Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Smart Vehicales Surveillance in Foggy Conditions Using Enhanced Deep Learning Algorithms

[1]Shubham Kumar Sain, [2]Khush Khandelwal, [3]Madan Mohan Agarwal
[1,2,3]Birla Institute of Technology, Mesra, Ranchi

Abstract: - Foggy conditions pose significant challenges for vehicles due to reduced visibility, making it difficult for drivers to see other vehicles, pedestrians, and road signs, which increases the risk of accidents. Fog also distorts depth perception, complicating the ability to accurately judge distances and speeds of other vehicles, leading to potential collision. For this we use deep learning methods, while deep convolutional neural networks excel at eliminating fog, they also need to be able to handle photos taken in actual meteorological situations with patches of cloud cover or fog. Blur is harder to categorize in the actual world, and decreasing map or picture quality will result in output results with inconsistent colours or less content. Additionally, the model's complexity will rise with additional convolutional block stacking. Deep learning methods for fog image processing can be plagued by over fitting in addition to the challenge of gathering enough training data. This can restrict the capabilities of the model and make it difficult to use it practically in real-world situations.

This proposed a combined method for removing fog from surveillance images using WaveletFormerNet, a Transformer-based wavelet network designed for real-world non-homogeneous dense fog scenarios, this transformer method use the wavelet transform method. It also uses Multi-Object Detection with Enhanced YOLOv2 and LuNet Algorithms to detect objects. When these methods are combined, they can better handle the intricacies of hazy surroundings, improving both visibility and object detection precision. The effectiveness of the proposed technique is proven through rigorous testing, highlighting its potential to enhance the operation of monitoring systems in difficult weather situations.

Keywords: *Convolutional neural networks (CNN), Transformer-based wavelet network, YOLOv2, LuNet Algorithms, Real-world meteorological conditions*

1. Introduction

Haze is a frequent atmospheric phenomena that ruins and distorts pictures. Many computer vision applications, including remote sensing processing [1, 2, 7] and video analysis and identification [3, 4], need the use of image de-hazing algorithms. Picture de-hazing is a crucial low-level picture recovery work and a pre-processing step for high-level vision tasks, and it has been a hot research area in computer vision and image processing in recent years. The traditional atmospheric scattering model (ASM) [8, 9] has been employed in several prior de-hazing techniques [5, 6, 7] to describe the hazy image deterioration process by Eq. (1):

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad \text{.....(1)}$$

Where I(x) and J(x) are the degraded images and the clear images, respectively. A represents global atmospheric light; $t(x) = e^{-} \beta d(x)$ is the transmission map, where β and d(x) represent scene depth and atmospheric scattering characteristics, respectively. The primary goal of early prior-based dehazing techniques, which have advanced significantly, is to manually estimate the global atmospheric light A and the medium transmission map t(x). These prior-based techniques [1, 2, 5, 7] must, however, be more in line with practice because they typically call for laborious iterative optimization and manually created priors. Although it is commonly recognized that environmental variables like temperature, humidity, and altitude affect the creation of haze, it might be difficult to represent hazy pictures with a simple model. Consequently, when handling complicated scenarios like non-

homogeneous and dense fog conditions, these prior-based approaches may lead to estimated mistakes. It is unable to handle very white movies and photos with dense white regions presents difficulties in accurately calculating attenuation [5, 12]

However, clean images are necessary for surveillance since blurry or foggy images prevent us from seeing objects, but simply clearing the image won't fix the issue. We also require an object detection technique. Understanding several aspects of object recognition and haze removal, such as the number of objects, their distance from the item, their density of fog, and the real-time environmental conditions, is necessary. Therefore, in order to fix our problem, we must maintain these features.

In this research, we propose to use image clearing, or the removal of haze or fog from an image, to identify an item for surveillance in a single model. Utilizing Multi-Object Detection with Enhanced YOLOv2 and LuNet Algorithms [11] and WaveletFormerNet, a Transformer-based wavelet network [11] designed for real-world non-homogeneous dense fog conditions. When these methods are combined, they can better handle the intricacies of hazy surroundings, improving both visibility and object detection precision. The effectiveness of the technique is proven through rigorous testing, highlighting its potential to enhance the operation of monitoring systems in difficult weather situations.

We use a transformer-based wavelet network called WaveletFormerNet [10], which was created especially for non-homogeneous fog situations that arise in real life. By introducing the WaveletFormer and IWaveletFormer blocks, this novel technique minimizes the loss of texture detail while maintaining picture quality. WaveletFormerNet is a lightweight technique that effectively gathers multi-frequency information by utilizing parallel convolution inside Transformer blocks. Furthermore, to improve feature extraction, a feature aggregation module (FAM) is presented, which captures long-range relationships among data at several levels. With the use of frequency information, WaveletFormerNet offers an end-to-end wavelet reconstruction network that successfully addresses picture dehazing problems in challenging actual circumstances. Wide-ranging tests carried out on artificial and real-world datasets confirm WaveletFormerNet's efficacy. Adding object detection is crucial to further improve surveillance capabilities. In order to comprehend and analyze surveillance film more effectively, object identification is essential for distinguishing item kinds and establishing their distances from one another. More complete surveillance systems are made possible by this added capabilities, which enhances threat identification, situational awareness, and decision-making. For better accuracy and efficacy in real-world conditions, surveillance systems can combine object detection with fog removal techniques like WaveletFormerNet.

Additionally, we employ LuNet and Enhanced YOLOv2 algorithms for Multi-Object Detection in order to detect objects.[11] Numerous applications, such as intelligent transportation systems, robot navigation, video surveillance, and video analytics, benefit from multiple object tracking (MOT) in videos. Even though a lot has been learned since the first research, visual tracking of multiple items is still difficult because of ambient noise, occlusions in measurements, variable object counts, and similarity in appearance across objects. In order to identify moving objects before exchanging information, the study concentrated on three important processes: feature extraction, object detection, and classification. This paper suggests a deep reinforcement learning and LuNet-based multi-object video identification technique. Initially, the improved "you only look once" version 2 (YOLOv2) picks up a lot of items. In this study, a YOLOv2 base network modified by decreasing the metrics and using LuNet in their place. The LuNet network is utilized in the improved model's feature extraction process to pull out the most anticipated qualities from the picture. Additionally, the LuNet design of the underlying network makes the model small. This method uses the MOT20 vehicle benchmark dataset to assess the performance of the strategy against many state-of-the-art methods.

2. Related works

2.1. Image dehazing

Traditional prior-based approaches and learning-based methods are the two main groups into which the currently available picture dehazing techniques fall.

2.1.1Traditional methods

The majority of prior-based dehazing techniques [1-5, 7, 12, 14] estimate the transmission map using clear and fuzzy pictures and then recover the haze-free images using ASM. The dark channel prior (DCP) was introduced by He et al. [5, 12], based on the assumption that low-intensity values are frequently present in at least one channel in the picture patches of haze-free outdoor photos. Zhu et al. [7] suggested using colour attenuation prior (CAP) to estimate the scene depth as reliable prior information in order to overcome the disparity in brightness and saturation of hazy pictures. However, these algorithms' performance is intrinsically limited by the particular situation, and if the scenario does not fulfil these priors, they can result in undesired colour distortions. But WaveletFormerNet [10] can reconstruct pictures with more resolution because it makes use of the complementing benefits of previous- and deep learning-based methods.

2.1.2. Deep learning methods

Deep learning methods have been put up recently to address the issue of dehazing underwater images. The restoration of underwater photographs using these approaches has yielded encouraging results. Three categories can be used to group them: (i) Transformer-based techniques, and (ii) CNN-based techniques.

2.1.2.1. CNN-based techniques

In recent years, a broad variety of CNN-based techniques [4] have taken center stage. In order to estimate t(x) utilizing a coarse-scale network and local optimization, Ren et al. [4] recommended the use of MSCNN. Reiterating ASM, Li et al. [6, 7, 8] suggested AODNet to learn each fuzzy picture and its t(x). All these techniques, however, rely on ASM, and the dehazing outcomes are frequently skewed in terms of color. To reduce the bottleneck issue that conventional multi-scale approaches have

The training of these supervised approaches requires a high amount of data pairs, which is the reason for their great performance. More significantly, these methods are nearly exclusively learned on synthetic pictures, making them poorly suited for real-world image dehazing.

2.1.2.2 Transformer-based methods

Transformer-based models [10] have recently produced good results for dehazing photos. The WaveletFormer and IWaveletFormer blocks are used in this process to prevent texture detail loss and preserve image quality. The lightweight mechanism's multi-frequency information is captured by the parallel convolution in the Transformer blocks. We provide a feature aggregation module (FAM) to further improve WaveletFormerNet's feature extraction capabilities by capturing long-range relationships among data of various levels. We introduce an end-to-end wavelet reconstruction system called WaveletFormerNet.

2.2. Object detection

2.2.1. R-CNN

R-CNN is the approach to detect and count vehicles. Although this technique can accelerate the detection process, it has lower detection accuracy than other traditional method, most importantly these methods are incapable of detecting distant object

2.2.2. Enhanced YOLOv2

An improved YOLOv2 algorithm [13] was presented by Malik Javed Akhtar et al. (2022) to identify and recognize vehicles in surveillance films. In this post, the YOLOv2 main network is updated by using DenseNet in place of fewer parameters. The model is more compact due to the dense architecture of the underlying network. Since all layers have direct connections, DenseNet-201 is used as the base network in this study. This helps to get pertinent data from the first layer and deliver it to the last layer. The Kaggle and KITTI datasets were used to train the model, while the Pascal VOC and MS COCO datasets were used to cross-validate its performance.

2.2.3. Enhanced YOLOv2 and LuNet

The use of LuNet and Enhanced YOLOv2 algorithms in surveillance videos is suggested by T. Mohandoss and J. Rangaraj [11]. In the Yolov2 algorithm. In this study, a real-time video dataset known as MOT20 was collected and converted into minuscule video frames. The video/image frame of the recognized object will provide recommended layouts and noise detection for different moving objects. The noise is eliminated and smoothed with a Kalman filter. The filter makes use of noise measurements gathered over time to anticipate future observations and assess the model's parameters. Forecasts, measurements, and updates based on forecasts and comparisons at every level are all possible with this filter. Mathematical estimators may be used to anticipate and update the state of various linear processes. In order to improve performance, the YOLOv2 network predicts and categorizes bounding box categories using binary cross-entropy loss rather than multiple labels. This effort replaced the amount of metrics in the YOLOv2 base network with LuNet, so altering it. This study uses LuNet technology for feature extraction in the upgraded model to extract the most anticipated qualities from the picture. Additionally, the LuNet design of the underlying network makes the solution compact. This study employs LuNet as the base network, which enables us to harvest important features from the first layer and transmit them to the last layer due to the direct connections between all layers.

2.3. Proposed method

YOLO (You Only Look Once) object detection, like many other computer vision algorithms, relies heavily on the quality of the input images. Higher-quality images generally provide clearer and more distinguishable visual features, which can aid in the accurate identification and localization of objects. Conversely, low-quality images with noise, blurriness, or other imperfections can hinder the detection process, potentially leading to slower performance and reduced accuracy in predicting object locations and classifications. Therefore, ensuring good quality data inputs is crucial for optimizing the performance of YOLO-based object detection systems.

To address the limitations inherent in individual methods for object detection, we've devised a new approach that merges two distinct techniques. Recognizing that the quality and quantity of images are pivotal factors, we've devised a strategy that leverages the strengths of both methods. Initially, we employ an image fog removal technique to enhance the clarity of the visual data. Subsequently, this refined image is utilized for object detection, thereby augmenting the accuracy of predictions. By combining these methodologies, we aim to achieve superior results by optimizing both image quality and the effectiveness of object detection algorithms.

This paper offers a combined method for removing fog from surveillance images using WaveletFormerNet, a Transformer-based wavelet network designed for real-world non-homogeneous dense fog scenarios; this transformer method use the wavelet transform method. It also uses Multi-Object Detection with Enhanced YOLOv2 and LuNet Algorithms to detect objects. The combination of these methods known as “ClearDetect”, can better handle the intricacies of hazy surroundings, improving both visibility and object detection precision. The effectiveness of the proposed technique is proven through rigorous testing, highlighting its potential to enhance the operation of monitoring systems in difficult weather situations.

3. Methodology of “ClearDetect”

The WaveletFormerNet's [10] construction is shown in Fig. 1. WaveletFormerNet's encoding and decoding are both based on the WaveletFormer and IWaveletFormer block; however, DWT and IDWT, respectively, are used in place of down sampling and up sampling in the encoding and decoding segments. While the WaveletFormer and IWaveletFormer block serves as the foundational block of the network, primarily integrating the wavelet transform and Swin Transformer, we don't utilize these two pre-existing tools directly; instead, we enhance them. The wavelet transform is utilized to convert the features into the frequency domain. Then, the frequency information is employed to direct WaveletFormerNet in order to retrieve the image's texture and structural details. Additionally, the receptive field brought on by the Swin Transformer is reduced by the parallel convolution. The WaveletFormer and IWaveletFormer block's structure also lessens the details brought on by issues with down sampling loss and other issues. Additionally, we suggest a Feature Aggregation Module (FAM) that combines various feature information levels to preserve picture resolution and improve the network's receptive field. In

order to extract features in various receptive fields, we lastly modify an Atrous Spatial Pyramid Pooling (ASPP) module in the network and adjust dilated convolution with different expansion rates (rate = 3, 6, 9).

(A) WaveletFormer and IWaveletFormer block

WaveletFormer and IWaveletFormer Blocks use DWT and IDWT to decompose the images from the frequency domain point of view, respectively, and the feature maps are used as inputs to the Transformer module with parallel convolution.

(i) Frequency decomposition of images

Fig. 2 illustrates the detailed structure of the WaveletFormer block, adopting frequency information to guide the network in reconstructing a clear image. We can observe that the input image FDWTin can be divided into the low- and high-frequency details separated into four different frequency sub bands: the low-frequency band FLL, the horizontal sub band FLH, the vertical sub band FHL and the high-frequency sub band FHH on the diagonal edge of the original image. This mechanism alleviates detail and colour loss and provides a better balance between network processing efficiency and image recovery performance. For the 2D discrete wavelet transform, we import the pytorch_wavelets package and use Daubechies wavelet basis functions.

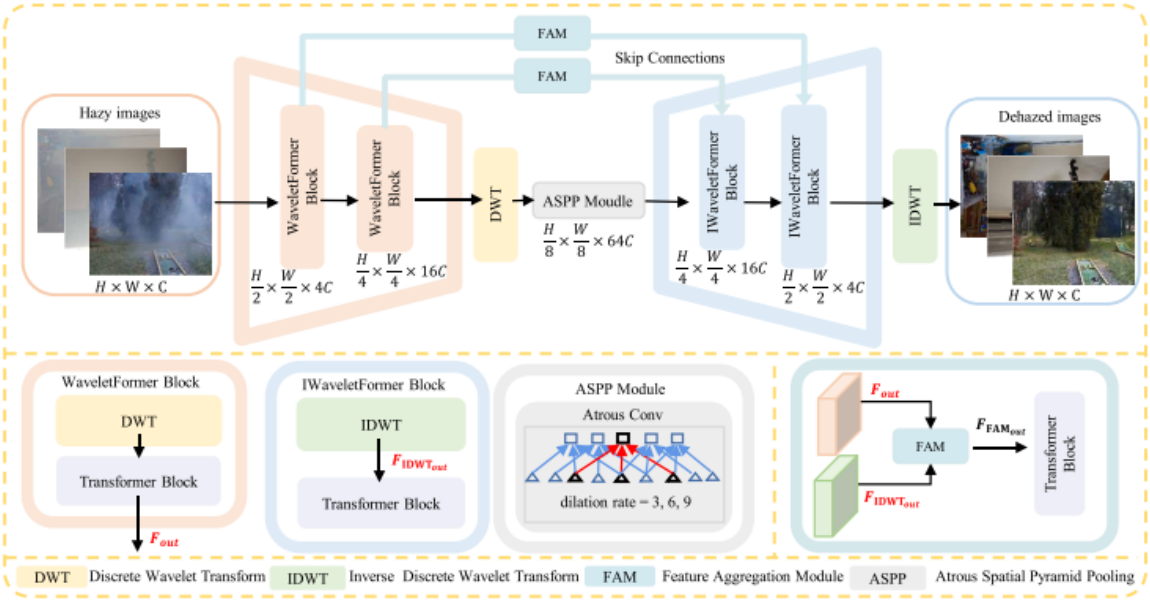

Fig. 1. The schematic illustration of the WaveletFormerNet. WaveletFormer block and IWaveletFormer block consist of DWT and IDWT and Transformer block respectively, and IDWT is the reverse process of DWT.[10]

(ii) Parallel convolution in Vision Transformer

According to the mechanism, given an input feature map X∈ ℝb×h×w×c, we project X to Q, K, V (query, key, value), and we compute the attention function for a set of queries simultaneously and pack them into a matrix Q; so that the computed output matrix can be described as:

Attention (Q, K, V) = Softmax (QK^T/√ dk) V ……. (2)

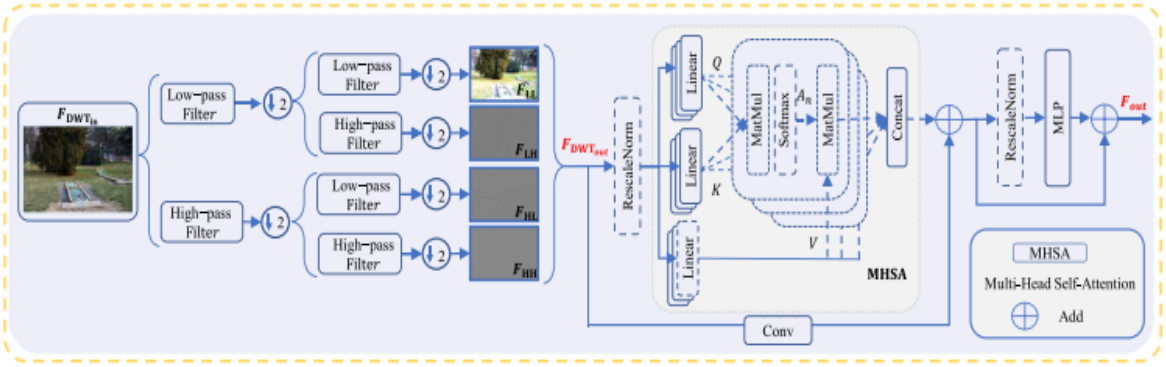

Fig. 2. The architecture of WaveletFormer and IWaveletFormer blocks. Note: The WaveletFormer block and the IWaveletFormer block have the same structure: they utilize DWT and IDWT to substitute down sampling and up sampling, respectively. [10]

(iii) Feature aggregation module

As the key component of the FAM, the Multi-Head Cross-Attention (MHCA) introduces the feature map Fout of the WaveletFormer block and the feature map FIDWTout from IWaveletFormer block into the MHSA for processing, the computed weight values Y to be rescaled by the sigmoid activation function. The resulting feature tensor Z will be summed with the feature map Fout to obtain the high-level feature tensor. In addition, the feature tensor X produced by FIDWTout is also obtained at a high-level feature tensor after operations such as ReLu. Finally, we concatenate these two high-level feature tensors as the FMHCAout. Therefore, FFAMout can be expressed as:

FFAMout = MLP (MHCA (Fout, FIDWTout)) +Fout ……. (3)

The feature aggregation module (FAM) introduces the feature map Fout of the WaveletFormer block and the feature map FIDWTout from IWaveletFormer block for processing. Fig. 3 illustrates that FAM is a link between the encoding and decoding stages, guiding our WaveletFormerNet to generate images with more crisp textures and rich features.

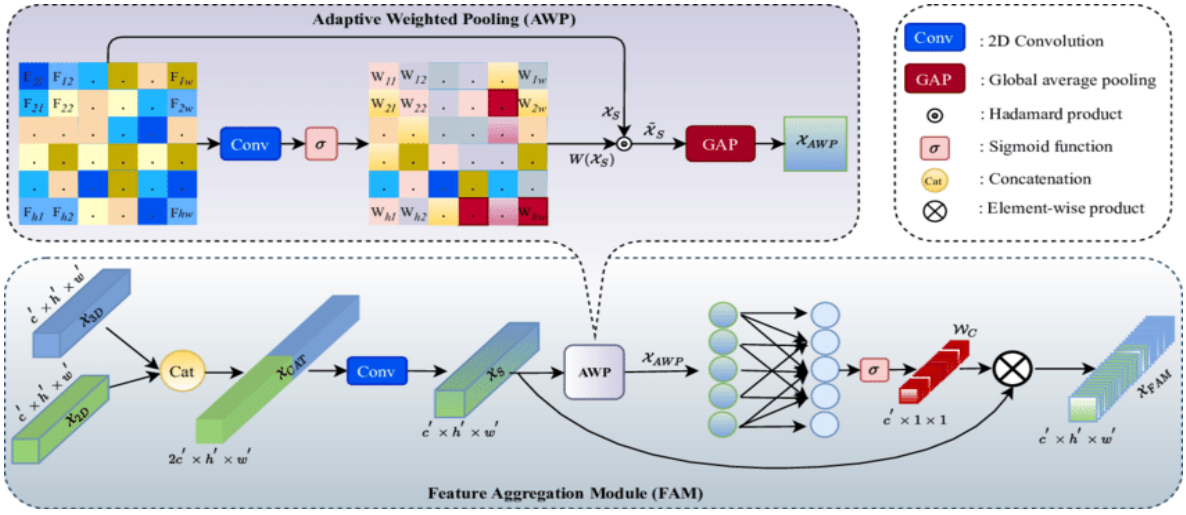

Fig. 3.Architecture of feature aggregation module employed in the WaveletFormerNet [10]

(iv) Training datasets

We extensively and comprehensively evaluated our model and compared SOTA methods on real-world and synthetic datasets in the same experiment setting. [10]

(a) Real-world datasets

We use the following four datasets to evaluate our experiments: the NTIRE 2018 image dehazing dataset (I-Haze), the outdoor NTIRE 2018 image dehazing dataset (O-Haze), a benchmark for image dehazing with dense-haze and haze-free images (Dense-Haze), and the NTIRE 2020 dataset for non-homogeneous dehazing challenge (NH-Haze).

(b) I-Haze and O-Haze

They contain 25 and 35 hazy images (size 2833 × 4657 pixels) respectively for training. Both datasets contain 5 hazy images for validation along with their corresponding ground truth images. We used training data for training and validation data for the test.

(c) Dense-Haze

It contains 45 hazy images (size 1200 × 1600 pixels) for training, 5 hazy images for validation and 5 more for testing with their corresponding ground truth images. We have performed training on training data and tested our model with test data.

(d) NH-Haze

It contains 45 hazy images (size 1200 × 1600 pixels) for training. We selected 40 pairs of data for training and the rest for testing.

(e) Real-world datasets expansion

Authors use the same training strategy and dataset expansion process for all four real-world datasets. Specifically, we randomly cropped the original images into square patches of 512 × 512 pixels; these patches are not identical for every epoch. To augment the training data, we implemented random rotations (90, 180, or 270 degrees) and random horizontal flips when processing the training data. This step allows these small real-world datasets to be expanded into larger datasets that are more efficient and more suitable for data-driven methods of training experiments.

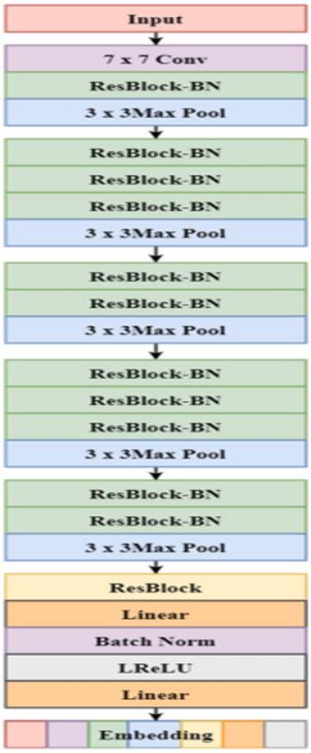

Fig. 4. Architecture of feature extractor [10].

(B) Object detection

(i) Dataset

The MOT20 benchmark, which comprises of 8 new sequences depicting crowded and difficult settings, is used in this study. The dataset used for the new benchmark was selected with care to test detectors and trackers in high-traffic environments. Certain new sequences feature pedestrian densities of 246 per frame in comparison to earlier challenges. Eight sequences were made for this study, half of which were utilized for testing and the other half for training. Test sequence annotations are kept confidential to avoid approaches being too biased toward particular sequences. Three different sets were used to film the scenes. Every scene was shot in many takes, utilizing both the train and test sets. But one of the situations was reserved for testing in order to see how well the procedure generalized. The new data contains around three times more bounding boxes for training and testing than MOT17. Ten times greater than the initial baseline density, each sequence was captured in high quality from an elevated vantage point, with an average pedestrian density of 246 per frame.

(ii) 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 over fitting, 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) to extract spatial data and a Recurrent Neural Network (RNN) 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 HASTIDS, 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) , followed by pooling and resampling to remove I 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 over fitting, 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 to extract object appearance features. LuNet's input is a 128 × 64 × 128 × 64 image patch.

The network employs LeakyReLU as the activation function for robust optimization, multiple 3 × 33 × 3 max pooling, and two-stride instead of stride convolution. Fig. 4 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).

(iii) Object detection using YOLOv2

It is a development over YOLO. To increase YOLO's speed and detection accuracy, YOLOv2 applies new ideas and incorporates judgments from previous training challenges. YOLOv2 has six stages, which are addressed as follows, detailed diagram in fig.5.

- ***Batch normalization (BN):*** Each mini-batch's mean and variance are calculated and used for activation. After that, the activations are normalized using a zero mean and a one standard deviation for each mini batch. Lastly, the same distribution is used to sample the components of each mini-batch. We call this process "batch normalization." It produces the same distribution of activations.
- ***High-resolution classifier:*** The input resolution of (224 × 224) is used by the YOLO backbone. In YOLOv2, the input resolution has been improved to (448 × 448). In order to handle the new resolution input of the object detection job, the network must be adjusted. Consequently, a single-resolution picture (448 × 448) and ten epochs were used to make specific modifications to the classification network in YOLOv2, increasing the average accuracy (mAP) by 1%.
- ***Anchor box convolution:*** As previously said, Faster RCNN forecasts bounding boxes by first generating region proposals, which are parameterized with respect to this anchor box. YOLOv2 makes use of this estimating technique. Next, we project the class and object scores for every bounding box prediction. While mAP decreased by 0.3%, withdrawals increased by 7%.
- ***Anchor box size and aspect ratio prediction:*** The LuNet method is used by the YOLOv2 to train the bounding boxes to get higher priors. The anchor box's centre is then determined by using this backdrop. Using the clustering data, estimate the anchor box's dimensions and aspect ratio. The method improves the accuracy of detection.
- ***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.
- ***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 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.

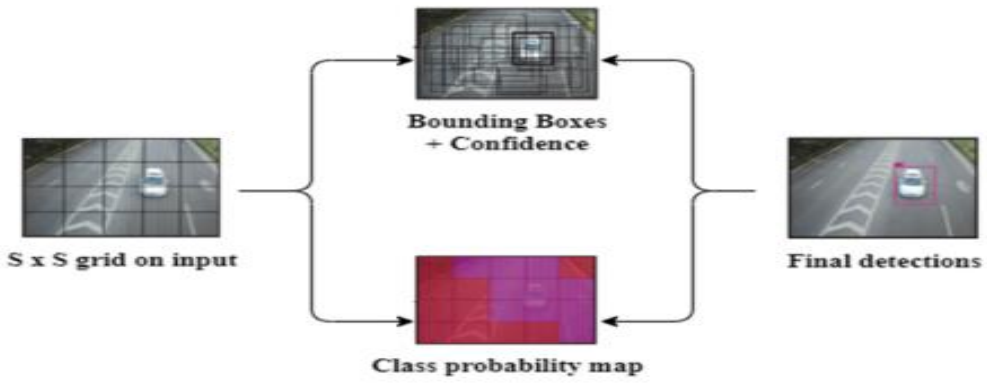

Fig.5. Yolov2 Object detection diagram [13]

(iv) Object detection using enhanced YOLOv2-LuNet

This research develops an enhanced multi-object detection technology based on the enhanced YOLOv2-LuNet [11] 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 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 centre of a labelled object lies in a specific zone, the region will be used to predict the object. Using this for object detection after clearing the image or video frames can helps to achieve better results.

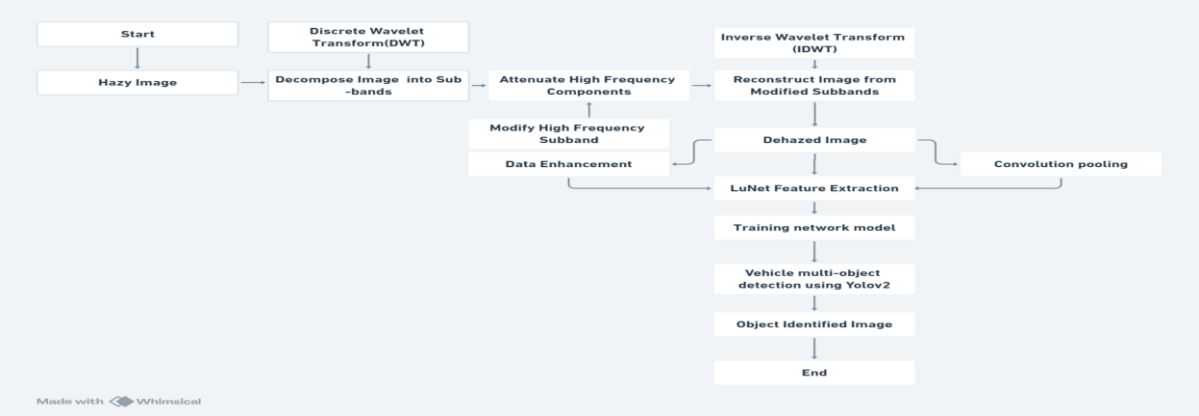

Fig.6.flow chart for the complete process from start to end.

Fig.7.1.Example Original image

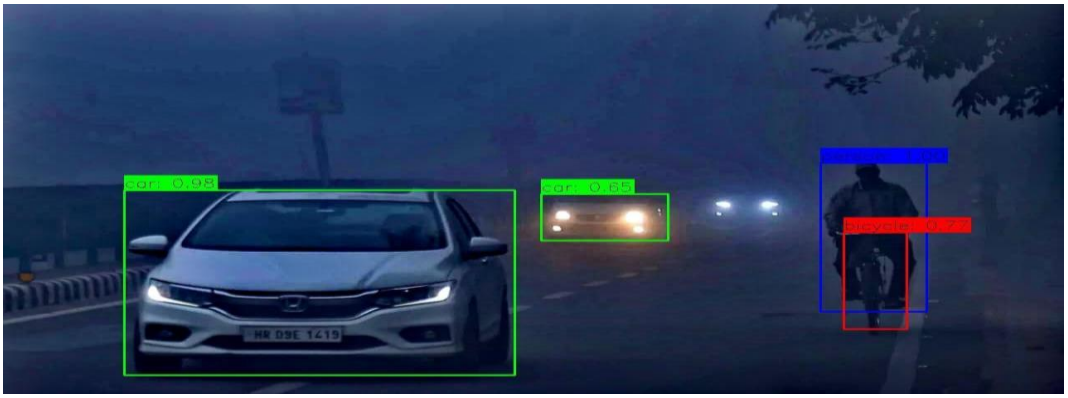

Fig.7.2.Example image after Dehazing and object detection

Fig.7.3.Example Original image

Fig.7.4.Example image after Dehazing and object detection

4. Results and discussion

In Fig.6. we have the complete flowchart from dehazing to object detection and in the Fig. 7.1 and fig.7.3 we have an example image of the vehicle in the foggy environment which make the identification of object very hard therefore, In this work, an image dataset for traffic analysis is used, proposed method “clear detect" clear the foggy image using wavelet transform and YOLOv2-LuNet processing is depicted in Fig. 7.2 and Fig.7.4. The metrics examined are accuracy, precision, recall, ground truth (GT), detection rate (DET), true positive rate (TP), and false positive rate.

(a)Accuracy

Accuracy measurement aids in determining how well the wavelet transformer and LuNet classifier detects objects in image. 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 proposed approach “cleardetect”. The proposed method’s accuracy values are evaluated against current methods by employing feature masking in image. The existing technologies’ accuracy ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 75%, 80%, 84%, 88%, and 88%, respectively. An accuracy of 94% was observed in the case of the proposed model. Fig. 8 depicts that the proposed 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 image data than other models.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100.0	66.0	70.0	80.0	81.0	80.0	90.0
200.0	69.0	72.0	81.0	83.0	82.0	92.0
300.0	71.0	74.0	82.0	85.0	85.0	92.5
400.0	73.0	79.0	83.0	87.0	86.0	93.5
500.0	75.0	80.0	84.0	88.0	88.0	94.0

Table 1: Accuracy of the methods

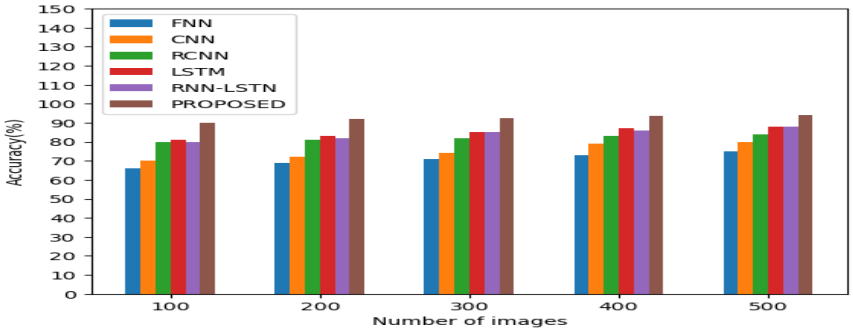

Fig.8. This shows the Accuracy of the traditional method and proposed method.

(b)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%, 74%, 85%, 83%, and 83%, respectively. Fig. 9 depicts that the proposed method has a 96% 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.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100	62	65	73	75	76	90
200	64	66	76	74	78	91
300	66	68	79	79	80	93
400	68	69	81	81	82	95
500	70	74	85	83	83	96

Table 2: Precision of the methods

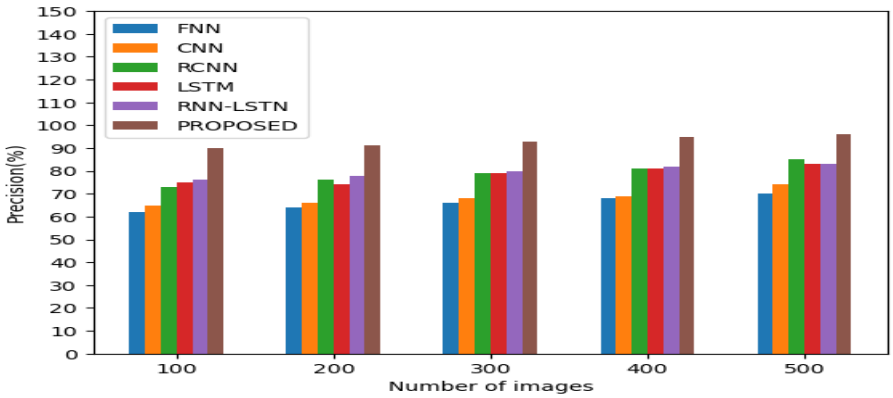

Fig.9. This shows the Precision of the traditional method and proposed method.

(c) Recall

Recall is the proportion of true positive objects the model successfully detects out of all the objects in the images. It ensures the model detects as many relevant objects as possible, which is critical for thorough image 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 proposed method is depicted in Table 3. Fig. 10 depicts that the proposed 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 58%, 65%, 73%, 75%, and 84%, respectively. A recall of 93% was observed in the case of the proposed model. When the number of images increased so does recall value.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100	52	60	65	66	75	80
200	54	61	66	69	79	84
300	55	63	69	71	80	88
400	56	64	70	73	82	90
500	58	65	73	75	84	93

Table 3: Recall of the methods

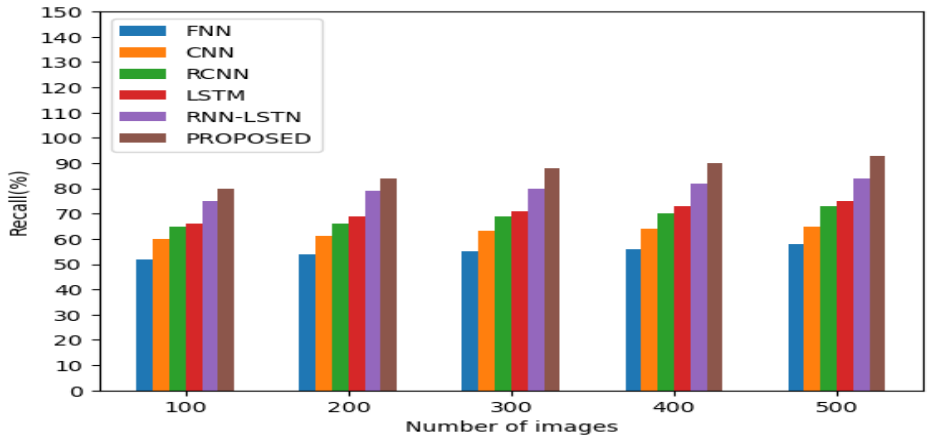

Fig.10. This shows the Recall of the traditional method and proposed method.

(d) True Positive (TP)

True positives are cases in which the model correctly identifies and localizes objects of interest in images. 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. 11 depicts that the proposed method has a 90% maximum TP value. The proposed approach yielded the true positive value (90%).The existing technologies' TP ratings for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 56%, 59%, 62%, 66%, and 68%, respectively. A TP of 90% was observed in the case of the proposed model.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100	48	50	56	56	60	73
200	50	52	54	58	61	80
300	53	55	60	60	63	83
400	53	57	63	63	66	86
500	56	59	62	66	68	90

Table 4: True positive values of the methods

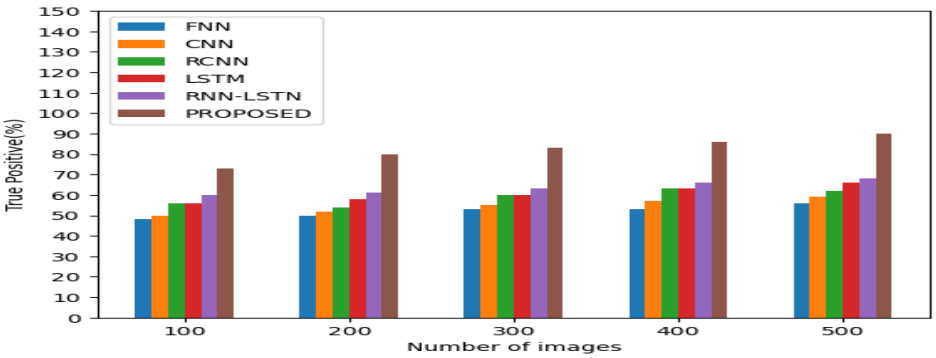

Fig.11. This shows the True Positive values of the traditional method and proposed method.

(e) 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. 12 depicts that the proposed method has 89% FP value.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100	50	52	58	60	60	69
200	52	56	60	62	64	74
300	54	60	63	64	66	80
400	55	61	66	68	70	85
500	57	64	68	71	75	89

Table 5: False Positive of the methods

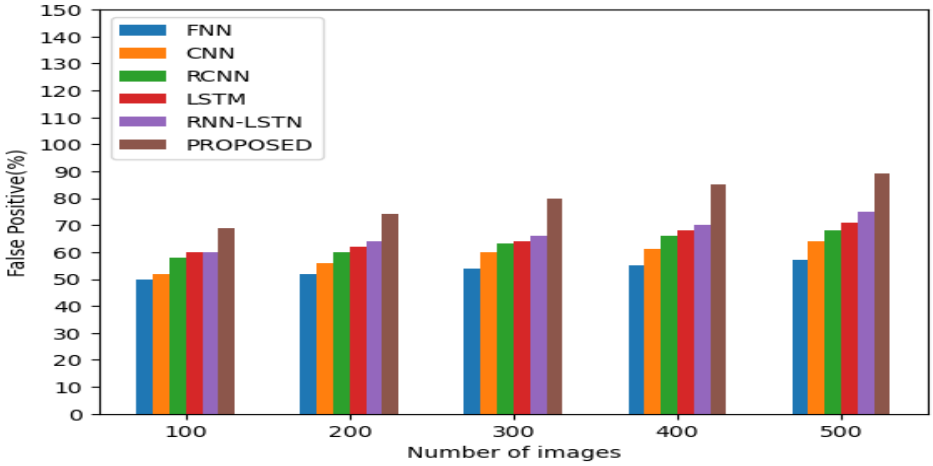

Fig.12. This shows the False Positive of the traditional method and proposed method.

(f) 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 proposed technique.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100.0	63.0	70.0	70.0	76.0	78.0	80.0
200.0	65.0	72.0	72.0	77.0	79.0	83.0
300.0	67.0	73.5	74.0	80.0	82.0	87.0
400.0	69.0	75.0	79.0	83.0	84.0	89.0
500.0	71.0	78.0	83.0	84.0	86.0	91.0

Table 6: Ground Truth of the methods

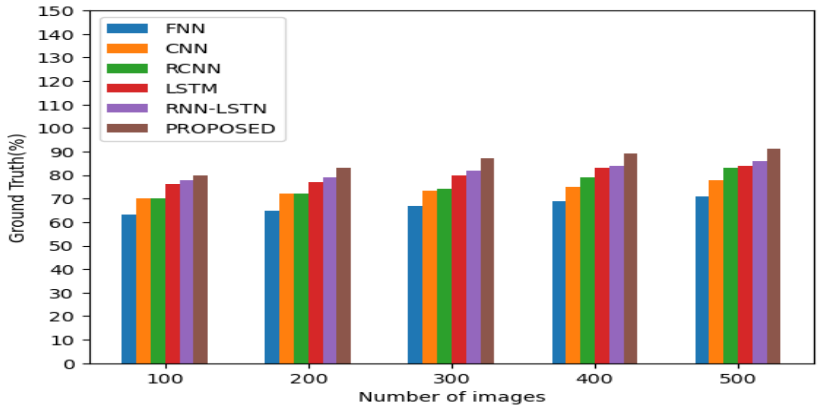

Fig.11. This shows the Ground Truth of the traditional method and proposed method.

(g) Detection rate

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 and images.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 (89%). The existing technologies' detection rates for FFNN, CNN, RCNN, LSTM, and RNN-LSTM are 53%, 60%, 61%, 70%, and 76%, respectively. Comparisons indicate that it outperforms existing methods.

Number of images	FNN	CNN	RCNN	LSTM	RNN_LSTN	PROPOSED
100	40	50	52	60	65	67
200	46	52	54	62	66	76
300	48	58	56	63	70	81
400	50	59	59	66	72	83
500	53	60	61	70	76	89

Table 11: Detection rate of the methods

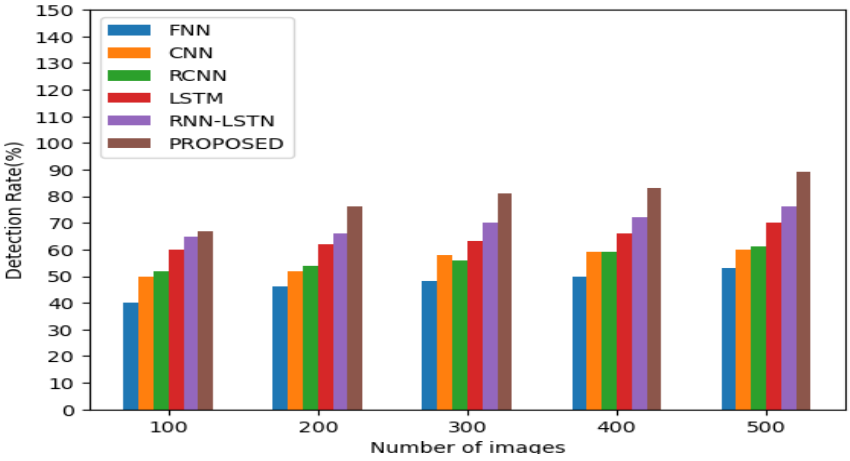

Fig.12. This shows the Detection rate of the traditional method and proposed method.

5. Conclusion

By merging cutting-edge picture enhancement and object identification algorithms, the proposed "ClearDetect" approach offers a notable improvement in object detection under cloudy environments. It solves the problems caused by non-homogeneous thick fog by using a Transformer-based wavelet network called WaveletFormerNet, which improves image clarity. The preprocessed picture is then used as input for the improved YOLOv2 and LuNet multi-object detection algorithms, which significantly increase object recognition accuracy and precision.

The ClearDetect approach has proven to be effective in greatly reducing the impacts of fog via extensive testing and assessment, which enhances vision and the functionality of object detecting systems during inclement weather. This combined strategy produces better outcomes by maximizing image quality and utilizing the advantages of cutting-edge detection algorithms.

To sum up, the ClearDetect technique presents a strong way to improve surveillance systems' performance in difficult conditions, guaranteeing more accurate and dependable object identification. Improved operational efficacy in real-world applications is made possible by the combination of advanced object identification algorithms with wavelet-based picture augmentation, especially in monitoring systems that are subjected to hazy or foggy situations.

Refrences

[1] Jiang, K., Wang, Z., Yi, P., Jiang, J., Xiao, J., & Yao, Y. (2018). Deep distillation recursive network for remote sensing imagery super-resolution. Remote Sensing, 10(11), 1700.
[2] Kumari, A., & Sahoo, S. K. (2024). A new fast and efficient dehazing and defogging algorithm for single remote sensing images. Signal Processing, 215, 109289.
[3] Rasti, P., Uiboupin, T., Escalera, S., & Anbarjafari, G. (2016). Convolutional neural network super resolution for face recognition in surveillance monitoring. In Articulated Motion and Deformable Objects: 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings 9 (pp. 175-184). Springer International Publishing.
[4] Wang, Z., Yi, P., Jiang, K., Jiang, J., Han, Z., Lu, T., & Ma, J. (2018). Multi-memory convolutional neural network for video super-resolution. IEEE Transactions on Image Processing, 28(5), 2530-2544.
[5] He, K., Sun, J., & Tang, X. (2010). Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence, 33(12), 2341-2353.
[6] Zhu, Q., Mai, J., & Shao, L. (2015). A fast single image haze removal algorithm using color attenuation prior. IEEE transactions on image processing, 24(11), 3522-3533.
[7] More, V. N., & Vyas, V. (2022). Removal of fog from hazy images and their restoration. Journal of King Saud University-Engineering Sciences.
[8] McCartney, E. J. (1976). Optics of the atmosphere: scattering by molecules and particles. New York.
[9] Narasimhan, S. G., & Nayar, S. K. (2002). Vision and the atmosphere. International journal of computer vision, 48, 233-254.
[10] Zhang, S., Tao, Z., & Lin, S. (2024). WaveletFormerNet: A Transformer-based wavelet network for real-world non-homogeneous and dense fog removal. Image and Vision Computing, 146, 105014.
[11] Mohandoss, T., & Rangaraj, J. (2024). Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms in Surveillance Videos. e-Prime-Advances in Electrical Engineering, Electronics and Energy, 8, 100535.
[12] Dwivedi, P., & Chakraborty, S. (2023). Single image dehazing using extended local dark channel prior. Image and Vision Computing, 136, 104747.
[13] Akhtar, M. J., Mahum, R., Butt, F. S., Amin, R., El-Sherbeeny, A. M., Lee, S. M., & Shaikh, S. (2022). A robust framework for object detection in a traffic surveillance system. Electronics, 11(21), 3425.
[14] Saini, H., Agarwal, M. M., Govil, M. C., & Sinha, M. (2021). Design of fuzzy controlled routing protocol to save energy in ad hoc networks. International Journal of Services Technology and Management, 27(1-2), 51-71.