Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.058903

ARTICLE

Context-Aware Feature Extraction Network for High-Precision UAV-Based Vehicle Detection in Urban Environments

Yahia Said[1,*], Yahya Alassaf[2], Taoufik Saidani[3], Refka Ghodhbani[3], Olfa Ben Rhaiem[4] and Ali Ahmad Alalawi[1]

[1]Department of Electrical Engineering, College of Engineering, Northern Border University, Arar, 91431, Saudi Arabia

[2]Department of Civil Engineering, College of Engineering, Northern Border University, Arar, 91431, Saudi Arabia

[3]Faculty of Computing and Information Technology, Northern Border University, Rafha, 91911, Saudi Arabia

[4]College of Science, Northern Border University, Arar, 91431, Saudi Arabia

*Corresponding Author: Yahia Said. Email: yahia.said@nbu.edu.sa

Received: 23 September 2024 Accepted: 07 November 2024 Published: 19 December 2024

ABSTRACT

The integration of Unmanned Aerial Vehicles (UAVs) into Intelligent Transportation Systems (ITS) holds transformative potential for real-time traffic monitoring, a critical component of emerging smart city infrastructure. UAVs offer unique advantages over stationary traffic cameras, including greater flexibility in monitoring large and dynamic urban areas. However, detecting small, densely packed vehicles in UAV imagery remains a significant challenge due to occlusion, variations in lighting, and the complexity of urban landscapes. Conventional models often struggle with these issues, leading to inaccurate detections and reduced performance in practical applications. To address these challenges, this paper introduces CFEMNet, an advanced deep learning model specifically designed for high-precision vehicle detection in complex urban environments. CFEMNet is built on the High-Resolution Network (HRNet) architecture and integrates a Context-aware Feature Extraction Module (CFEM), which combines multi-scale feature learning with a novel Self-Attention and Convolution layer setup within a Multi-scale Feature Block (MFB). This combination allows CFEMNet to accurately capture fine-grained details across varying scales, crucial for detecting small or partially occluded vehicles. Furthermore, the model incorporates an Equivalent Feed-Forward Network (EFFN) Block to ensure robust extraction of both spatial and semantic features, enhancing its ability to distinguish vehicles from similar objects. To optimize computational efficiency, CFEMNet employs a local window adaptation of Multi-head Self-Attention (MSA), which reduces memory overhead without sacrificing detection accuracy. Extensive experimental evaluations on the UAVDT and VisDrone-DET2018 datasets confirm CFEMNet's superior performance in vehicle detection compared to existing models. This new architecture establishes CFEMNet as a benchmark for UAV-enabled traffic management, offering enhanced precision, reduced computational demands, and scalability for deployment in smart city applications. The advancements presented in CFEMNet contribute significantly to the evolution of smart city technologies, providing a foundation for intelligent and responsive traffic management systems that can adapt to the dynamic demands of urban environments.

KEYWORDS

Smart cities; UAVs; vehicle detection; traffic management; intelligent transportation systems; anchor-free detection; high-resolution network; context-aware feature extraction; multi-head self-attention

1 Introduction

The modern urbanization process causes increased travel needs, which have an impact on the economy and people's lives in cities causing numerous road accidents and significant traffic congestion [1]. Detecting heterogeneous traffic flow helps to decrease traffic congestion and accidents on urban highways. Information collection is highly required to support ITS for traffic management. Unfortunately, most security cameras are set in permanent locations and have detection limits such as ranges and blind spots. As a result, deploying security cameras to detect traffic objects in a timely and effective manner is difficult.

The mature technology of drones or Unmanned Aerial Vehicles (UAVs) has accelerated their use in ITS. UAV-based monitoring systems provide benefits over typically fixed security cameras, such as large field of vision. Low cost, simplicity of deployment, portability, and inexpensive maintenance are all advantages. Furthermore, when UAVs identify objects on various road sections, they may efficiently avoid obstacles without interfering with road traffic. Besides, the incorporation of intelligent image processing techniques into UAVs can identify the position of objects precisely, minimize maintenance and installation labor costs, and increase the intelligence of traffic surveillance [2].

The use of Unmanned Aerial Vehicles (UAVs) for real-time traffic monitoring in urban environments has grown significantly in recent years. However, current vehicle detection methods face numerous challenges due to the complex nature of urban environments, where vehicles are often small, densely packed, and partially occluded. In urban traffic scenarios, UAV imagery frequently captures small vehicles in densely packed configurations, making it difficult for traditional detection models to distinguish between closely situated vehicles. These models often struggle with occlusion caused by buildings or other vehicles and the size variability due to the altitude of UAVs and varying perspectives in urban landscapes. Most existing models rely heavily on standard convolutional architectures or Feature Pyramid Networks (FPNs). While these methods perform well in standard detection tasks, they fail to capture the fine-grained details necessary for small object detection, especially in crowded scenes.

Another critical challenge is the inability of existing models to capture adequate contextual information for accurate vehicle detection. In dense urban environments, vehicles are often partially obscured by surrounding infrastructure or other objects. Traditional detection methods that focus on local spatial features struggle to differentiate between vehicles and their complex backgrounds. Many existing models lack mechanisms for incorporating global contextual information, leading to poor detection performance in cluttered urban environments. While self-attention mechanisms have shown promise in improving feature extraction by capturing long-range dependencies, they come with high computational costs. This is particularly problematic in UAV-based systems where resources are often limited, making traditional multi-head self-attention mechanisms impractical for real-time vehicle detection. The computational burden arises from applying attention globally, which increases memory usage and limits the feasibility of such models in resource-constrained environments like UAVs. For effective vehicle detection, robust extraction of both spatial and semantic features is essential. However, many traditional models fail to balance these two aspects, resulting in diminished accuracy when vehicles are partially occluded or vary in scale. Conventional architectures do not offer efficient mechanisms to simultaneously capture both spatial and semantic features, limiting their ability to perform well in complex urban scenes.

In comparison to wall cameras, vehicles (e.g., cars, vans, bicycles, etc.) in UAV images span significantly smaller high-density pixel areas. As a result, using UAVs to detect vehicles presents two main challenges:

- How to rapidly process and interpret images obtained from UAVs.
- How to reliably recognize vehicles of high density and tiny size. As a result, the precision and speed of the detection algorithm are crucial for traffic monitoring.

Traditional hand-crafted features-based algorithms depend mostly on the knowledge of key feature points [3] and objects' edges. The sliding window technique was deployed to establish the position of the item by searching the entire image. When utilizing UAVs for traffic surveillance, these algorithms are incapable of recognizing tiny and dense objects. Recent advances in GPU (Graphics Processing Unit) processing capability, the introduction of big data [4], and deep learning [5] have accelerated building powerful object detection methods.

Object detection algorithms based on deep learning may be classified into two types: two-stage detection based on the regional proposal [6] and single-stage detection based on regression [7]. Although the two-stage detection algorithms have good detection accuracy, their detection speed is limited, restricting real-time applicability in UAVs processing systems. On the other hand, the single-stage detection techniques are extremely fast. Yet, these algorithms perform badly when it comes to detecting tiny objects in images captured by UAVs.

In recent years, one-stage object detection models have been divided into two categories: anchor-based [8,9] and anchor-free [10]. Faster-RCNN (Fast Region-based Convolutional Network) represents anchor-based detectors that work by regressing bounding boxes and categorizing classes. As a result, designing the anchor box's hyper-parameters such as aspect ratio, size, and the number of anchors influence significantly the detection performance. Anchor-free detection models, as opposed to anchor-based detectors, identify objects directly from input images without processing a vast number of proposed regions. Furthermore, by eliminating the manual design in terms of anchor hyper-parameters, anchor-free detectors present strong generalization power on varied detection tasks. Either anchor-based or anchor-free model, the majority of current detectors learn object features using CNN (Convolutional Neural Networks)-based feature extraction networks. Nevertheless, unlike ordinary object recognition tasks, vehicle detection in aerial images involves more small and occluded samples. The capacity of the model to learn from hard data results in a substantial influence on detection performance.

There are primarily two prominent strategies for improving the networks' capacity to learn all required features. The first is to create a high-resolution heatmap for precise local discriminating of occluded and small objects. The second is to extract accurate semantic information for larger visual interpretation, which improves prediction performance. Yet, because lower resolution frequently provides higher semantic information, it is challenging to strike a fair balance between rich semantic information and high resolution.

Since rich semantic information and high-resolution representations can be learned while well-assuring information transfer between feature maps with high- and low-resolution, the newly suggested high-resolution network (HRNet) [11] demonstrated comparable performance for object detection. Unfortunately, it was discovered that HRNet struggles in learning tough situations, particularly for severely occluded and small-scale objects. Spatial context has been involved in recent research (represents the position of an existent item [12]) in object detection; nonetheless, the effectiveness of this aspect in object re-identification [11] stimulates its investigation in vehicle detection. Reference [13] asserts that connections between an object in the image provide reliable spatial representations. That was essential for resolving HRNet's noisy and meaningless spatial representations. The basic MSA operation demonstrates a remarkable capacity for acquiring exact spatial connections, owing to the significant success of Transformer Networks in Computer Vision [14]. As a result, the

MSA was deployed to vehicle detection in order to extract spatial context. Referring to the theory mentioned above, we argue that extracting semantic and spatial context results in more discriminative representation learning. So, it is possible to effectively overcome the limitations of HRNet for small and occluded vehicles. The fundamental obstacles to detecting small objects in aerial images captured using UAV are the low resolution, which conveys limited information, and the slow debasement of critical characteristics that define small objects throughout the convolutional neural network's down-sampling process.

To address the challenges of small and densely packed vehicle detection, CFEMNet introduces the Context-aware Feature Extraction Module (CFEM). This module fuses multi-scale feature learning early in the network, allowing the model to dynamically combine spatial and semantic information during feature extraction.

To handle the main problems in detecting vehicles using UAVs, basic but effective CFEM was presented. CFEM improves the model's ability to detect small, occluded vehicles by leveraging context from the surrounding environment, making it more effective in crowded urban settings. Unlike traditional multi-scale learning methods that rely solely on convolutional layers, CFEMNet incorporates a novel combination of Self-attention and Convolution layers within the Multi-scale Feature Block (MFB). This fusion allows the network to capture both local and global context, ensuring comprehensive feature extraction. This method enhances the model's ability to recognize vehicles in complex environments by integrating both local spatial features (via convolution) and global relationships (via self-attention), leading to more accurate detection.

To overcome the computational inefficiencies of traditional attention mechanisms, CFEMNet employs a local window adaptation of Multi-head Self-Attention (MSA). This technique applies self-attention within localized regions, significantly reducing memory and computational overhead while maintaining high detection accuracy. This adaptation makes the model efficient and scalable, enabling real-time UAV-based traffic monitoring without compromising on detection performance.

The Equivalent Feed-Forward Network (EFFN) Block is introduced to ensure robust extraction of both spatial and semantic features. By balancing these two types of features, the EFFN Block enhances the model's capacity to detect vehicles that vary in scale and appearance, even in complex and occluded environments.

The EFFN Block ensures accurate vehicle detection across different scales and occlusion levels, providing superior performance in challenging urban settings.

The main objective was to deploy convolution operation for extracting deep semantic context and MSA operation to extract spatial context. The newly suggested CFEM comprises an MFB followed by EFFN Block that are applied after the fourth stage of HRNet feature pyramid. Based on the "Shift" theory [15] and certain studies [16], a parallel mechanism was proposed to combine MSA and convolution. The MFB was proposed to integrate convolution and MSA with low computing costs. To be more specific, the input feature maps were passed through a 1×1 convolution to retrieve features and passed to the MSA path and the convolution path. As a result, we may extract as many common parameters as feasible from the parallel routes, lowering the number of parameters significantly. Then, two parallel pathways were created based on their distinct paradigms in order to learn context information from HRNet more effectively. To achieve the deep semantic context, group convolution was adopted for the convolution path. To get an accurate spatial context, we built an MSA based on position encoding with relative-distance-aware, as proposed in [17]. Besides, local window MSA was proposed to address the memory consumption disadvantage in high-resolution feature maps. The proposed design greatly reduced the required memory while barely affecting precision. After

that, robust semantic representation was obtained and offered shared features in the MFB output feature maps via the EFFN Block. CFEM resizes the modified multi-scale feature maps to the same resolution and concatenates them as the detection head's input. Lastly, like in traditional anchor-free detectors, vehicle identification is structured in the detection head by the center point and passing the concatenated feature maps through convolution for scale prediction task. The proposed CFEMNet is formed by mounting the proposed detection head and the CFEM on the HRNet. So, the limitation of HRNet's was elevated by improving the ability of feature extraction for small and occluded vehicle instances.

The following points summarize the contributions of the proposed work:

- To address HRNet's low detection performance for substantially small and occluded vehicles in aerial images, we proposed a novel CFEM that uses a parallel architecture to extract spatial context and semantic context via convolution and MSA.
- Presenting a local window MSA to considerably minimize the memory requirement of the MSA path in CFEM. We also studied the effect of aspect ratio windows on the vehicle shape representation abilities.
- We explored the influence of the two MFB on detection performance.
- The performance of the proposed CFEMNet was demonstrated by extensive experiments on two difficult benchmarks (UAVDT and Visdrone-DET2018 datasets).

The remainder of this paper is organized as follows: Section 2 is allocated for the presentation of the literature review. The proposed approach is presented and detailed in Section 3. Experiments and results are presented and discussed in Section 4. In Section 5, conclusions and future works are provided.

2 Related Works

Small object detection in aerial images is a difficult but critical challenge in traffic management. The fundamental difficulty related to small objects is low resolution and poor information. The capacity of the convolutional neural network to convey features will eventually degrade as it is down-sampled. Furthermore, increasing the depth of the convolutional layer results in decreasing the fine-grained information of the object in feature maps but the semantic information increases. Additional research reveals that object localization shallow feature maps with fine-grained information in is more, whereas in object classification [18] depends more on deep feature maps with rich semantic information. As a result, in a deep neural network, learning and fusing multi-scale data is critical to solving tiny object detection [19].

Scale Normalization for Image Pyramids (SNIP) [20] efficiently exploited all of the training data through training the model at each side of the pyramid. FPN [21] built a lateral connection that combined deep features with semantic information and shallow features with spatial information. Path Aggregation Network (PANet) [22], which is based on FPN, provided increased the bottom-up path to reduced features transfer path while employing the shallow-level features for the exact positioning information to enhance small object recognition performance.

Liu et al. [23] developed Multi-branch Parallel Feature Pyramid Networks (MPFPN) to regain rejected deeper layers' features, allowing for more plentiful feature extraction from small-size objects. To forecast and fine-tune the position and size of small items, Zhang et al. [24] proposed multi-scale characteristics using the interleaved cascade architecture. In addition, unlike the previously described network structures designed manually, the Neural Architecture Search (NAS) was used to search for

and design the structure of FPNs automatically, as in Neural Architecture Search Feature Pyramid Network (NAS-FPN) [25] and Automatic Feature Pyramid Network (Auto-FPN) [26].

In the context of vehicle detection using UAVs, many works have been proposed [27] to achieve the desired performance. Kainz et al. [28] proposed a vehicle detection and tracking method using UAV for a web-based traffic monitoring system. The proposed method starts by defining the region of interest by the user. Then a motion detection sensor is deployed to detect moving vehicles. After that, the YOLOv4 model [29] was used for detecting vehicles and other traffic objects. Finally, detected vehicles are tracked and counted. The tracking algorithm was based on the Euclidian distance of the detection box in two consecutive frames. The proposed method was evaluated on a custom scenario and evaluated based on the detection rate in a given region of interest.

A vehicle detection system was proposed in [30]. The main idea was to transform the input image from the red-green-blue (RGB) space color to the Hue-Saturation-Value (HSV) space color. This operation was deployed to improve sample diversity and to adapt lighting conditions adaptability. Then, the Single Shot Multi-Box Detector (SSD) model [31] was improved for vehicle detection. A custom loss function was proposed to enhance the feature extraction process. The evaluation of the proposed system on real videos collected using UAV shows reliable performance compared to old machine learning algorithms.

Makrigiorgis et al. [32] proposed a vehicle detection method in aerial images provided by UAV for use in traffic monitoring systems. The proposed method combined road segmentation and vehicle detection to enhance the overall performance by increasing the focus on relevant vehicles. YOLOv3 [33] and YOLOv4 [29] were adopted for the detection task. For the segmentation task, many models have been investigated such as ERFNet [34], PAN [35], and DeepLabv3 [36]. The evaluation of the mentioned models for different tasks showed encouraging results that prove their potential use for real traffic monitoring systems.

Bakirci et al. [37] investigated vehicle detection using the YOLOv8 algorithm. The emphasis was targeting aerial images taken by a customized autonomous UAV, which provides a novel opportunity to put this algorithm to use in the real world. The dataset used to train and test the algorithm was collected from a varied group of traffic imagery taken by UAVs. The research methodically used a specially built and programmed drone to change fly paths, altitudes, orientations, and camera angles in an attempt to increase the variety of vehicle images. Intentionally targeting the algorithm's adaptability across many contexts, this technique is intended to improve its generalization skills. A thorough comparison analysis was carried out to assess the algorithm's performance, with an emphasis on the YOLOv8n and YOLOv8x sub-models from the YOLOv8 series. The sub-models were tested extensively using the proposed dataset in a variety of lighting and environmental scenarios.

After a deep investigation of the literature review, it was discovered that almost all works focus on deploying general object detection models for vehicle detection in aerial images provided by UAVs. However, this led to degraded performance due to low detection capacities for small objects. To solve the problem of detecting small objects in aerial images, we proposed the integration of MSA in a convolutional neural network model through a CFEM. The proposed design enhanced the learning capability and the detection ability of small objects.

3 Proposed Approach

3.1 HRNet Background

HRNet [11] was initially designed for human pose estimation. Fig. 1 presents an illustration of the HRNet. It was designed to keep high-resolution feature maps throughout the feature extraction process. This was done by connecting high-resolution to low-resolution feature maps in a parallel manner. Besides, features are repeatedly exchanged at different resolutions. It was demonstrated that HRNet can do well on a variety of different visual tasks. HRNet recently demonstrated promising performance since it excels at both low-level spatial and high-level semantic representations. Nonetheless, HRNet struggles in localizing a high number of small and occluded objects. To summarize, it was discovered that HRNet presents numerous disadvantages for small and occluded objects:

- HRNet is not sufficiently deep, limiting the ability to extract semantic features.
- During the information transmission process, misaligned and meaningless spatial features are created, which are harmful for accurate detection of small and occluded objects.

To overcome the mentioned limitation, several improvements were proposed including the devilment of an anchor-free framework and an attention module to enhance the focus of the network on specific objects which are vehicles in this work.

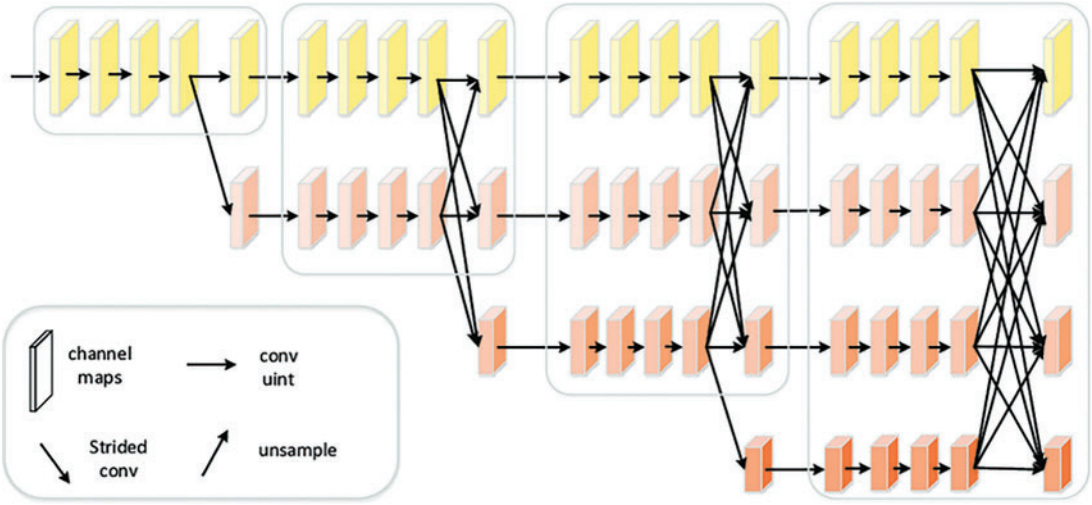

Figure 1: Architecture of the HRNet

3.2 Proposed CFEMNet

To address HRNet's inadequacies, we present CFEMNet, a basic but effective anchor-free network that makes detectors more resilient to small and occluded vehicles by learning context features more effectively. Fig. 2 illustrates the general architecture of the CFEMNet. CFEMNet will be described in depth by detailing the different components: CFEM, Feature Fusion, Detection Head, and Optimization.

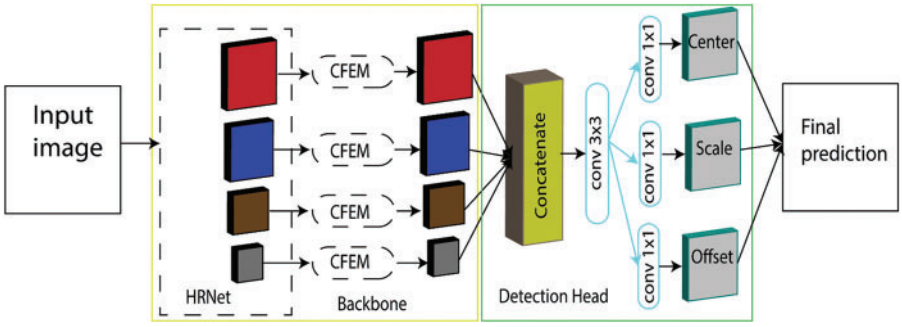

Figure 2: The proposed CFEMNet

3.3 Context-Aware Feature Extraction Module

The main goal of the proposed CFEM is to learn the accurate spatial context as well as the deep semantic context from HRNet. Two major components made the proposed CFEM: a new MFB that integrates the MSA path and the convolution path with a parallel design and low computational cost, and an EFFN Block developed to obtain a relevant semantic information and offer cross-channel information communications for the MFB-output Block's feature maps. Fig. 3 depicts the architecture of the CFEM.

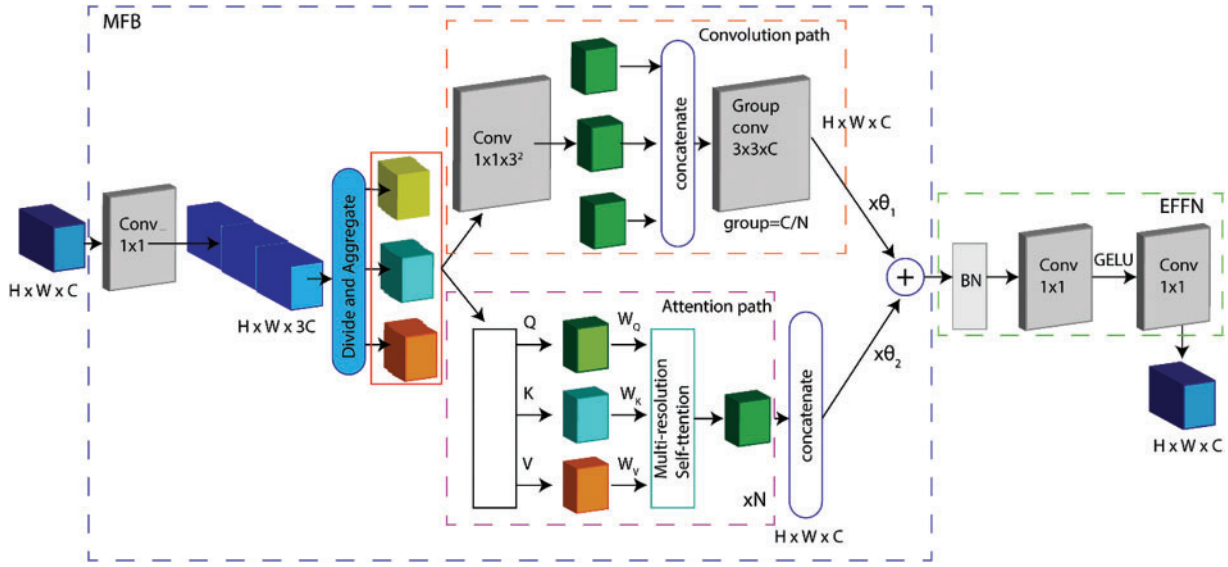

Figure 3: The proposed CFEM. Conv: convolution, BN: batch normalization, and Group conv: group convolution

A novel CFEM was proposed to merge the MSA path and the convolution path in a parallel architecture with minimal processing cost in order to acquire accurate spatial context and deep semantic context. There is a variety of existing works such as [38] that present the use of parallel path combinations between convolution and MSA. However, the proposed CFEM presents many benefits compared to [38]. First, in the convolution path, channels were deepened to extract the strong semantic context to produce the deep semantic context, whereas [38] map channels using identity instead of

deepening the channel. Second, the awareness of relative distance was adopted for coding the position to acquire the precise spatial context, which was not used in [38]. Third, different from downstream vision tasks discussed in [38], vehicle detection in aerial images is heavily reliant on features with high resolution. To address the issue of high-resolution feature memory consumption, we offer a unique local window MSA that conforms to vehicle feature distribution quirks, which may considerably minimize memory consumption while scarcely affecting detection performance. As a result, the CFEM coupled with HRNet is designed carefully for better understanding the context of small and occluded vehicles in aerial images.

The workflow of the CFEM is divided into two stages. In the first stage, 1 × 1 convolution was applied in order to get deep semantic information by extending the input feature maps to 3C feature maps. Then, N pieces were obtained by dividing the 3C feature maps and a rich collection of features were aggregated including (H × W × C/N) × 3C feature maps forwarded to the next stage with parallel structure. In the second stage, Convolution and MSA follow two parallel pathways. Intermediate features were passed to both paths then restructure them to satisfy the required process. A 1 × 1 convolution layer is utilized to transform the 3N intermediate features to 32 projected feature maps for the convolution path with kernel size 3. Using the shift and addition procedures introduced in [15], the next feature Convolutional maps may be processed in the same manner that regular convolution do, and local receptive field generates relevant information. Unfortunately, the aforementioned procedures are difficult to vectorize, which significantly reduces actual efficiency. Wasteful shift and addition operations were avoided by using the group convolution with learnable kernel weights described in [38] was applied. Particularly, generated feature maps were concatenated to generate the 9 × C/N channel feature map. Kernel weights given in [38] were carefully selected to initialize the group convolution layers. Lastly, the previously prepared group convolution extracts deep semantic context and generates the output feature map with C channel. We split the intermediate features for the MSA route into N groups of three features each. Then we consider the three feature components as queries Q $\in \mathbb{R}^{H\times W\times \frac{C}{N}}$, keys k $\in \mathbb{R}^{H\times W\times \frac{C}{N}}$, and values V $\in \mathbb{R}^{H\times W\times \frac{C}{N}}$. Also, the number of heads was the same as the number of groups N. According to the newly built Multi-Resolution Self-Attention Module shown in Fig. 4, feature maps with C/N channels of the MSA were collected and to generate C channels output feature map to retrieve the accurate spatial context by concatenating feature maps. Finally, feature maps generated from the convolution path and the MSA path are combined based on two dynamic weights θ_1 and θ_2. Those weights are adjustable and learnable through the backpropagation process. The output features maps of the CFEM can be computed as (1).

$$F_{out} = \theta_1 F_{conv} + \theta_2 F_{MSA} \tag{1}$$

The dynamic weights θ_1 and θ_2 measures the bias of different features toward the MSA or the convolution path. We can retrieve the deep semantic and exact spatial context of HRNet output features using MFB, which can compensate for HRNet's inability to extract obstructed and small-scale vehicle features.

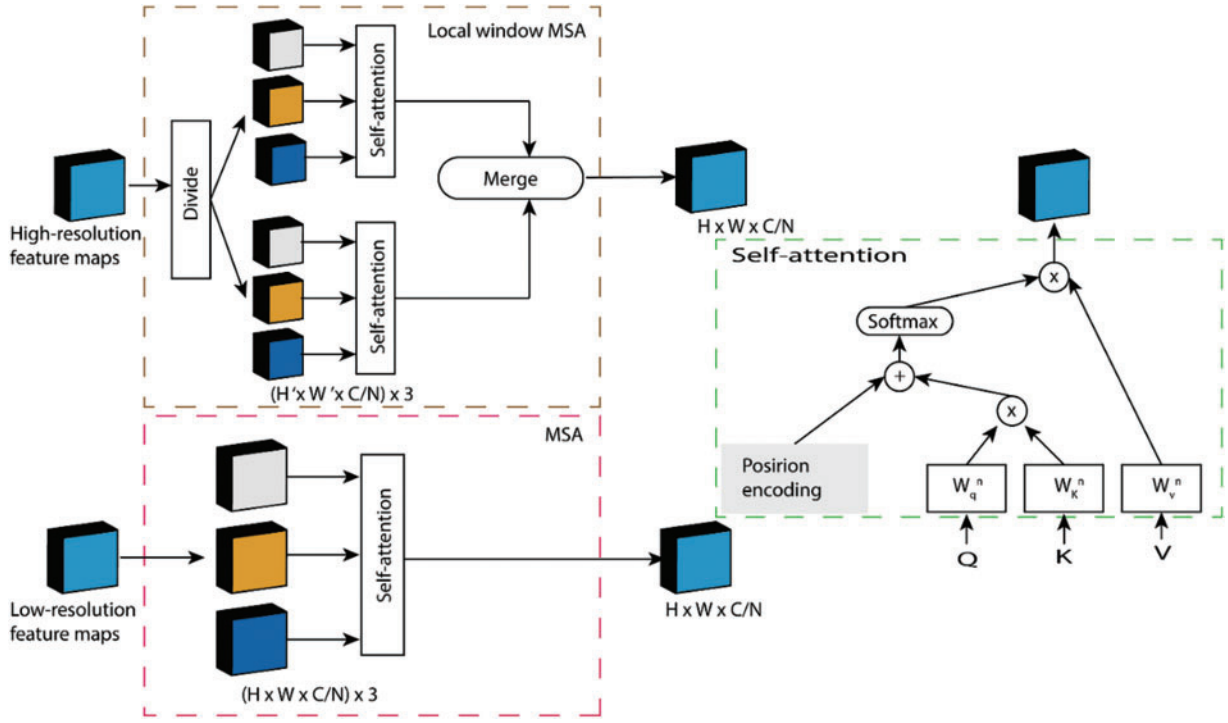

Figure 4: The proposed MSA and local window MSA

3.4 Multi-Head Self-Attention

An MSA with robust context capture capabilities and a global receptive field was adopted for efficient acquisition of context embedded in low-resolution feature maps at stage 4 of HRNet-32 with downsampling rates of 16 and 32. As shown in Fig. 4, N heads with Q, K, and V inputs are required for MSA. The final output can be computed as (2).

$$MAS\,(Q,K,V) = concat\ (head_1, head_2, \ldots, head_N)$$

$$head_n = softmax\left(\frac{\left(QW_q^n\right)\left(KW_k^n\right)}{\sqrt{\frac{C}{N}}} + B\right)\left(VW_v^n\right) \tag{2}$$

W_q^n, W_k^n, and $W_v^n \in \mathbb{R}^{\frac{C}{N}\times C}$ are the projection metrics for the queries, keys, and values, respectively. B is location encoding with relative distance, which is capable of associating positional awareness across objects.

3.5 Proposed Local Window MSA

Despite having robust context capture capacity and a wider receptive field, the high-resolution vehicle recognition operation consumes extensive of memory. As a result, for HRNet 32 stage 4 high-resolution feature maps, we suggest replacing the MSA by a local window MSA. Subsequently, it has the potential to dramatically minimize memory consumption while demonstrating negligible loss in the performance. The local window MSA size was created with a certain aspect ratio of 1:2, this

configuration resulted in minimal memory consumption and produced a good performance. In effect, the inputs $Q \in \mathbb{R}^{H\times W\times \frac{C}{N}}$, keys $K \in \mathbb{R}^{H\times W\times \frac{C}{N}}$, and values $V \in \mathbb{R}^{H\times W\times \frac{C}{N}}$ were divided into non-overlapping windows. The local windows can be defined as (3).

$$\begin{aligned}
&\{Q \to \{Q_1, Q_2, \ldots, Q_p\} \in \mathbb{R}^{H'\times W'\times \frac{C}{N}} \\
&K \to \{K_1, K_2, \ldots, K_p\} \in \mathbb{R}^{H'\times W'\times \frac{C}{N}} \\
&V \to \{V_1, V_2, \ldots, V_p\} \in \mathbb{R}^{H'\times W'\times \frac{C}{N}}\}
\end{aligned} \tag{3}$$

where H' is the height of the local window and W' is its width. Unlike existing works that proposed local windows with fixed sizes, the proposed local window was selected based on an aspect ratio of 1:2. The suggested local window is suited for vehicle aspect ratio abnormalities, which facilitates the extraction of vehicle features. After applying the MSA on each window separately, the output is merged to produce the final feature maps. Fig. 4 presents the difference between the MSA and the local window MSA.

3.6 Equivalent Feed-Forward Network

Since MSA executes self-attention in non-overlapping channels, the communication between channels is ignored. Aiming to improve the depth in order to acquire robust semantic representation and share features across channels, a 1×1 convolution layer was applied. An EFFN Block was developed based on the feed-forward network provided in the Transformer network by combining two 1×1 convolutions, a GELU (Gaussian Error Linear Unit) activation function, batch normalization, and a residual connection. Basically, The EFFN Block has four times the number of channels as the input channels. The proposed EFFN was connected to the output of the MFB and its output is the final output of the CFEM.

3.7 Features Fusion

While combining deeper feature maps improves detection performance, The CFEM context features were scaled to the same resolution by passing them via a deconvolution layer. Following that, concatenated scaled feature maps are provided to the detection head. Possible complex feature fusion techniques were performed to gain more performance, but this will negatively affect the computation complexity.

3.8 Detection Head

The resultant concatenated feature maps are fed into the detection head to generate the prediction results. Initially, a 3×3 convolutional layer is added to lower the number of channels to 256 in the feature maps. Then, 1×1 convolution layers are used to construct the center heatmap, scale map, and offset map. Following that, vehicle bounding boxes may be generated automatically using the corresponding scales of the center heatmap in the scale map. Lastly, the offset prediction branch improves detection performance by gently changing vehicle center locations.

To train the proposed detection head, ground truth labels are required. Based on existing bounding boxes annotation, the center and the offset of each vehicle were generated. To generate the center points, the center of each bounding box is selected. As it is impossible to find the exact center point of a vehicle, a 2D Gaussian distribution P was centered by the center point of the labeling bounding box. The center point can be computed as (4).

$$M_{ij} = max_{n=1,2,\ldots,k} P(i, j; x_n, y_n, \sigma_{wn}, \sigma_{hn}) \tag{4}$$

$$P(i,j;x_n,y_n,\sigma_{wn},\sigma_{hn}) = e^{-(\frac{(i-x)^2}{2\sigma_w^2}+\frac{(j-y)}{2\sigma_h^2})}$$

where k represents the number of detected vehicles; (i,j) is the position of the center; (x_n, y_n, h_n, w_n) are the coordinates of the center point, the width, and height of the nth vehicle; σ_w, σ_h are the variance of the 2D Gaussian distribution related to the width and height of the vehicle respectively.

The height and/or width of the vehicle can be used to establish the scale ground truth. Line annotation was proposed in [39], bounding box with a consistent aspect ratio may be produced. Our network estimates the height of each vehicle based on the line annotation and constructs the desired aspect ratio bounding box with the. The ground truth scale is computed as (5).

$$S_{ij} = \log(h_n) \tag{5}$$

To minimize uncertainty, a radius of 2 negatives of vehicle centers is additionally allocated $\log(h_n)$.

The center position was corrected before remapping by appending offset branch following [40]. The offset ground truth is defined as (6).

$$O_{ij} = \left(\frac{x_n}{r} - \left\lfloor \frac{x_n}{r} \right\rfloor, \frac{y_n}{r} - \left\lfloor \frac{y_n}{r} \right\rfloor\right) \tag{6}$$

3.9 Optimization and Loss

To enhance the prediction capability, the imbalance between negative and positive samples was solved. For the classification task, the Focal Loss [40] was used to train the detecting head's center position prediction. The loss function for center prediction is defined as (7).

$$\begin{aligned} &\{L_{center} = -\frac{1}{N}\sum_{i=1}^{\frac{w}{r}}\sum_{j=1}^{\frac{h}{r}} \alpha_{ij}(1-\hat{p}_{ij})^{\gamma}\log(\hat{p}_{ij}) \\ &\hat{p}_{ij} = \begin{cases} p_{ij}\ if\ y_{ij}=1 \\ 1-p_{ij}\ otherwise \end{cases} \\ &\alpha_{ij} = \begin{cases} 1\ if\ y_{ij}=1 \\ (1-M_{ij})^{\beta} otherwise \end{cases}\} \end{aligned} \tag{7}$$

where p_{ij} is the prediction probability of position (i,j) if it is a center for a vehicle or not. $y_{ij}=1$ refers to the positive prediction of the vehicle's center at position (i,j). The gaussian mask M_{ij} was used in the training process to minimize the impact of negative samples. β and γ are the hyperparameters that measure the focus on negative and positive samples. Extensive experiments proved that $\gamma = 2$ and $\beta = 4$ achieved the best performances. The scale and offset were solved as regression problems based on the smooth L1 loss. The loss function for the scale and the offset are defined as (8) and (9).

$$L_{scale} = -\frac{1}{N}\sum_{n=1}^{N} smooth\,L1(s_n - \bar{s}_n) \tag{8}$$

$$L_{offset} = -\frac{1}{N}\sum_{n=1}^{N} smooth\,L1(o_n - \bar{o}_n) \tag{9}$$

where s_n and o_n are nth the predicted scale and offset and $\bar{s}_n$ and $\bar{o}_n$ are their corresponding ground truth, respectively. The final loss function for the proposed network is the weighted sum of the different loss

functions. The loss function can be computed as (10).

$$L = \rho_c L_{center} + \rho_s L_{scale} + \rho_o L_{offset} \tag{10}$$

where ρ_c, ρ_s, and ρ_o are the weights for the center loss, the scale loss, and the offset loss, respectively. Experiments proved that $\rho_c = 0.01$, $\rho_s = 1$, and $\rho_o = 0.1$ are the best value for this work.

4 Experiments and Results

4.1 Experimental Environment and Evaluation Metrics

To demonstrate the robustness of the proposed model, it was evaluated using two challenging datasets. The VisDrone-VDT2018 dataset contains totaling 33,366 frames generated from 79 video sequences. The dataset provides three non-overlapping subsets: training set (56 video clips totaling 24,198 frames), validation set (7 video clips totaling 2846 frames), and testing set (16 video clips with 6322 frames). The video sequences were shot in numerous cities under diverse weather and lighting circumstances. The dataset was manually annotated which results in an average of 149,9k instances for each category. Five main classes were provided which are pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. In this work, we only considered the vehicle classes including car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle.

The UAVDT benchmark provides 100 video sequences chosen from over 10 h of footage shot with a UAV platform at various urban locales, representing many frequent situations such as squares, arterial streets, toll stations, highways, crossings, and T-junctions. The videos are captured at 30 frames per second (fps) with a resolution of 1080 × 540 pixels and JPEG format. In the UAVDT benchmark dataset, over 2700 vehicles were annotated in about 80,000 frames with 0.84 million bounding boxes. Three classes (car, bus, and truck) referring to vehicles were considered.

Experiments were conducted on a desktop with the Ubuntu operating system featured with Intel i7 CPU and Nvidia GTX 960 GPU. The Pytorch deep learning framework was used to develop the proposed model with CUDA (Compute Unified Device Architecture) acceleration and cuDNN (NVIDIA CUDA® Deep Neural Network library) support. OpenCv library was used for data manipulation.

To ensure better performance, the backbone (HRNet-W32) was first pre-trained on the ImageNet dataset then the pre-trained weights were transferred to the proposed model. The AdamW was adopted for loss optimization. The size of the proposed local window in MSA was set to 20 × 40. Many data augmentation techniques were applied to increase the generalization power of the model including Mosaic data, brightness variation, horizontal flip, cropping, and random scaling. The batch size was fixed to 4 due to the limited GPU memory. The initial learning rate was 0.002 and the weight decay was 0.0001. The proposed model was trained for 30 epochs each epoch has a number of iterations equal to the number of frames divided by 30. Table 1 presents a summary of the hyperparameter configuration.

Table 1: Hyperparameters configuration for training the proposed CFEMNet

Hyperparameter	Configuration
Learning rate	0.002
Batch size	4
Weight decay	0.0001
Epochs	30

(Continued)

Table 1 (continued)

Hyperparameter	Configuration
Learning rate scheduler	ReduceLROnPlateau (patience = 5, factor = 0.5)
Optimizer	AdamW ($\beta1 = 0.9$, $\beta2 = 0.999$, $\varepsilon = 1e{-}8$)
Cross-validation	K-fold (K = 5)
Early stopping	Enabled (patience = 5)

Evaluating an object detection model requires precise evaluation metrics. Usually, two evaluation metrics are used which are precision and recall. The precision indicates the performance of the proposed model in predicting correct detections. The recall provides information on how many positive detections were correctly predicted. The precision (P) and recall (R) can be computed as (11) and (12). The F1 score can be obtained by mixing the precision and recall to provide information on model quality. The F1 score can be computed as (13).

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = 2 \times \frac{R \times P}{P + R} \tag{13}$$

where TP is true positive samples that are correctly detected, FN is false positive samples that are not detected, and FP is false positive samples that are incorrectly detected.

Besides, the mean average precision (mAP) is a widely used metric for evaluating the performance of object detection models. As its name mentions, it is based on the average precision (AP) of N classes across the complete dataset. To consider a prediction as positive, the IoU (Intersection over Union) between the predicted box and the ground truth box must not exceed a given threshold. In this work, the IoU threshold was set to 0.5. The mAP can be computed as (14).

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{14}$$

4.2 Results and Comparison

The detection efficacy of the proposed CFEMNet was validated through the achieved results. Despite dense objects or small cars on the road, the suggested restructuring allows the CFEM-Net to recognize and locate traffic-moveable items in images taken by UAVs.

In comparison to HRNet, the suggested CFEMNet method performs better at recognizing tiny objects in UAVs. As compared to the findings of the CFEMNet model, HRNet has numerous missed detections and cannot detect densely dispersed tiny objects in aerial images captured by a UAV. In contrast, the proposed CFEMNet can detect densely dispersed tiny objects with more accuracy than HRNet, resulting in much better confidence scores. Table 2 presents the achieved average precision (AP) for each class and the mAP of the proposed model compared to the HRNet model on the Visdrone-DET2018 dataset. Based on the archived results, the proposed CFEMNet outperforms the

HRNet in terms of mAP and presents better detection of small objects such as bicycles and tricycles. For an input size of 608 × 608, the CFEMNet achieved an improvement of more than 3% and for an input size of 800 × 800, an improvement of more than 4% was achieved.

Table 2: Achieved mAP on the Visdrone-DET2018 dataset

Model	Input size	AP (%)								mAP (%)
		Car	Van	Bus	Truck	Motor	Bicycle	Awning-tricycle	Tricycle	
HRNet	608	69.4	38.8	49.2	31.3	30.1	9.6	12.2	20.4	32.6
	800	79.8	47.3	60.5	37.4	44.1	14.1	15.7	28.3	39.9
CFEMNet	608	73.7	41.7	52.6	32.8	35.4	11.2	14.6	25.5	35.9
	800	82.6	51.1	62.4	39.9	48.8	19.0	20.3	33.8	44.7

Table 3 presents the achieved F1 score of the proposed model compared to the HRNet. Referring to the results presented in Table 3, the proposed CFEMNet presents an improvement of more than 5% for different input sizes. The achieved F1 scores proved the good quality of the proposed model compared to the HRNet model.

Table 3: Achieved F1 score on the Visdrone-DET2018 dataset

Model	Input size	F1 score (%)								
		Car	Van	Bus	Truck	Motor	Bicycle	Awning-tricycle	Tricycle	Average
HRNet	608	64.1	39.9	51.2	36.4	39.6	16.2	23.7	30.6	37.4
	800	69.8	46.5	59.5	39.8	48.1	23.5	25.9	38.3	43.9
CFEMNet	608	65.0	41.8	52.8	38.5	42.4	16.2	24.6	35.7	39.6
	800	70.6	48.1	59.2	42.3	50.8	25.3	26.3	39.9	45.3

For evaluation on the UAVDT dataset, the same configuration was applied. Table 4 presents the achieved AP for each class and the mAP. The achieved mAP proved the superiority of the CFEMNet compared to the HRNet, especially for dense and small objects. An improvement of more than 8% was achieved in terms of mAP on the UAVDT dataset which indicates the efficacy of the proposed improvement applied to the HRNet model.

Table 5 shows the achieved F1 scores on the UAVDT dataset. Based on the achieved results, the proposed CFEMNet outperforms the HRNet by more than 6%. The F1 score proved the quality of the proposed model. The performance of the proposed CFEMNet model in terms of F1 score for smaller objects is greatly enhanced due to the use of CFEM and the new feature fusion technique. This result demonstrates that the suggested CFEMNet is primarily geared at detecting small objects, and the enhancements may successfully increase the model's detection accuracy for small objects.

Table 4: Achieved mAP on the UAVDT dataset

Model	Input size	AP (%)			mAP (%)
		Car	Bus	Truck	
HRNet	608	41.7	36.2	34.5	37.4
	800	51.3	43.7	49.4	48.1
CFEMNet	608	47.7	42.8	44.8	45.1
	800	59.2	52.2	52.9	54.7

Table 5: Achieved F1 score on the UAVDT dataset

Model	Input size	F1 score (%)			
		Car	Bus	Truck	Average
HRNet	608	49.7	43.3	44.6	45.8
	800	58.6	47.8	51.2	52.5
CFEMNet	608	57.3	43.8	45.6	48.9
	800	59.9	50.9	54.1	54.9

To validate the detection accuracy and speed of small objects, comparative studies with various object detection methods are performed on the Nvidia GTX 960 GPU using the VisDrone2018 validation set. All the presented models have been reproduced using our experimental environment and evaluated their performance in terms of mAP and FPS (Frames Per Second). Table 6 displays the experimental outcomes. Because of its basic structure and lightweight construction, the detection speed of the CFEMNet is comparable to that of YOLOv3, which can meet real-time constraints for vehicle detection in traffic management systems. Furthermore, the CFEMNet outperforms the state-of-the-art algorithms included in this comparison for detecting small and dense objects.

Table 6: Comparison against state-of-the-art models

Model	Speed (FPS)	mAP (%)
YOLOv3 [33]	8.2	39.9
YOLOv4 [29]	10.5	40.6
YOLOv8 [41]	7.6	41.3
SSD [31]	6.3	37.2
CFEMNet (ours)	7.1	44.7

CFEMNet's design optimizes computational complexity through a combination of an efficient backbone architecture, innovative feature extraction techniques, and adaptive attention mechanisms. These advantages make it a compelling choice for real-time vehicle detection in UAV applications, offering a balance between high performance and manageable computational demands. As urban environments continue to evolve, CFEMNet's capabilities can significantly contribute to the development of smart transportation systems that require efficient processing of visual data.

4.3 Limitations and Discussion

CFEMNet is designed to enhance vehicle detection in urban environments using UAV imagery, incorporating multi-scale feature learning and attention mechanisms to overcome challenges such as occlusion and dense traffic. However, like any deep learning model, CFEMNet's performance is influenced by external factors such as image quality, weather conditions, image resolution, and UAV viewing angles. These factors can significantly affect model accuracy and generalization, leading to specific limitations in real-world applications.

Low-quality images (e.g., due to noise, compression artifacts, or motion blur) degrade the performance of deep learning models because they limit the model's ability to extract key features. In the case of UAV imagery, variations in image quality can arise from factors such as drone movement (causing camera shake) or low-light conditions.

CFEMNet's detection accuracy relies on the ability to extract both spatial and semantic features from high-resolution images. Poor-quality images can obscure fine-grained details, making it difficult for the model to differentiate between vehicles, especially in densely populated traffic scenes. If input images presented, CFEMNet may struggle with detecting small, occluded, or partially visible vehicles, as the critical edges or textures needed for accurate segmentation and classification become blurred. Besides, handling traffic in low-light conditions where noise or lack of contrast between vehicles and background significantly affects feature extraction.

Weather conditions like rain, fog, snow, or strong sunlight can significantly degrade the quality of UAV-captured images. Rain or fog can blur images, reducing visibility, while strong sunlight or shadows can create regions of overexposure or deep contrast. Such conditions introduce occlusions and distortions in the imagery, making it harder for CFEMNet to extract the fine-grained spatial relationships needed for vehicle detection, especially in urban traffic where vehicles may already be partially occluded or packed closely together. The multi-scale feature extraction and attention mechanisms in CFEMNet, while capable of addressing occlusion to some extent, may not be sufficient to overcome these extreme conditions without additional information.

UAVs capture images from different viewing angles depending on flight paths and altitudes. The top-down view is typical for UAV-based traffic monitoring, but variations in angle (e.g., oblique views) can introduce perspective distortion that affects the apparent shape and size of vehicles. These distortions can confuse the model and lead to misclassification or inaccurate localization of vehicles. Additionally, occlusions can be more severe at oblique angles, where one vehicle may completely block the view of another. CFEMNet's ability to fuse features across scales can partially mitigate occlusion, but severe perspective distortion can significantly reduce model accuracy.

CFEMNet's detection accuracy may suffer when the UAV captures images at non-orthogonal angles where vehicles may appear distorted or elongated, making it harder for the model to recognize them as distinct objects. Occlusion problems are exacerbated at oblique angles, leading to missed detections, particularly for vehicles hidden behind larger ones.

To mitigate this limitation, CFEMNet would benefit from robust training based on data augmentation techniques such as artificially degrading image quality can help it generalize better to low-quality inputs. One way to overcome this limitation is to integrate weather-invariant techniques into the model by training with a variety of weather conditions through data augmentation (simulating rain, fog, shadows). Also, incorporating thermal or infrared imagery, which is less affected by adverse weather conditions than visible light, to complement UAV images and provides more reliable features in low-visibility environments.

Factors like image quality, weather conditions, resolution, and UAV viewing angle present significant challenges for CFEMNet's ability to generalize across diverse and dynamic environments. These limitations point to potential areas for future research, including robust preprocessing, multi-modal data integration, super-resolution techniques, and geometric transformation mechanisms to better handle variations in input data. Addressing these issues would enable CFEMNet to become a more versatile and reliable model for real-world UAV-enabled vehicle detection, particularly in the unpredictable conditions of smart city environments. A possible solution to the limitation related to different viewing angles is incorporating 3D-aware features or geometric transformations that can normalize objects to a canonical view, compensating for perspective distortion. Moreover, training CFEMNet on datasets that include images captured from multiple angles can improve the model's ability to generalize across different viewing conditions.

4.4 Ablation Study

To validate the efficacy of our model, ablation experiments on the Visdrone-DET2018 dataset were performed. Initially, the impact of combining convolution and MSA paths on detection outcomes. Then, we illustrate the influence of the suggested CFEM module on the performance enhancement of our detector utilizing different resolution features in the feature pyramid. Lastly, an ablation study was conducted on the local window MSA's window size setting in MFB to find best value. For all ablation experiments, the input size was fixed to 800.

To investigate the influence of the parallel design of the MSA path and the convolution path in the MFB on detection performance, the convolution path was first removed then the MSA path was removed. As shown in Table 7, the performance of the model with the attention path is comparable to that of the final model, indicating that the MSA path is prevalent in the MFB. This demonstrates that the MSA path is more advantageous compared to the convolution path in collecting the exact spatial context for obstructed and small vehicle situations in aerial images.

Table 7: Ablation study on the impact of the parallel design of the convolution and attention paths in the MFB

Model	FLOPs (G)	Parameters (M)	mAP (%)	Speed
With convolution path	137.3	29.4	42.6	11.5
With attention path	137.8	29.6	43.2	9.4
With both paths	138.2	29.9	44.7	7.1

Moreover, the model with a convolution path, with an attention path, and with both paths have an almost identical number of parameters, demonstrating that convolution and MSA may be incorporated with minimal computational cost based on the proposed design.

To determine the impact of the proposed CFEM with varied resolutions in the feature pyramid on the performance of the proposed model, the proposed local window MSA applied after the high-resolution features in the feature pyramid and the MSA applied after the low-resolution features in the feature pyramid, were removed. Table 8 compares the performance of the above detectors. CFEMNet with local window because low-resolution feature maps link to high-level semantic representation, MSA focuses on extracting semantic context. Furthermore, CFEMNet with Multi-head Self-Attention (MSA) focuses on spatial context extraction since high-resolution feature maps largely correlate to low-level spatial information. The average of the parameters θ_1 and θ_2 in the test

set to corroborate the achieved results. CFEMNet with local window MSA, the result reveals that $\theta_1 = 0.5831$ of the convolution paths is greater than $\theta_2 = 0.3942$ of the MSA paths, indicating that the convolution path is more relevant. And the outcome in CFEMNet with MSA is exactly the reverse.

Table 8: Impact of the proposed CFEM on the performance

Model	mAP (%)	$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$
Local window MSA	44.4	0.5831	0.3942
MSA	44.2	0.4016	0.5724
CFEMNet	44.7	0.5014	0.4848

The proposed model was trained with various scales of the local window on the Visdrone-DET2018 dataset to analyze the impact of the window size in MFB on detection performance. It is worth mentioning that a bigger window size should result in a broader receptive field and better context-capturing capabilities, but this will result in exponentially increasing memory use. So, the proposed local window MSA is intended to significantly reduce memory use while sacrificing low-performance degradation. For testing our approach efficacy, varied size windows experiments were conducted. Table 9 compares our CFEMNet performance with various scales of window size in MFB. As can be shown, increasing the window area improves performance. Consequently, Experiments 2, and 5 with the greatest window area perform the best. Nevertheless, larger sizes consume too much memory and do not correspond to the peculiarities of vehicle features. The proposed 1:2 designs can easily analyze the vehicle's overall information with a certain aspect ratio while using significantly less memory (Experiment 3), achieving great performance. As shown, the proposed 20×40 window outperforms the 40×20 window (comparison between Experiments 3 and 4) and performs roughly equally to a larger window (comparison between Experiments 3 and 5) with 0.9 GFLOPs (Giga Floating-point Operations Per Second) less. As a result, in order to achieve an effective balance between memory consumption and performance, we adjusted the size of local window in the MFB to 1:2.

Table 9: Impact of the local window MSA size

Experiment	Scale	size	mAP (%)	FLOPs
1	1:1	20×20	42.3	137.9
2	1:1	40×40	45.5	139.1
3	1:2	20×40	44.7	138.2
4	2:1	40×20	42.1	138.2
5	1:4	20×80	45.1	139.1

5 Conclusion

Most existing algorithms have difficulty detecting objects in UAV vision due to their small size and dense dispersion. This work proposes a new CFEMNet model that predicts the location of small vehicles with high density by using shallower-layer characteristics with rich fine-grained information. We present a basic yet effective CFEMNet for vehicle detection in crowded areas. The proposed CFEM

adopts a parallel design of the convolution and MSA paths. The MFB combines convolution and MSA at a cheap computing cost to extract exact context via the MSA path deep semantic context via the convolution path. Additionally, for high-resolution vehicle detection and efficiently balancing accuracy, speed and memory usage, we provided a specific local window. Our CFEMNet can acquire a high-level knowledge of substantially small and occluded vehicle situations based on HRNet by processing the context features provided by CFEM using the anchor-free detection head, effectively overcome HRNet's low feature extraction capacity for hard samples. The performance of the proposed model was proved through two challenging datasets. The achieved results demonstrated the robustness of the proposed CFEMNet for vehicle detection in aerial images provided by UAVs. We discovered that our model missed numerous challenging vehicle occurrences in some highly crowded situations during the visual examination of the testing findings. When the occlusion is too extreme, it appears impossible to collect the distinguishable properties of challenging vehicle situations by retrieving the whole context information. in future works, we will focus more on solving heavy occlusion challenges by providing more context information related to the occluded vehicle.

Acknowledgement: The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University for funding this work through research group No. (RG-NBU-2022-1234).

Funding Statement: This research was funded by the Deanship of Scientific Research at Northern Border University, Arar, Saudi Arabia through research group No. (RG-NBU-2022-1234).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yahia Said, Yahya Alassaf; data collection: Taoufik Saidani, Refka Ghodhbani, Olfa Ben Rhaiem, Ali Ahmad Alalawi; analysis and interpretation of results: Olfa Ben Rhaiem, Refka Ghodhbani, Ali Ahmad Alalawi; draft manuscript preparation: Yahia Said, Yahya Alassaf, Taoufik Saidani. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be made available on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] P. Zhao and H. Hu, "Geographical patterns of traffic congestion in growing megacities: Big data analytics from Beijing," *Cities*, vol. 92, pp. 164–174, Jul. 2019. doi: 10.1016/j.cities.2019.03.022.

[2] H. El-Sayed, M. Chaqfa, S. Zeadally, and D. Puthal, "A traffic-aware approach for enabling unmanned aerial vehicles (UAVs) in smart city scenarios," *IEEE Access*, vol. 7, pp. 86297–86305, 2019. doi: 10.1109/ACCESS.2019.2922213.

[3] W. Sun, X. Zhang, X. He, Y. Jin, and X. Zhang, "A two-stage vehicle type recognition method combining the most effective Gabor features," *Comput. Mater. Contin.*, vol. 65, no. 3, pp. 2489–2510, 2020. doi: 10.32604/cmc.2020.012343.

[4] C. P. Juan, O. H. Mondragon, and W. M. Mayor-Toro, "Fast and precise: Parallel processing of vehicle traffic videos using big data analytics," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12064–12073, 2021.

[5] R. Ayachi, M. Afif, Y. Said, and M. Atri, "Traffic signs detection for real-world application of an advanced driving assisting system using deep learning," *Neural Process. Lett.*, vol. 51, pp. 837–851, Feb. 2020. doi: 10.1007/s11063-019-10113-2.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.
[7] C. Li *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
[8] R. Ayachi, M. Afif, Y. Said, and A. B. Abdelali, "Real-time implementation of traffic signs detection and identification application on graphics processing units," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 7, 2021, Art. no. 2150024. doi: 10.1142/S0218001421500245.
[9] R. Ayachi, Y. Said, and M. Atri, "A convolutional neural network to perform object detection and identification in visual large-scale data," *Big Data*, vol. 9, no. 1, pp. 41–52, Mar. 2021. doi: 10.1089/big.2019.0093.
[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019, pp. 6569–6578.
[11] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020. doi: 10.1109/TPAMI.2020.2983686.
[12] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 20–25, 2009, pp. 1271–1278.
[13] H. Zhang *et al.*, "Faster R-CNN based on frame difference and spatiotemporal context for vehicle detection," *Signal Image Video Process.*, vol. 18, no. 10, pp. 7013–7027, Oct. 2024. doi: 10.1007/s11760-024-03370-3.
[14] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan and L. Zhang, "Dynamic DETR: End-to-end object detection with dynamic attention," in *Proc. of the IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 10–17, 2021, pp. 2988–2997.
[15] W. Chen, D. Xie, Y. Zhang, and S. Pu, "All you need is a few shifts: Designing efficient convolutional neural networks for image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 7241–7250.
[16] X. Pan *et al.*, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 815–825.
[17] A. Srinivas, T. -Y. Lin, N. Parmar, J. Shlens, P. Abbeel and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 19–25, 2021, pp. 16519–16529.
[18] W. Sun, X. Zhang, S. Shi, and X. He, "Vehicle classification approach based on the combined texture and shape features with a compressive DL," *IET Intell. Transp. Syst.*, vol. 13, no. 7, pp. 1069–1077, Nov. 2019. doi: 10.1049/iet-its.2018.5316.
[19] F. Tian *et al.*, "Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1751–1766, Apr. 2021. doi: 10.1109/TCSVT.2021.3080928.
[20] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita and F. Herrera, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," *Knowl.-Based Syst.*, vol. 194, pp. 1–10, Apr. 2020, Art. no. 105590.
[21] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 2117–2125.
[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 18–23, 2018, pp. 8759–8768.
[23] Y. Liu, F. Yang, and P. Hu, "Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks," *IEEE Access*, vol. 8, pp. 145740–145750, 2020. doi: 10.1109/ACCESS.2020.3014910.
[24] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in UAV vision based on cascade network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019.

[25] G. Ghiasi, T. -Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 7036–7045.
[26] H. Xu, L. Yao, W. Zhang, X. Liang, and Z. Li, "Auto-FPN: Automatic network architecture adaptation for object detection beyond classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019, pp. 6649–6658.
[27] E. V. Butilă and R. G. Boboc, "Urban traffic monitoring and analysis using unmanned aerial vehicles (UAVs): A systematic literature review," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 620. doi: 10.3390/rs14030620.
[28] O. Kainz, M. Dopiriak, M. Michalko, F. Jakab, and I. Nováková, "Traffic monitoring from the perspective of an unmanned aerial vehicle," *Appl. Sci.*, vol. 12, no. 16, 2022, Art. no. 7966. doi: 10.3390/app12167966.
[29] A. Bochkovskiy, C. -Y. Wang, and H. -Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[30] X. Wang, "Vehicle image detection method using deep learning in UAV video," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–8, Jan. 2022. doi: 10.1155/2022/8202535.
[31] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Comput. Vis.-ECCV 2016: 14th European Conf.*, Amsterdam, The Netherlands, Springer International Publishing, Oct. 11–14, 2016, pp. 21–37.
[32] R. Makrigiorgis, N. Hadjittoouli, C. Kyrkou, and T. Theocharides, "AirCamRTM: Enhancing vehicle detection for efficient aerial camera-based road traffic monitoring," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, Jan. 4–8, 2022, pp. 2119–2128.
[33] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," in *Computer Vision and Pattern Recognition*, Berlin/Heidelberg, Germany: Springer, 2018, vol. 1804, pp. 1–6.
[34] R. Eduardo, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, 2017.
[35] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," arXiv:1805.10180, 2018.
[36] L. -C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[37] M. Bakirci, "Enhancing vehicle detection in intelligent transportation systems via autonomous UAV platform and YOLOv8 integration," *Appl. Soft Comput.*, vol. 164, Jan. 2024, Art. no. 112015. doi: 10.1016/j.asoc.2024.112015.
[38] J. Chen, P. Wu, X. Zhang, R. Xu, and J. Liang, "Add-Vit: CNN-Transformer hybrid architecture for small data paradigm processing," *Neural Process. Lett.*, vol. 56, no. 3, p. 198, Apr. 2024. doi: 10.1007/s11063-024-11643-8.
[39] O. Havryliuk, "Automated annotation scheme for extending bounding box representation to detect ship locations," M.S. thesis, Sch. of Innov. Design Eng., Mälardalen Univ., Vasteras, Sweden, 2023.
[40] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 22–29, 2017, pp. 2980–2988.
[41] R. Varghese and M. Sambath, "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *Proc. 2024 Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, Mumbai, India, IEEE, Feb. 18–20, 2024, pp. 1–6.