circumstances. An object relation module is used to improve the pooled features of the target ROI utilizing support box selection to increase detection performance even further. The technique obtains 81.0% mAP in video object detection, according to simulation outcomes on the ImageNet VID dataset [15].

Wael Mahdi Bridge et al. (2021) suggested that the model is based on a unique form of recurrent neural network (RNN) with appropriate features derived from objects to forecast the direction of moving objects via cameras, improving tracking accuracy and lowering computing costs. Long short-term memory (LSTM), an RNN with a unique structure, is a conventional and promising approach for tackling machine learning (ML) problems in the video data framework. The algorithm's underlying structure uses frame history information and depth classification methods to generate robust predictions, which will improve tracking in various visual challenges, like the difficulty of correlating observation data of the same object in a hybrid camera network [16].

G. Kiruthiga et al. (2021) advocated employing probabilistic neural networks (PNN) to improve object detection in surveillance image analysis. Below, the work began to investigate the fundamental concepts of neural networks and perceptual networks and some fundamental theories that are sometimes overlooked and can help to understand why DL is becoming more popular in various applications. Surveillance image processing is undoubtedly the most impacted area by this exponential improvement, particularly in image recognition and detection. According to simulation results, CNN-PNN delivers enhanced simulation results in optimal object recognition in video surveillance or video streaming systems [17].

Xiao Yuxuan et al. (2020) suggested a low-light object-detecting night vision detector (NVD) with a specially constructed feature pyramid and context fusion model. This method has found specific answers for low-light object detection by conducting extensive experiments on the real low-light open-available dataset ExDARK and the chosen normal-light corresponding COCO. Also, this method has drawn some practical findings for concern and has found some valuable answers for object detection in low illumination. This method enhances detection performance by 0.5% to 2.8% across the conventional COCO evaluation criteria over the baseline model [18].

2.1. Problem statement

Most of the above multi-object tracking approaches denote objects with raw pixels and low-level constructed characteristics like histograms of oriented gradients (HOG), Haar-like features, and local binary patterns (LBP). Despite their computing efficiency, hand-crafted features have significant limits because they cannot capture the more complicated qualities of things. If the image includes various colors, existing algorithms detect them inaccurately. Aside from that, one of the most typical issues is that if the backdrop lighting shifts, it can be misconstrued as a foreground object. A few approaches also need help to identify shadows. Camouflage issues can arise when the foreground and background objects are too close together. Non-static background modeling is another difficulty. In high-traffic areas, many foreground objects frequently hide the background. The persistent foreground and background are difficult to distinguish due to ongoing change [19].

Although there are several traditional methods for recognizing objects in surveillance videos, the fundamental issue is that previous approaches need to be more precise and are also prohibitively expensive. Researchers are currently working on identifying vehicle objects using DL algorithms. In this work, the enhanced YOLOv2 method with LuNet, a type of DL technology, was employed to identify multiple objects in surveillance video.

3. Proposed methodology

The work focuses on developing feature selection and classification methods for addressing current issues in detecting moving objects captured by event cameras. In the enhanced model, the LuNet network is used for feature extraction to extract the most expected characteristics from the image. The enhanced "you only look once" version 2 (YOLOv2) initially detects numerous objects. The suggested method evaluates frame entropy and minimizes energy consumption. Furthermore, the classification of the dataset is implemented utilizing enhanced YOLOv2 and LuNet architecture to minimize the computational duration of the proposed enhanced YOLOv2. The comparison of real-time data sets accomplishes classification. Fig. 1 shows the basic architecture for the proposed algorithm.

A real-time video dataset called MOT20 was gathered and transformed into tiny video frames in this work. Suggested layouts and noise detection for various moving objects will be provided in the video/image frame of the identified object. A Kalman filter removes and smooths the noise. The suggested filter evaluates the model's parameters and forecasts future observations using noise measurements collected over time.

The filter can make forecasts, collect measurements, and then update depending on the predictions and comparisons at each level. The status of many linear processes can be predicted and updated using mathematical estimators. The YOLOv2 network uses binary cross-entropy loss instead of multiple labels to categorize and predict bounding box categories to increase performance. This work changed the YOLOv2 base network by lowering the number of metrics and substituting it with LuNet. In the enhanced model, this work applies LuNet technology for feature extraction to extract the most expected characteristics from the image. Furthermore, the suggested approach is compact because of the underlying network's LuNet architecture. Because of the direct connections between all levels, this work uses LuNet as the base network, which allows us to harvest vital details from the initial layer and transfer them to the last layer.

3.1. Dataset

This work uses the MOT20 benchmark, which consists of 8 new sequences illustrating highly crowded and challenging scenes. The dataset was initially proposed at the 4th BMTT MOT Challenge Workshop at the Computer Vision and Pattern Recognition Conference (CVPR) 2019, and it allows users to assess the cutting-edge techniques for MOT in extremely crowded scenarios [20].

The new benchmark's dataset was carefully chosen to test trackers and detectors in extremely busy scenes. Compared to previous challenges, some new sequences have pedestrian densities 246 per frame. For this work, eight sequences were created, half used for training and

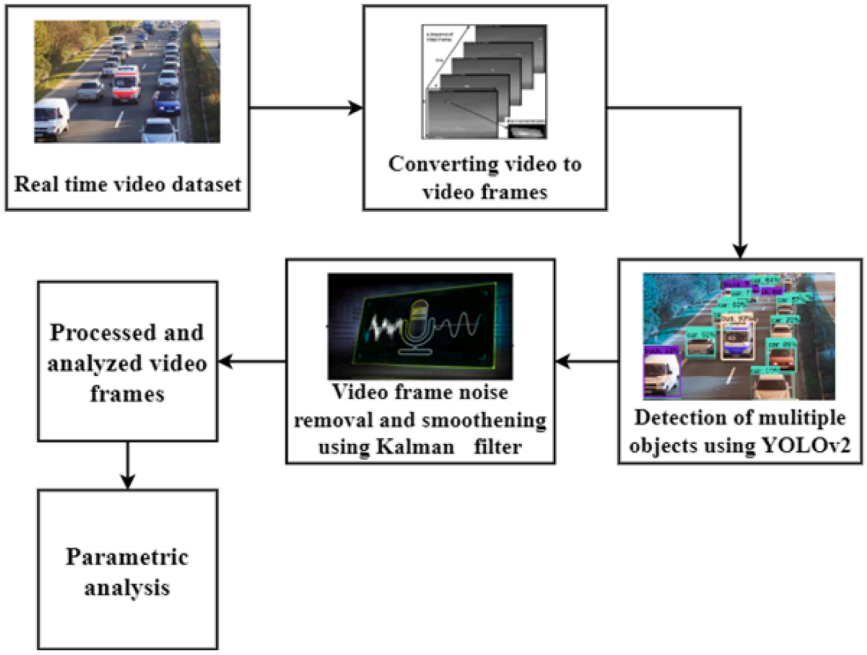

Fig. 1. Basic Block Diagram of Object Detection.

the other half for testing. Annotations of test sequences are not published to prevent methods from becoming overly tuned to specific sequences. The sequences were filmed on three separate sets. Several sequences were shot for each scene and spread across the train and test sets. However, one of the scenarios was set aside for testing to challenge the method's generalization ability. Compared to MOT17, the new data has roughly three times as many bounding boxes for training and testing. All sequences were shot in high resolution from an elevated vantage point, with an average pedestrian density of 246 per frame, ten times higher than the initial baseline density.

3.2. *Feature extraction using LuNet*

The purpose of a feature extractor in vehicle detection is to convert the original input image into a set of representative features to capture vehicle detection information. These features are fed into later stages of the detection process, like classification or bounding box regression. In general, feature extraction involves reducing the dimensionality of input data while retaining relevant information. This reduction makes subsequent calculations more efficient and lowers the risk of overfitting, which occurs when the model learns noisy or irrelevant patterns from the data. Both bounding box and class prediction are based on features extracted from images. In this work, the LuNet network is the backbone of the YOLOv2 model for feature extraction in-vehicle object detection.

The modified version of HAST-IDS is called LuNet. HAST-IDS is a multilayer framework that uses a Convolutional Neural Network (CNN) [21] to extract spatial data and a Recurrent Neural Network (RNN) [22] to capture network data's temporal characteristics. HAST-IDS works by stacking all RNN layers after stacking the CNN layer stack. In contrast, LuNet stresses the hierarchical structure of CNN and RNN layers.

CNN hierarchy takes precedence over the RNN hierarchy in HAST-IDS, which may result in the loss of temporal data inherent in the original input data, resulting in inefficient RNNs. Moreover, LuNet synchronizes RNN and CNN DL in many phases to efficiently capture network traffic's spatial and temporal data. Each step is carried out with the help of a LuNet block, which combines CNN and RNN blocks. The total number of filters utilized in the RNN/CNN framework calculates the model's learning granularity. CNN produces a feature map, which is then processed by the ReLu (activation function) [23–24], followed by pooling and resampling to remove irrelevant input. Batch normalization alleviates the covariance shift problem, which can occur owing to dynamic changes in the range of input values from one layer to another to improve learning. Moreover, to achieve superior learning results, use trainable parameters to tune and update network weights during the learning process. As the granularity of one LuNet block goes from coarse-grained to fine-grained, new layers must be added to modify the final size of one level that will likely be used as input to the following level.

In overfitting, the network has learned enough from the training data to limit its capacity to detect biases in new samples. After the RNN+CNN framework, LuNet employs a dropout layer with a default value of 0.5. Finally, CNN and global average pooling layers retrieve spatial and temporal characteristics learned from the LuNet frameworks.This work employs an improved ResNet-v2 network named LuNet [25] to extract object appearance features. LuNet's input is a $128 \times 64 \times 128 \times 64$ image patch. The network employs LeakyReLU as the activation function for robust optimization, multiple $3 \times 33 \times 3$ max pooling, and two-stride instead of stride convolution—Fig. 2 depicts the feature map in the last re-block of the average pooling layer. This model retrieves the object's 128-dimensional embedding features from the final multilayer perceptron (MLP) layer. Compared to previous feature extraction networks, this network is lightweight (5M parameters).

3.3. *Object detection using YOLOv2*

YOLOv2 [15] is an evolution from YOLO. YOLOv2 takes decisions from prior training challenges and implements new principles to improve YOLO's speed and detection accuracy. Six stages are incorporated in YOLOv2 and are discussed as follows:

i. Batch normalization (BN):

The mean and variance for each mini-batch are determined and utilized for activation. The activations are then normalized for all mini-batch by employing a zero mean and a standard deviation of one. Finally, each mini-batch's elements are sampled by using the same distribution. This procedure is known as batch normalization [23]. It generates the same activation distribution.

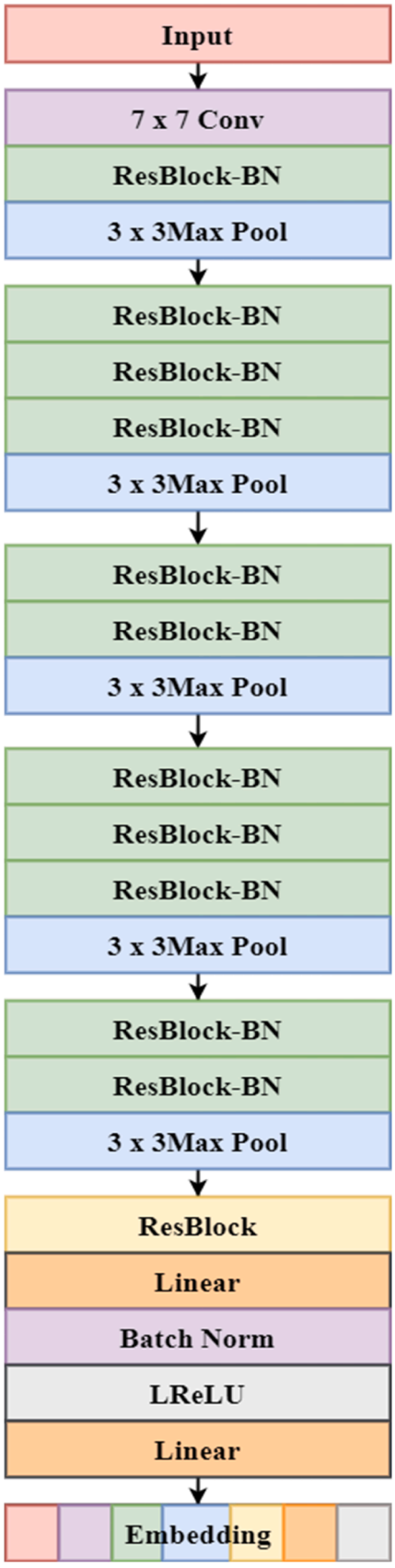

Fig. 2. Architecture of feature extractor.

ii. High-resolution classifier:

The YOLO backbone employs the (224 × 224) input resolution. The input resolution in YOLOv2 has been enhanced to (448 × 448). As a result, the network must be modified to accommodate the new resolution input of the object detection task. As a result, specific changes were made to the classification network in YOLOv2 with a single-resolution image (448 × 448) and ten epochs, raising 1% of the average precision (mAP).

iii. Anchor box convolution:

As previously stated, Faster RCNN generates region suggestions using an anchor box as a reference, which are then parameterized relative to this proposed anchor box to forecast bounding boxes. YOLOv2 employs this estimation method. The class and object scores are then projected for each predicted bounding box. Withdrawals climbed by 7%, while mAP declined by 0.3%.

iv. Anchor box size and aspect ratio prediction:

The proposed YOLOv2 employs the LuNet approach, which trains the bounding boxes to acquire increased priors. This background is then utilized to define the anchor box's center location. Predict the size and aspect ratio of the anchor box utilizing the clustering information. The proposed technique increases detection precision.

v. Fine-grained features:

As previously stated, YOLO trains with images (224 × 224). The YOLO design has been tweaked to form the Yolov2 architecture. YOLOv2 is retrained using higher resolution images (448 × 448) to pinpoint tiny objects. YOLOv2 uses higher and lower resolution features throughout this retraining process by stacking nearby data in various channels and raises 1% of the detection MAP.

vi. Multi-scale training:

To allow the system to run reliably on images of varying sizes, a new image of size {320, 352,..., 608} is selected every ten (randomly selected) batches. That is, the same network can be detected at multiple resolution levels. For instance, the proposed YOLOv2 reaches 40 fps at higher resolutions and 78.4% mAP, whereas YOLO obtains 63.4% mAP and 45 fps on VOC 07. YOLOv2 attains great detection accuracy while operating swiftly; however, the process is confined to high-resolution and multi-class object recognition.

3.4. Object detection using enhanced YOLOv2-LuNet

This research develops an enhanced multi-object detection technology based on the enhanced YOLOv2-LuNet model and a target tracking system based on the complex moving window Kalman filter. This method allows for the efficient monitoring of several moving objects in perplexing settings. MOT is a dataset of real-time video frames captured with this model. The proposed filter helps to eliminate and smoothen the noise. The video frames are processed and evaluated once the noise has been removed. The detected object frame will contain an enhanced YOLOv2 model that identifies numerous moving objects. YOLOv2 is a real-time object detection system that takes in an image and directly provides the object position and confidence score.

In YOLOv2, sliding windows are not used for feature extraction, and the classifier is removed. Thus, this work proposes using LuNet as the primary network for object detection in the upgraded YOLOv2 version of this study because of its superior performance. The input image is divided into many areas by this approach. When the center of a labeled object lies in a specific zone, the region will be used to predict the object.

YOLO v2 object detection method diagram is shown in Fig. 3.

The anchor box anticipates the bounding box (b box) in the YOLOv2 method, and the last fully linked layer is eliminated. Convolutional layers and pooling layers make up the network structure. The image size has been reduced from 448 × 448 to 416 × 416. When the image is reduced 32 times, the final feature size is 13 × 13. A center grid anticipates objects that will fall in the center of the image. To obtain a network with superior detection and recognition skills for many objects, the work changed the network parameters and ran multiple trials using the network framework of YOLOv2-voc.

Initially, this work does high-resolution pre-training on the MOT20 dataset. Fine-tuning technology is utilized to train the vehicle dataset defined by the pre-trained LuNetnetwork based on categorization. During the training phase, the proposed method enriches the data with random scaling, saturation, and exposure to test the usefulness and resilience of the proposed technique. These techniques were employed to broaden the scope of our sample. The entire image is then sent into the neural network only once. The neural network separates the image into regions, predicts the borders and likelihood of each zone, and assigns weights to all frames depending on probability. Finally, the proposed work only acquires test results with confidence scores more significant than a specific threshold. In our experiments, the threshold is set to 0.25. Fig. 4 depicts the entire training procedure.

4. Results and discussion

The proposed technique is evaluated using the simulation tool called MATLAB. In this work, a video dataset called MOT20 is used.The suggested YOLOv2-LuNet processing is depicted in Fig. 5. The metrics examined are accuracy, precision, mean absolute position (MAP), ground truth (GT), detection rate (DET), true positive rate (TP), and false positive rate (FP).

4.1. Accuracy

Accuracy measurement aids in determining how well the proposed LuNet classifier detects objects in video frames. It sheds light on the model's ability to correctly identify and localize objects of interest.It refers to the degree of understanding between a noise and actual value evaluation. Table 1 illustrates an accuracy analysis of the suggested approach.The suggested method's accuracy values are evaluated against current methods by employing feature masking in video sequence frames. The existing technologies' accuracy ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 76%, 82%, 85%, 86%, and 87%, respectively. An accuracy of 94% was observed in the case of the proposed model. Fig. 6 depicts that the suggested method has a 94% maximum accuracy value.Comparisons show that it outperforms traditional methods because it uses architectural innovations and layer combinations to better capture temporal dependencies and spatial features in video data than other models.

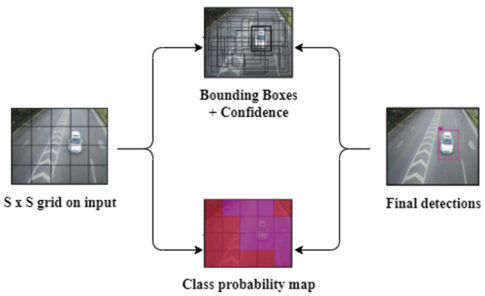

Fig. 3. YOLO v2 objectdetection method diagram.

4.2. Precision

It is the degree to which repeated noise measurements yield identical outcomes under comparable conditions. Precision is the ratio of true positives (objects correctly detected and localized) to the sum of true positives and false positives (objects incorrectly detected).The precision evaluation of the proposed method is depicted in Table 2.The assessment of precision outcomes shows that the proposed approach achieves greater accuracy than the cutting-edge techniques. The existing technologies' precision ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 70%, 75%, 84%, 86%, and 88%, respectively. Fig. 7 depicts that the suggested method has a 95% maximum precision value.Comparisons show that it outperforms existing methods by effectively reducing the number of false positive detections, resulting in a higher precision score.

4.3. Recall

Recall is the proportion of true positive objects the model successfully detects out of all the objects in the video frames. It ensures the model detects as many relevant objects as possible, which is critical for thorough video analysis. Recall is calculated by dividing true positives (correctly detected objects) by the sum of true positives and false negatives.

The ratio of appropriate images acquired overall is referred to as recall. The recall analysis for the suggested method is depicted in Table 3. Fig. 8 depicts that the suggested method has a 92% maximum recall value. Compared to existing techniques, the proposed approach outperforms existing procedures and methods for object tracking and classification. The existing technologies' recall ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 59%, 66%, 72%, 74%, and 84%, respectively. A recall of 92% was observed in the case of the proposed model. When the number of images increased so does recall value. Comparisons indicate that it outperforms traditional methods because

4.4. True Positive(TP)

True positives are cases in which the model correctly identifies and localizes objects of interest in video frames. The measurement of true positives directly assesses the object detection system's detection accuracy. The number of correctly matched detections determines the true positive count. Each correctly matched detection increases the number of true positives.

True positive analyses are the criterion necessary to evaluate tracker performance. The first stage is to see if each suggested outcome is a TP that matches the underlying goal. Table 4 assesses the proposed method's TP. Fig. 9 depicts that the suggested method has a 90% maximum TP value.The proposed approach yielded the highest true positive value (90%).The existing technologies' TP ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 56%, 58%, 63%, 65%, and 67%, respectively. A TP of 90% was observed in the case of the proposed model. Comparisons indicate that it outperforms existing methods.

4.5. False Positive (FP)

The initial stage establishes whether each predicted result is an FP. False positives happen when the classifier incorrectly identifies background or unrelated objects as the objects of interest.Table 5 depicts the results of the FPs. In image classification and object detection applications, false positives denote the number of objects classified or detected by the proposed method per second. It can be used to calculate a model's average processing speed. Fig. 10 depicts that the suggested method has 89% FP value.

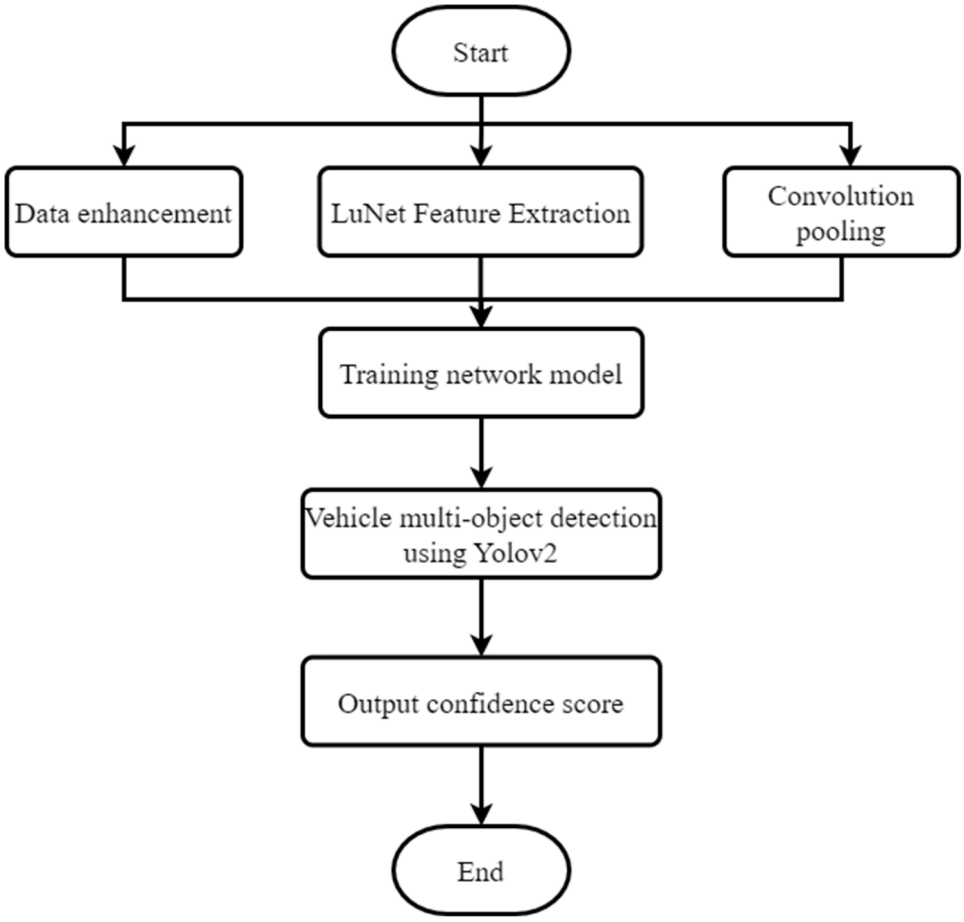

Fig. 4. Vehicle detection flowchart.

Fig. 5. Examples of proposed vehicle detection with our method.

Table 1
Accuracy evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	66	72	79	80	82	90
200	70	74	81	82	84	92
300	72	76	82	84	85	92.5
400	74	80	84	85	88	93.5
500	76	82	85	86	87	94

4.6. Ground truth (GT)

GT annotations provide a standardized reference for comparing object detection models or algorithms. During model evaluation, ground truth annotations are compared to model predictions to calculate performance metrics like accuracy, precision, and recall.GT is the knowledge collected in the field. Image information can be linked to real-time characteristics, and real-world materials can be used in the field. Table 6 depicts the outcomes of a GT on the applicability of the suggested technique.

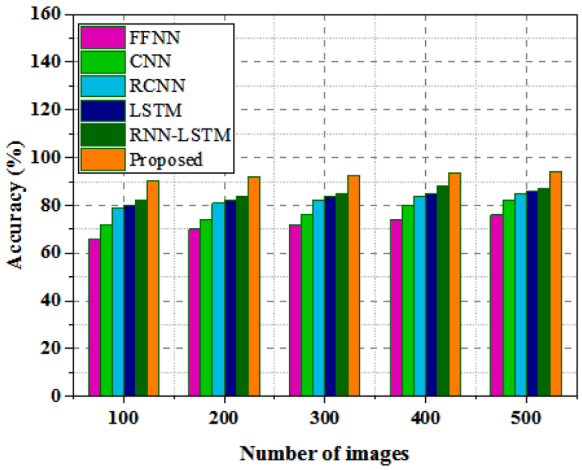

Fig. 6. Accuracy analysis.

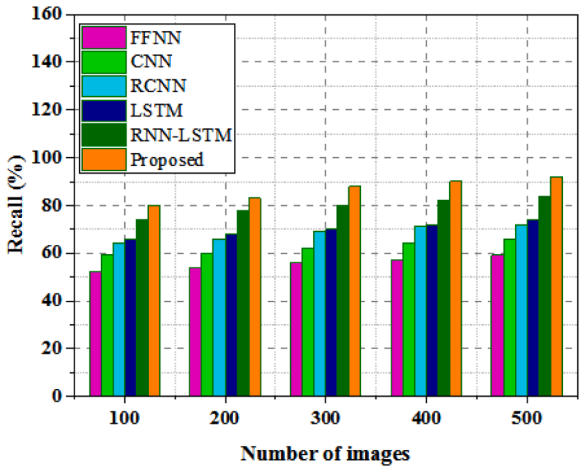

Fig. 8. Precision Analysis.

Table 2
Precision evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	62	66	71	74	77	89
200	64	67	75	76	78	90
300	65	69	78	80	82	92
400	68	70	80	83	85	94
500	70	75	84	86	88	95

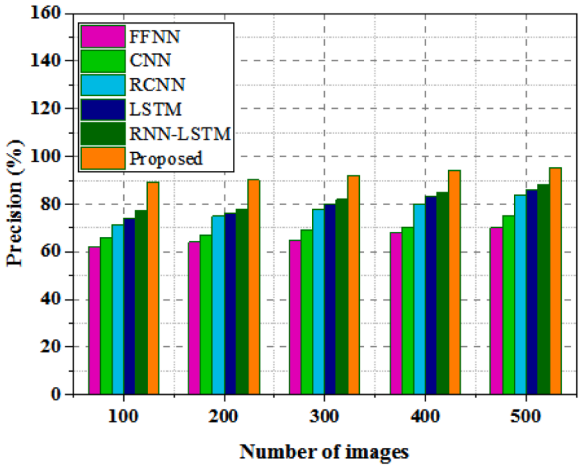

Fig. 7. Precision analysis.

Table 3
Recall evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	52	59	64	66	74	80
200	54	60	66	68	78	83
300	56	62	69	70	80	88
400	57	64	71	72	82	90
500	59	66	72	74	84	92

Fig. 11 depicts that the suggested method has a 91% maximum GT value.The existing technologies' GT ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 70%, 80%, 83%, 85%, and 87%, respectively. A GT of 91% was observed in the case of the proposed model.

Table 4
True Positive evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	48	50	54	56	59	73
200	50	52	56	58	61	79
300	52	56	60	60	63	83
400	53	57	62	62	65	85
500	56	58	63	65	67	90

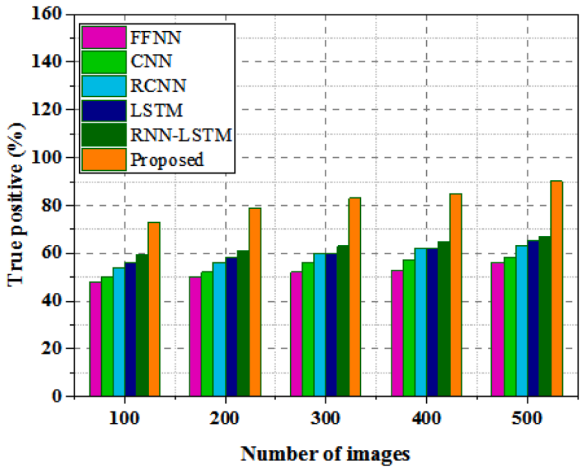

Fig. 9. TP analysis.

Table 5
False Positive evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	48	52	57	59	61	69
200	50	55	60	62	65	75
300	52	58	62	64	67	80
400	55	60	66	68	70	88
500	57	62	68	70	74	89

Comparisons indicate that it outperforms existing methods.

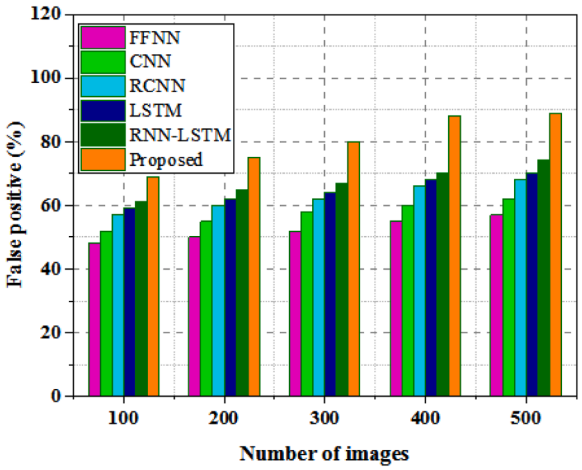

Fig. 10. FPanalysis.

Table 6
Ground truth evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	62	68	72	75	77	81
200	64	72	74	77	79	83
300	66	76	79	81	83	85
400	68	78	81	82	84	89
500	70	80	83	85	87	91

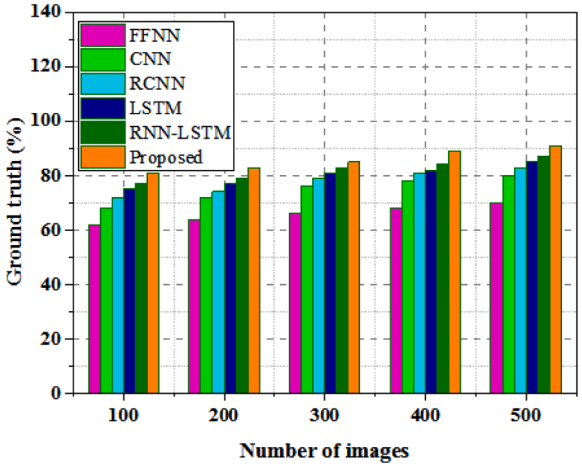

Fig. 11. Ground truth analysis.

4.7. Detection rate

The detection rate is the ratio of true positives to total ground truth objects. It denotes the proportion of objects correctly detected by the system among all the objects in the video frames.The simulation outcomes of the detection rate are shown in Table 7. In terms of detection values, the proposed technique outperforms all previous strategies. Fig. 12 depicts that the proposed approach yielded the highest detection rate (88%). The existing technologies' detection rates for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 52%, 60%, 61%, 70%, and 75%, respectively. Comparisons indicate that it outperforms existing methods.

Table 7
Detection rate evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	42	49	52	58	64	67
200	46	52	54	60	67	77
300	48	56	56	63	70	80
400	50	58	58	67	72	83
500	52	60	61	70	75	88

4.8. MAP

The mean absolute position (MAP) refers to location data, including coordinate information associated with a single image. The MAP is calculated as the average of the average precision (AP) values for all object classes. It converts a scalar value that represents the model's overall detection performance.Table 8 depicts the examination of the MAP in The suggested method outperforms the existing approach when the proposed MAP score is compared to the cutting-edge MAP score. Fig. 13 depicts that the proposed approach yielded the highest MAP (90%). The existing technologies' MAP for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 66%, 69%, 76%, 80%, and 82%, respectively. Comparisons show that it outperforms existing methods because the training strategy used for LuNet may prioritize optimizing MAP, resulting in a model that is inherently better at producing high-quality detections across multiple object categories.

4.9. Quantitative Results

Based on quantitative studies, integrating static and dynamic models can improve detection performance. Traditional methods fail to identify the relevance of video objects by merging a static foreground network with a dynamic highlight network. Because this modeling method is trained on static foreground information, it produces more accurate forecast values than other methods. Based on earlier research, this approach can imply that reducing training data reduces performance and that the proposed method is data-driven. The values computational time of the suggested and traditional approaches are analyzed in Table 9.

The proposed procedure is faster than the others. This procedure was discovered to save computing time while also eliminating substantial bottlenecks in execution efficiency. In most cases, video prominence is prevented by motion or edge data computation.

Figs. 14 and 15 shows the outcomes and include the suggested model's static and dynamic impacts with other traditional approaches. When compared to static and dynamic processes and other models, the proposed approach using static and dynamic processes lowers mean absolute position and computational costs. Since calculation duration is

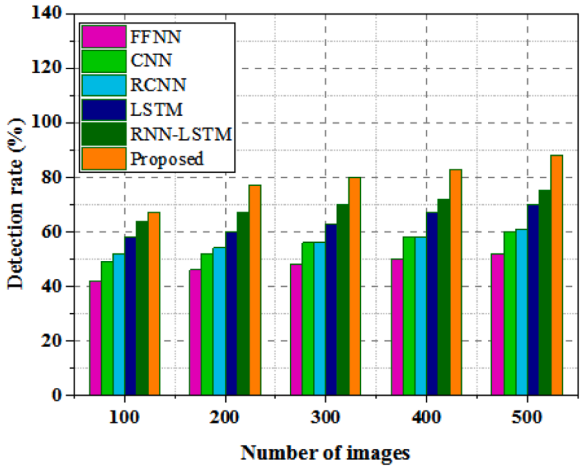

Fig. 12. Detection rate analysis.

Table 8
MAP Evaluation.

Number of images	FFNN	CNN	RCNN	LSTM	RNN-LSTM	Proposed
100	52	58	62	66	69	70
200	56	60	68	70	72	74
300	59	62	70	72	74	78
400	62	65	72	76	78	82
500	66	69	76	80	82	90

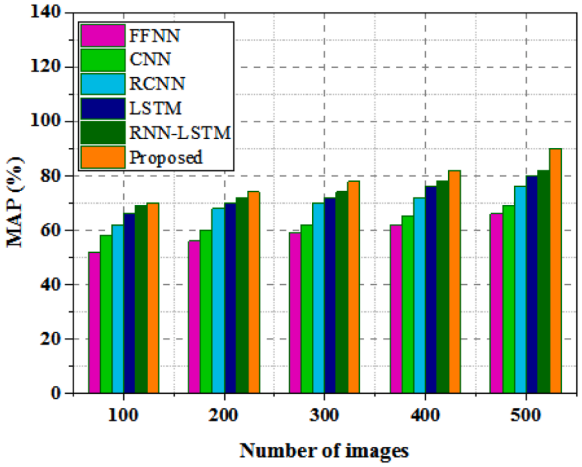

Fig. 13. Mean absolute position analysis.

Table 9
Mean absolute position and Computational time evaluation.

Algorithms	MAE			Computation time(s)		
	Static	Dynamic	Static and dynamic	Static	Dynamic	Static and dynamic
FFNN	0.42	0.44	0.40	83	80	75
CNN	0.30	0.32	0.26	77	75	60
RCNN	0.23	0.25	0.21	55	52	45
LSTM	0.13	0.14	0.09	50	40	33
RCNN-LSTM	0.12	0.13	0.08	38	33	27
Proposed	0.10	0.11	0.07	33	28	23

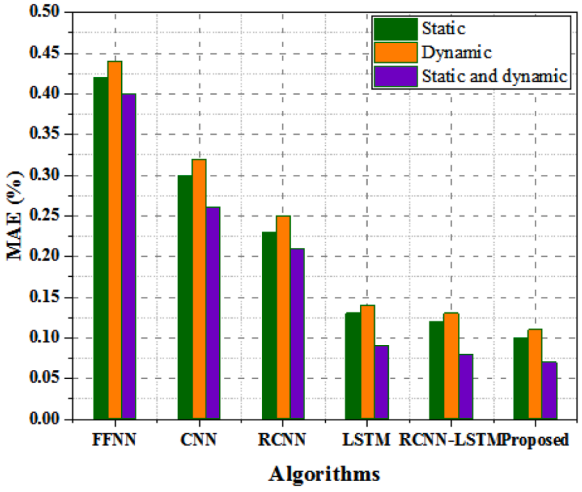

Fig. 14. MAE suggested models over existing DL models.

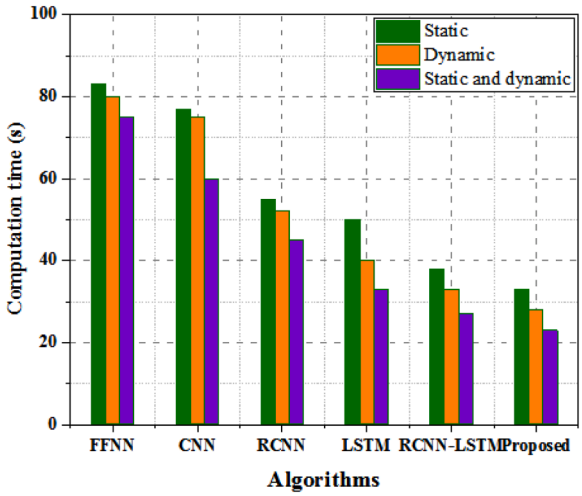

Fig. 15. Suggested models vs. Existing DL models' Computational load values.

reduced, less network is available to process incoming information.

4.10. FPGA Performance

The 100 MHz Xilinx Zynq-7020 FPGA's hardware configuration and capabilities are compared to the algorithm results. This work utilizes logic simulation software incorporated into ISIM for testing purposes. We perform a preliminary examination of timing results and latency requirements after synthesis and implementation to ensure the behavior is fulfilled.

4.11. Power Consumption

Table 10 analyzes the proposed method's power usage to the existing algorithms. The event camera in the proposed method utilizes minimum watts of 0.33 W. Algorithmic performance accounts for only 0.33 W of the device's dynamic power consumption.

The hybrid computational abilities of the Xilinx Zynq module were used in this investigation. It is, however, constrained by the significant delay of frame-based models. The Zynq tool is a diverseand robust development framework. Nevertheless, it may use sleep modes and non-volatile memory, is far more efficient than helpful, and consumes significantly less total power than a Smart Fusion field programmable gate array. Moreover, with proper hardware selection and assembly work, this structure has numerous possibilities for low power consumption (< 1W).

4.12. Comparison with Existing Models

Different tests were carried out to assess the evaluation of the suggested model. This dataset has just three example classes: buses, autos, and trucks. The proposed method outperforms other current methods. The statistics are summarized in Table 11. When the suggested LuNet architecture is employed as the base network in enhanced YOLOv2, the

Table 10
Existing object detection algorithms' power consumption and latency against the suggested technique.

Algorithm	Power(Watts)	Latency(ns)
Proposed	0.42	435
RNN-LSTM	0.46	578
LSTM	0.77	690
RCNN	0.91	745
CNN	1.07	770
FFNN	1.2	825

Table 11
Comparison of suggested models with conventional models.

Algorithm	Accuracy	Sensitivity	Specificity	Precision	G-mean	F-Measure
YOLOv2-Darknet19	93.21	93.6	93.2	93.14	93.4	92.7
YOLOv3-Darknet53	93.32	94.3	94.8	93.4	93.5	93.8
Improved YOLOv3-Net-DenseNet-121	94.5	94.7	95.0	94.3	94.6	94.2
Improved YOLOv2-Net-201-DenseNet-201	94.6	95.5	95.2	94.8	95.2	94.9
Improved YOLOv2 (proposed)- LuNet	95.2	95.8	95.9	95.9	95.6	95.0

best accuracy of 95.2% is obtained. Every layer gathers data from the prior layer and forwards it to the next layer. The classification layer, in particular, connects to the preceding layers and extracts the most significant data for vehicle detection.

The proposed method's accuracy, sensitivity, specificity, precision, g-mean, and f-measure are 95.2%, 95.8%, 95.9%, 95.9%, 95.6%, and 95.0%, respectively. The comparison demonstrates that the suggestedtechnique outperforms the existing methods such as YOLOv2-Darknet19, YOLOv3-Darknet53, Improved YOLOv3-Net-DenseNet-121, and Improved YOLOv2-Net-201-DenseNet-201 in terms of accuracy, f-measure,sensitivity, specificity, precision, and g-mean.

5. Conclusion

This work proposes a multi-object visual tracking model using hybrid LuNet and enhanced YOLOv2 that addresses limitations of traditional algorithms, such as the fact that features created manually cannot capture more complicated object features, tracking fails if the number of objects changes, and so on. This work proposes an enhanced YOLOv2 object detector to detect multiple objects. The proposed technique's feature extraction network is LuNet, which replaces darknet18 in the original YOLOv2. Furthermore, the suggested method is more compact and employs more representative features because of the dense connections between layers. Specifically, every subsequent layer in the proposed base network is connected directly to all preceding layers until the classification layer. We run extensive experiments to assess the performance of the suggestedtechnique, and our model outperforms the state-of-the-art in average accuracy. The experimental outcomes show that the proposed multi-object tracking method improves the algorithm's robustness and accuracy. To achieve higher accuracy, we intend to update and tune the proposed method in the future and maps for vehicle detection and classification. In addition, we plan to test our proposed architecture for other object detection applications, like anomalous activity identification.

CRediT authorship contribution statement

T. Mohandoss: Writing – original draft, Conceptualization. **J. Rangaraj:** Writing – review & editing, Validation, Methodology.

Declaration of competing interest

The authors declare that they have no conflict of interests.

Data availability

Data will be made available on request.

References

[1] X. Zhang, Y. Ling, Y. Yang, C. Chu, Z. Zhou, Center-point-pair detection and context-aware re-identification for end-to-end multi-object tracking, Neurocomputing 524 (2023) 17–30.

[2] S. Guo, S. Wang, Z. Yang, L. Wang, H. Zhang, P. Guo, Y. Gao, J. Guo, A Review of Deep Learning-Based Visual Multi-Object Tracking Algorithms for Autonomous Driving, Appl. Sci. 12 (2022) 10741.

[3] A. Pearce, J.A. Zhang, R. Xu, K. Wu, Multi-Object tracking with mmWave Radar: A Review, Electronics 12 (2023) 308.

[4] Cao, J.; Weng, X.; Khirodkar, R.; Pang, J.; Kitani, K. Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv 2022, arXiv:2203.14360.

[5] S.K. Pal, A. Pramanik, J. Maiti, P. Mitra, Deep learning in multi-object detection and tracking: State of the art, Appl. Intell. 51 (2021) 6400–6429.

[6] H. Suljagic, E. Bayraktar, N. Celebi, Similarity based person re-identification for multi-object tracking using deep Siamese network, Neural Comput. Appl. 34 (2022) 18171–18182.

[7] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, Francisco Herrera, Deep Learning in Video Multi-Object Tracking: A Survey, Neurocomputing (2019), https://doi.org/10.1016/j.neucom.2019.11.023.

[8] D. Balamurugan, S.S. Aravinth, P.C.S. Reddy, A. Rupani, A. Manikandan, Multiview Objects Recognition Using Deep Learning-Based Wrap-CNN with Voting Scheme, Neural Processing Letters 54 (2022) 1–27, https://doi.org/10.1007/s11063-021-10679-4.

[9] R. Alagarsamy, D. Muneeswaran, Multi-Object Detection and Tracking Using Reptile Search Optimization Algorithm with Deep Learning, Symmetry 15 (2023) 1194, https://doi.org/10.3390/sym15061194.

[10] S. Prabu, J.M. Gnanasekar, Realtime object detection through m-resnet in video surveillance system, Intelligent Automation & Soft Computing 35 (2) (2023) 2257–2271.

[11] M.J. Akhtar, R. Mahum, F.S. Butt, R. Amin, A.M. El-Sherbeeny, S.M. Lee, S. Shaikh, A Robust Framework for Object Detection in a Traffic Surveillance System, Electronics 11 (2022) 3425, https://doi.org/10.3390/electronics11213425.

[12] M.F. Alotaibi, M. Omri, S. Abdel-Khalek, E. Khalil, R.F. Mansour, Computational Intelligence-Based Harmony Search Algorithm for Real-Time Object Detection and Tracking in Video Surveillance Systems, Mathematics 10 (2022) 733, https://doi.org/10.3390/math10050733.

[13] X. Wang, C. Fu, Z. Li, Y. Lai, J. He, DeepFusionMOT: A 3D Multi-Object Tracking Framework Based on Camera-LiDAR Fusion With Deep Association, IEEE Robotics and Automation Letters 7 (3) (2022) 8260–8267, https://doi.org/10.1109/LRA.2022.3187264.

[14] M. Annamalai, M.P. Bala, Intracardiac Mass Detection and Classification Using Double Convolutional Neural Network Classifier, J. Eng. Res. 11 (2A) (2023) 272–280, https://doi.org/10.3390/s22103862.

[15] C. Zhang, Z. Xia, J. Kim, Video Object Detection Using Event-Aware Convolutional Lstm and Object Relation Networks, Electronics 10 (2021) 1918, https://doi.org/10.3390/electronics10161918.

[16] Manikandan Annamalai, Ponni Muthiah, An Early Prediction of Tumor in Heart by Cardiac Masses Classification in Echocardiogram Images Using Robust Back Propagation Neural Network Classifier, Brazilian Archives of Biology and Technology 65 (2022), https://doi.org/10.1590/1678-4324-2022210316.

[17] G. Kiruthiga, N. Yuvaraj, Improved Object Detection in Video Surveillance Using Deep Convolutional Neural Network Learning, International Journal for Modern Trends in Science and Technology 7 (2021) 104–108.

[18] R. Ali, A. Manikandan, J. Xu, A Novel framework of Adaptive fuzzy-GLCM Segmentation and Fuzzy with Capsules Network (F-CapsNet) Classification, Neural Comput & Applic (2023), https://doi.org/10.1007/s00521-023-08666-y.

[19] Venmathi, A.R, S. David, E. Govinda, K. Ganapriya, R. Dhanapal and A. Manikandan, "An Automatic Brain Tumors Detection and Classification Using Deep Convolutional Neural Network with VGG-19," *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, Coimbatore, India, 2023, pp. 1-5, doi:10.1109/ICAECA56562.2023.10200949.

[20] Dendorfer, Patrick & Rezatofighi, Hamid & Milan, Anton & Shi, Javen & Cremers, Daniel & Reid, Ian & Roth, Stefan & Leal-Taixé, Laura. (2020). MOT20: A benchmark for multi object tracking in crowded scenes.

[21] Mathiyalagan Palaniappan, Manikandan Annamalai, Advances in Signal and Image Processing in Biomedical Applications, 2019, https://doi.org/10.5772/intechopen.88759.

[22] Alex Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", arXiv:1808.03314v8 [cs.LG] 21 Dec 2020.

[23] Kolli, Srinivas & V., Praveen & John, Ashok & Manikandan, A. (2023). Internet of Things for Pervasive and Personalized Healthcare: Architecture, Technologies, Components, Applications, and Prototype Development. doi:10.4018/978-1-6684-8913-0.ch008.

[24] K. Sheikdavood, P. Surendar, A. Manikandan, Certain investigation on latent fingerprint improvement through multi-scale patch based sparse representation, Indian Journal of Engineering 13 (31) (2016) 59–64.

[25] Y.-S. Yoo, S.-H. Lee, S.-H. Bae, Effective Multi-Object Tracking via Global Object Models and Object Constraint Learning, Sensors 22 (2022) 7943, https://doi.org/10.3390/s22207943.

T. Mohandoss received his Bachelors Degree in Electronics and Communication Engineering with First Class from Anna University, India in 2008 and M.E. Degree in Process Control and Instrumentation Engineering with Distinction from Annamalai University, India in 2012. Presently he is working as an Assistant Professor in the Department of Electronics and Communication, Annamalai University (Deputed to Govt. College of Engg., Bargur). He is currently pursuing the Ph.D. degree at Annamalai University. His research interests include Statistical modelling methods, Image processing, Machine Learning etc.

J. Rangaraj received his Bachelors degree in Electronics and Communication with Distinction from Annamalai University, India in 2008 and Master degree in Process control and Instrumentation from Annamalai University, India in 2012. He received his Doctoral degree from Annamalai University, Tamil Nadu, India in 2019. Presently he is working as an Assistant Professor in the Department of Electronics and Communication, Annamalai University (Deputed to Govt. College of Technology, Coimbatore). His research interest includes Wireless Communication and Mobile Ad-Hoc Networks. He has around 40 papers in her credit in National and International level.