Topic: Fog removal and object detection for vehicles for surveillance using WaveletFormerNet and Enhanced YOLOv2 and LuNet Algorithms

# Abstract:

Although deep convolutional neural networks are incredibly good at removing fog, they must be able to process images captured in real weather conditions such as fog or patchy cloud cover. However, in the real world, blur is difficult to classify, and as map resolution or image resolution is reduced, this reduction will lead to color mismatch or loss of content in the output results. Additionally, further stacking of convolutional blocks will increase the complexity of the model. In addition to the difficulty of obtaining sufficient training data, deep learning techniques for fog image processing can also suffer from overfitting, which can limit the potential of the model and cause problems in its practical use in real-world scenarios. Considering these issues we use WaveletFormerNet, a Transformer-based wavelet network tailored for real-world fog image processing. By integrating discrete wavelet transform into Vision Transformer via WaveletFormer and IWaveletFormer blocks, it addresses texture loss and color distortion induced by subsampling. Parallel integration within the Transformer block reduces computational load while maintaining multi-frequency data capture. The Feature Aggregation Module (FAM) preserves image resolution, enhancing fog cloud image restoration. Extensive testing on real-world weather data demonstrates WaveletFormerNet's superiority over state-of-the-art methods, with promising results in real dust removal and application tests, confirming its potential for enhancing computer vision applications.

But fog removal alone won't fully address the challenges (object identification, distance of object from person, shape of object). We also require advanced object detection and tracking systems, particularly in surveillance and vehicular applications. This is where multi-object tracking (MOT) becomes crucial. By integrating LuNet and deep reinforcement learning with enhanced versions of YOLOv2, we enhance the ability to identify and track objects effectively. This improved approach considers factors such as occlusions, environmental noise, and diverse object appearances, ensuring superior performance in real-world situations.

Top of Form

This project presents an innovative approach to fog removal in surveillance videos using WaveletFormerNet, a Transformer-based wavelet network tailored for real-world non-homogeneous dense fog scenarios and also detecting the objects by leveraging Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms. The combination of these techniques addresses the complexities of foggy environments, enhancing visibility and object detection accuracy. Through extensive experimentation, the efficacy of the proposed methodology is demonstrated, showcasing its potential for improving surveillance systems' performance in challenging weather conditions.

Introduction:

Haze is a common atmospheric phenomenon that causes distortion and degradation of images. Image dehazing techniques are significant for many computer vision tasks, such as remote sensing processing [1,2,7] and video analysis and recognition [3,4]. In recent years, image dehazing has been a hot research topic in computer vision and image processing, serving as an essential low-level image recovery task and a pre-processing step for high-level vision tasks. Many previous dehazing methods [5,6,7] have used the classical atmosphere scattering model (ASM) [8,9] to characterize the degradation process of hazy images by

Eq. (1): I(x) = J(x)t(x) + A(1 − t(x)) …………………………………………………………………………………………………(1)

Where I(x) and J(x) are the degraded images and the clear images, respectively. A represents global atmospheric light; t(x) = e− βd(x) is the transmission map, where β and d(x) represent atmospheric scattering parameters and scene depth, respectively. The main idea of early priorbased dehazing methods is to estimate the medium transmission map t(x) and the global atmospheric light A by handcraft; these methods have made significant progress. However, these prior-based methods [1, 2, 5, 7] usually require time-consuming iterative optimization and manually designed priors; they need to be more consistent with the practice. As is well known, haze formation is related to natural factors, such as altitude, temperature, and humidity, but it is challenging to express hazy images with a simplistic model. Therefore, these prior-based methods may result in estimation errors when dealing with complex scenes such as the non-homogeneous and dense fog weather. , it can’t Handle highly white videos and images with dense white regions presents difficulties in accurately calculating attenuation [5,12]

But, For the surveillance we need clear image because we can’t see the object if these is haze or fog in the image but just clearing image will not solve our problem we need method to detect the object also and object detection and haze removing need many features to understand like the real time environmental conditions, distance of object, density of fog, distance of fog from the object, number of the object . therefore we need to keep these feature to solve our problem

In this paper we have idea of using image clearing or haze /fog removing from image and detect the object for the surveillance in the single model. Using WaveletFormerNet, a Transformer-based wavelet network tailored for real-world non-homogeneous dense fog scenarios and also detecting the objects by leveraging Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms [11]. The combination of these techniques addresses the complexities of foggy environments, enhancing visibility and object detection accuracy. Through extensive experimentation, the efficacy of the proposed methodology is demonstrated, showcasing its potential for improving surveillance systems' performance in challenging weather conditions.

we employ WaveletFormerNet[10], a transformer-based wavelet network designed specifically for real-world non-homogeneous fog conditions. This innovative approach introduces WaveletFormer and IWaveletFormer blocks to mitigate texture detail loss while preserving image resolution. Leveraging parallel convolution within Transformer blocks, WaveletFormerNet efficiently captures multi-frequency information, maintaining a lightweight mechanism. Additionally, a feature aggregation module (FAM) is introduced to enhance feature extraction by capturing long-range dependencies among information at different levels. WaveletFormerNet presents an end-to-end wavelet reconstruction network guided by frequency information, effectively tackling image dehazing issues in complex, real-world scenarios. Extensive experiments conducted on both synthetic and real-world datasets validate the effectiveness of WaveletFormerNet. To enhance surveillance capabilities further, incorporating object detection is essential. Object identification provides crucial context for better understanding and analyzing surveillance footage, including determining object distances and identifying object types. This additional functionality enables more comprehensive surveillance systems, allowing for improved threat detection, situational awareness, and decision-making. By integrating object detection alongside fog removal techniques like WaveletFormerNet, surveillance systems can achieve greater accuracy and effectiveness in real-world scenarios.

And for the object detection we use Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms.[11] 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.

## Related work:

### 1. Image dehazing

The existing methods for image dehazing are broadly classified into two categories: traditional prior-based methods and learning-based methods.

#### 1***.1Traditional methods***

Most prior-based dehazing methods [1,2,3,5,7,12] use hazy and clear images to estimate the transmission map, then use ASM to recover haze-free images. He et al. [5, 12] proposed the dark channel prior (DCP), assuming that the image patches of haze-free outdoor images often have low-intensity values in at least one channel. To address the difference in brightness and saturation of hazy images, Zhu et al. [7] proposed color attenuation prior (CAP) to estimate the scene depth as solid prior knowledge. However, the specific scenario inherently limits the performance of these methods, and they may lead to undesirable color distortions when the scenario does not satisfy these priors. In contrast, WaveletFormerNet [10] can reconstruct images with richer detail by leveraging the complementary advantages of prior- and deep learning-based methods.

#### 2. Deep learning methods

Recently, deep learning techniques have been proposed to tackle the problem of underwater image dehazing. These techniques have shown promising results in the restoration of underwater images. They can be classified into three categories: (i) CNN-based methods, (ii) Transformer-based methods.

1.2.1. CNN-based methods. A wide range of CNN-based methods [4] have dominated in recent years. Ren et al. [4] proposed MSCNN to estimate t(x) using a coarse-scale network followed by local optimization. Li et al. [6, 7, 8] reiterated ASM and proposed AODNet to learn each hazy image and its t(x). However, all of these methods rely on ASM, and the dehazing results are often color-distorted. To alleviate the bottleneck problem encountered in traditional multi-scale methods

However, behind the excellent performance achieved by these supervised methods, a large number of data pairs are required for the training; more importantly, these methods are almost trained on synthetic images , which cannot be well generalized to real-world image dehazing.

1.2.2 Transformer-based methods [10].

Recently transformer based models achieved the good results in the dehazing images its process include, the WaveletFormer and IWaveletFormer blocks to alleviate texture detail loss and maintain image resolution. The parallel convolution in the Transformer blocks captures the multi-frequency information in the lightweight mechanism. We present a feature aggregation module (FAM) to capture the longrange dependencies among information with different levels and further enhance the feature extraction capability of WaveletFormerNet. We present WaveletFormerNet, an end-to-end wavelet reconstruction

### 2. Object detection

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

#### 2.1. Enhanced YOLOv2

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 [13].

#### 2.2 Enhanced YOLOv2 and LuNet

T. Mohandoss a , J. Rangaraj[11] gives the idea of using Enhanced YOLOv2 and LuNet Algorithms in Surveillance Videos. In the yolov2 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 smooth’s 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.

# Working:

Fig.1 illustrate the structure of the WaveletFormerNet is Both encoding and decoding of WaveletFormerNet are based on the WaveletFormer and IWaveletFormer block, but the difference between the encoding and decoding segments is that downsampling and upsampling are replaced by DWT and IDWT, respectively. Although the WaveletFormer and IWaveletFormer block as the base block of the network mainly combines wavelet transform and Swin Transformer, we do not directly apply these two existing tools but improve them. We use the wavelet transform to transform the features to the frequency domain and use the frequency information to guide WaveletFormerNet to recover the structural and texture details of the image. In addition, the proposed parallel convolution also alleviates the receptive field caused by Swin Transformer. This structure of the proposed WaveletFormer and IWaveletFormer block also alleviates the details caused by downsampling loss and other problems. Furthermore, we propose a Feature Aggregation Module (FAM) to maintain image resolution and enhance the receptive field of the network, combining different levels of feature information. Finally, we adapt an atrous spatial pyramid pooling module (ASPP) in the network and adjust dilated convolution with different expansion rates (rate = 3, 6, 9), obtaining features in different receptive fields.

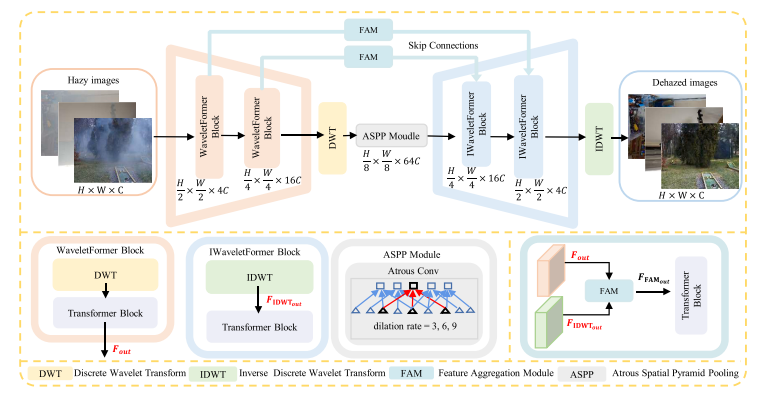


Fig. 1. The schematic illustration of the proposed WaveletFormerNet. WaveletFormer block and IWaveletFormer block consist of DWT and IDWT and Transformer block respectively, and IDWT is the reverse process of DWT.

1. 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.

2. Frequency decomposition of images

Fig. 1.2 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 subbands: the low-frequency band FLL, the horizontal subband FLH, the vertical subband FHL and the high-frequency subband FHH on the diagonal edge of the original image. This zmechanism alleviates detail and color 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.

3. Parallel convolution in Vision Transformer

According to the attention mechanism [50], 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)

And the Multi-Head Self-Attention (MHSA) [50] can be expressed as Eq. (3), where the projections are parameter matrices WQ i ∈ ℝdmodel×dk ,

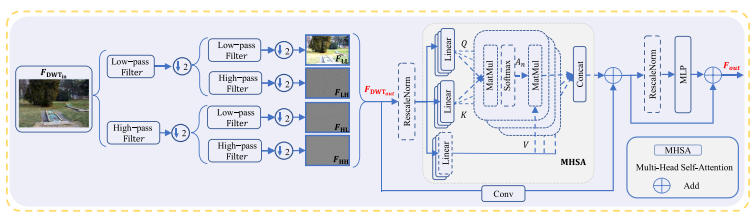


Fig. 1.2. The architecture of proposed WaveletFormer and IWaveletFormer blocks. Note: The WaveletFormer block and the IWaveletFormer block have the same structure: they utilize DWT and IDWT to substitute downsampling and upsampling, respectively.

## 3.1. Object detection[11]:

#### 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 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

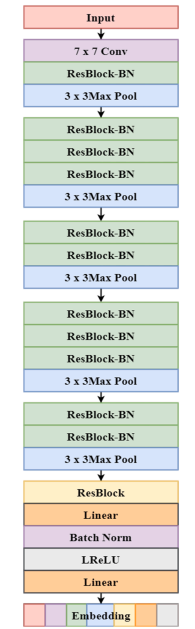
The purpose of a feature extractor in vehicle detection is to convert the original input image into a set of representative features to capture vehicle detection information. These features are fed into later stages of the detection process, like classification or bounding box regression. In general, 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 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 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 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. 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).

Fig. 2. Architecture of feature extractor.

3.3. Object detection using YOLOv2 [13]

It 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 minibatch 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. It generates the same activation distribution.

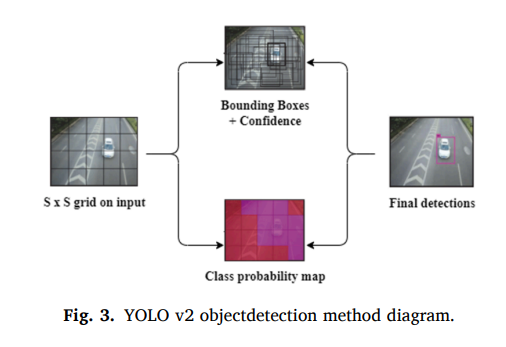
ii. High-resolution classifier: The YOLO backbone employs the (224 × 224) input resolution. The input resolution in YOLOv2 has been enhanced to (448 × 448). As a result, the network must be modified to accommodate the new resolution input of the object detection task. As a result, specific changes were 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.

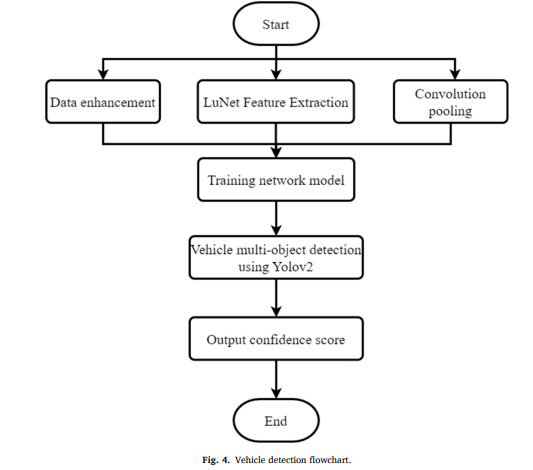
v. Fine-grained features: As previously stated, YOLO trains with images (224 × 224). The YOLO design has been tweaked to form the Yolov2 architecture. YOLOv2 is retrained using higher resolution images (448 × 448) to pinpoint tiny objects. YOLOv2 uses higher and lower resolution features throughout this 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 [11]

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 centre of a labelled object lies in a specific zone, the region will be used to predict the object.



[1] Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Jing Xiao, Yuan Yao, Deep distillation recursive network for remote sensing imagery super-resolution, Remote Sens. 10 (11) (2018) 1700.

[2] Apurva Kumari, Subhendu Kumar Sahoo, A new fast and efficient dehazing and defogging algorithm for single remote sensing images, Signal Process. 215 (2024) 109289.

[3] Pejman Rasti, Tonis Uiboupin, Sergio Escalera, Gholamreza Anbarjafari, 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, Springer, 2016, pp. 175–184.

[4] Zhongyuan Wang, Peng Yi, Kui Jiang, Junjun Jiang, Zhen Han, Lu Tao, Jiayi Ma, Multi-memory convolutional neural network for video super-resolution, IEEE Trans. Image Process. 28 (5) (2018) 2530–2544

[5] Kaiming He, Jian Sun, Xiaoou Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intel. 33 (12) (2010) 2341–2353.

[6] Qingsong Zhu, Jiaming Mai, Ling Shao, A fast single image haze removal algorithm using color attenuation prior, IEEE Trans. Image Process. 24 (11) (2015) 3522–3533.

[7] Vidya Nitin More, Vibha Vyas, Removal of fog from hazy images and their restoration,

Journal of King Saud University - Engineering Sciences, 2022, ISSN 1018-3639,

[8] Earl J. McCartney, Optics of the Atmosphere: Scattering by Molecules and Particles, New York, 1976.

[9] Srinivasa G. Narasimhan, Shree K. Nayar, Vision and the atmosphere, Int. J. Comput. Vis. 48 (3) (2002) 233.

[10]Shengli Zhang, Zhiyong Tao, Sen Lin,WaveletFormerNet : A Transformer-based wavelet network for real-world non-homogeneous and dense fog removal, mage and Vision Computing, Volume 146,2024,105014,ISSN 0262-8856,

[11]T. Mohandoss, J. Rangaraj,Multi-Object Detection using Enhanced YOLOv2 and LuNet Algorithms in Surveillance Videos, e-Prime - Advances in Electrical Engineering, Electronics and Energy, Volume 8, 2024, 100535, ISSN 2772-6711

[12]Pulkit Dwivedi, Soumendu Chakraborty, Single image dehazing using extended local dark channel prior, Image and Vision Computing, Volume 136,2023,104747, ISSN 0262-8856

[13] 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