Swin Transformer-Enhanced UAV Surveillance: A Multi-Modal Feature Optimization for High-Precision Road Vehicles Detection

Nouf Abdullah Almujally
Department of Information Systems, Princess Nourah bint Abdulrahman University, Saudi Arabia
Naalmujally@pnu.edu.sa

Ghulam Mujtaba
Faculty of Computing and AI
Air University, Pakistan
232697@students.au.edu.pk

Shuaa S. Alharbi
Department of Information Technology
Qassim University, Saudi Arabia
shuaa.s.alharbi@qu.edu.sa

Noif Alshammari
Department of Cyber Security
POSTECH University, South Korea
n.alshammari@mu.edu.sa

Ahmad Jalal
Department of Computer Science and Engineering
Korea University, South Korea
ahmadjalal@mail.au.edu.pk

Abstract: *Unmanned Aerial Vehicles (UAVs) have emerged as powerful platforms for intelligent traffic monitoring due to their high-resolution imaging and wide-area coverage. This paper introduces a robust vehicle detection and classification framework that employs a multi-modal feature optimization strategy to enhance detection accuracy in aerial environments. The proposed pipeline begins with Histogram Equalization for contrast enhancement, followed by semantic segmentation using DeepLabV3+ to accurately isolate vehicle regions. YOLOv10, a state-of-the-art real-time object detector, is then applied to localize vehicles with high precision. For feature extraction, we integrate three complementary modalities: Wavelet Transform Features (capturing multi-resolution frequency details), Gabor Filters (highlighting directional textures), and Speeded-Up Robust Features (SURF) (detecting keypoints and descriptors). A Genetic Algorithm (GA) is employed to optimize the extracted features by selecting the most discriminative subset, thus reducing redundancy. Final classification is performed using the Swin Transformer, a vision transformer that utilizes shifted window self-attention to model long-range spatial dependencies effectively. Experimental evaluations on two UAV benchmark datasets, Roundabout Aerial Images and VAID which demonstrate the superiority of our method, achieving classification accuracies of 97.71% and 98.57%, respectively. These results demonstrate the effectiveness, scalability, and real-world applicability of our approach in UAV-based vehicle monitoring, contributing to the advancement of autonomous aerial surveillance systems for intelligent transportation analytics and enhanced situational awareness in smart city applications.*

Keywords: *Pattern recognition, traffic management, YOLOv10, autonomous vehicles, intelligent traffic systems, track former, deeplabv3+.*

Received March 21, 2025; accepted September 10, 2025
https://doi.org/10.34028/iajit/23/1/13

1. Introduction

Unmanned Aerial Vehicles (UAVs) have emerged as pivotal tools in modern intelligent transportation systems, offering significant advantages such as high mobility, elevated perspectives, and the ability to rapidly capture extensive traffic data. Unlike fixed ground-based sensors, UAVs provide flexible and scalable monitoring capabilities over wide urban areas, highways, and intersections, making them particularly suitable for real-time traffic surveillance, congestion analysis, and autonomous navigation support [12]. Despite their advantages, analyzing UAV imagery for vehicle detection and classification remains a complex challenge due to several factors, like vehicles in aerial views often appearing small and visually ambiguous, environmental conditions such as varying illumination, motion blur, and occlusion further complicate detection, and dynamic traffic scenes introduce considerable background clutter.

These conditions make it difficult for traditional machine learning methods and early deep learning models to achieve consistent performance. While recent advancements in deep Convolutional Neural Networks (CNNs) and transformer-based architectures have improved detection accuracy in controlled settings, their effectiveness in aerial surveillance is limited. For example, YOLO-based detectors provide real-time performance but often miss small or partially occluded vehicles. Semantic segmentation models like DeepLabV3+ offer improved object boundary delineation but struggle in scenes with high object density or complex textures. Moreover, most existing approaches rely on single-type feature representations, typically either spatial or frequency-based, resulting in suboptimal generalization across diverse aerial environments. This exposes a critical gap in the literature: the lack of a unified framework that can robustly detect and classify vehicles from UAV imagery

by combining rich multi-modal features with advanced feature optimization and classification techniques.

To address these limitations, we propose a novel, multi-stage vehicle monitoring framework specifically designed for UAV-based aerial surveillance. Our method integrates advanced components across the entire perception pipeline to enhance performance at every stage. First, we apply histogram equalization to improve image contrast under varying lighting conditions. Next, semantic segmentation is performed using DeepLabV3+, which accurately isolates vehicle regions by leveraging atrous spatial pyramid pooling. For object detection, we incorporate YOLOv10, a state-of-the-art real-time detector optimized for aerial scenes with complex backgrounds. We then introduce a multi-modal feature extraction strategy, combining Wavelet Transform for multi-resolution analysis, Gabor Filters for directional texture encoding, and Speeded-Up Robust Features (SURF) for local keypoint-based descriptors. To eliminate redundancy and improve feature discriminability, a GA is employed for optimal feature selection. Finally, vehicle classification is conducted using the Swin Transformer, a hierarchical vision transformer that uses shifted window attention to model long-range spatial relationships efficiently and accurately. By combining all elements into a cohesive architecture, the proposed framework addresses the key challenges in UAV-based vehicle monitoring: small object detection, cluttered backgrounds, redundant feature representation, and classification under varying environmental conditions. The system is designed for real-world deployment, with strong applicability in smart city traffic analytics, automated law enforcement, and autonomous vehicular systems.

To summarize, the primary findings of our research can be outlined as follows:

1. We developed a comprehensive vehicle detection and classification framework that integrates advanced deep learning techniques with traditional feature extraction and optimization strategies for UAV-based aerial imagery.
2. Our preprocessing and segmentation pipeline, utilizing Histogram Equalization and DeepLabV3+, significantly enhances image quality and improves vehicle region extraction.
3. The adoption of YOLOv10 enables highly accurate vehicle detection, making it well-suited for UAV surveillance applications.
4. A multi-faceted feature extraction approach, incorporating wavelet transform features, gabor filters, and SURF, effectively captures essential vehicle characteristics.
5. Feature optimization using GA improves classification performance by selecting the most relevant features while reducing computational overhead.
6. The swin transformer-based classification model achieves superior accuracy, demonstrating its capability to handle complex aerial imagery.

2. Related Work

Researchers have done extensive research on UAV aerial surveillance for many years to build better traffic monitoring and identify vehicles. Studies about detecting targets from UAV cameras use traditional and new learning techniques to create better systems that can find things more accurately. At the research's start phase, scientists used manual sensor characteristics along with actual image processing, but classic vision methods showed they could not handle uneven vehicle sizes and environmental variations plus partial blockages. Advanced traffic analysis systems of UAVs work better using object detection technology from CNN and Transformer models. This section evaluates all important aspects of UAV vehicle detection research through its methods and newer detection strategies, along with system limitations.

2.1. Vehicle Detection and Classification Systems

UAV technology provides real-time traffic and security monitoring along with disaster response services by processing detailed images and moving unstaffed devices where needed. Standard vehicle detection systems built with static sensors and normal machine learning methods deal poorly with partial obscuration and changes in object size against environmental factors. Deep learning has made UAV-based vehicle detection and classification work better through the networks YOLO, Faster R-CNN, and Transformers because of their improved accuracy and performance. Hamzenejadi *et al*. [8] addressed the trade-off between detection accuracy and inference speed by modifying YOLOv5's network width and depth, achieving a 3.7% mAP50 improvement and a 6.1 FPS increase on the VisDrone and CARPK datasets while reducing model size by 44.6 MB. Similarly, Li *et al*. [14] enhanced YOLOv5-VTO by adding an extra prediction head for small-scale object detection and integrating a Bidirectional Feature Pyramid Network (BiFPN) to improve multi-scale feature fusion. Their use of Soft Non-Maximum Suppression (Soft-NMS) reduced false detections, leading to a 3.7% increase in mAP@0.5 and a 4.7% improvement in mAP@0.5:0.95, demonstrating the effectiveness of feature-based enhancements in UAV-based vehicle detection. Kumar *et al*. [13] tackled UAV-based Indian traffic surveillance, where dense and unstructured road conditions pose significant detection challenges. By employing the YOLOv8 model trained on a custom drone-captured dataset and incorporating preprocessing techniques like Gaussian filtering, resizing, normalization, and augmentation, they achieved 0.86 mAP50, validating the model's

applicability for real-world traffic monitoring and law enforcement. Beyond civilian applications, UAV-based vehicle detection is increasingly relevant for military surveillance. The research of Gupta *et al*. [7] introduced a publicly accessible military vehicle dataset including 6772 images that contain Military Trucks, Tanks, Aircraft, Helicopters, Civilian Cars, and Civilian Aircraft. The researchers tested Quantized SSD MobileNet v2 and Tiny YOLOv3 against each other and concluded that Tiny YOLOv3 delivered superior precision with better efficiency, thus making it a more suitable solution for UAV-based surveillance with limited resources. The researchers created mathematical equations for determining perfect flight paths and frame coverage dynamics, which optimized tasks during real-time reconnaissance operations. These latest research studies show how UAV deep learning systems grow stronger and confirm their ability to spot vehicles properly across various working conditions. Future work in drone vehicle detection must include two-step methods that combine different sensory inputs plus advanced edge computing for both day and nighttime operations. Future research should focus on enhancing UAV-based vehicle detection by integrating multi-modal sensor data, including LiDAR and thermal imaging, to improve detection performance in challenging environments such as low visibility and nighttime conditions. Addressing challenges related to occlusion, scale variation, and adverse weather conditions will be key to developing more robust and adaptable UAV surveillance systems for both civilian and defence applications

3. Methodology

Our UAV-based vehicle monitoring framework integrates deep learning techniques with feature optimization strategies for accurate vehicle detection and classification, as depicted in Figure 1. The process begins by using Histogram Equalization to make the images easier to see and by adjusting their overall brightness. DeepLabV3+ defines vehicle edges well during the region extraction step. YOLOv10 becomes our chosen object detection model because it processes vehicle locations swiftly through specialized aerial scene analysis. To build more detailed feature descriptions we take wavelet transform features, gabor filters, and SURF from both spatial and frequency-based data. The GA system tests and picks extraction results with optimal attributes that save processing time and improve detection accuracy. At last swin transformer handles vehicle classification using its hierarchical attention system to deliver high-quality results. Our system was tested on two datasets, achieving a classification accuracy of 97.71% on the roundabout aerial images dataset and 98.57% on the VAID dataset, highlighting its robustness and reliability.

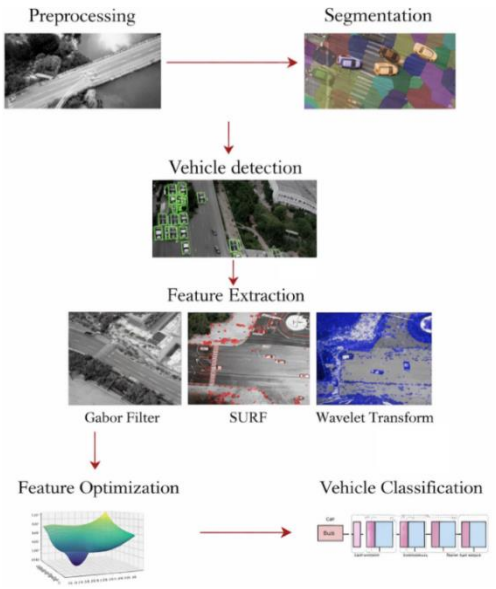

Figure 1. Proposed framework of UAVs traffic monitoring.

3.1. Preprocessing via Histogram Equalization

Preprocessing is a very important step in UAV-based vehicle monitoring, ensuring that raw aerial images are enhanced for improved feature extraction and detection accuracy. Histogram Equalization is employed to enhance image contrast by redistributing intensity values across the histogram, thereby improving visibility in varying lighting conditions [17]. Given an input UAV image, $I\,(x,y)$ with intensity values ranging from *0* to *L-1* the enhanced image $I_{HE}\,(x, y)$ is obtained using a novel adaptive histogram equalization function:

$$I_{HE}\,(x,y)^n = \frac{L-1}{N \times M} \sum_{I=0}^{I(X,Y)} p^{\,\alpha i} \tag{1}$$

Whereas P^i represents the probability density function of intensity i, N and M denote the image dimensions, and α is an adaptive enhancement factor that dynamically adjusts contrast based on local intensity distributions. Unlike traditional HE, this formulation ensures adaptive contrast enhancement tailored to UAV images, reducing over-enhancement artifacts while preserving essential details. This preprocessing step significantly refines image quality, improving segmentation accuracy in DeepLabV3+, leading to more precise vehicle detection using YOLOv10, and optimizing feature extraction performance for subsequent classification. The Output of the Said algorithm are shown in Figure 2.

Figure 2. Preprocessing via histogram equalization.

3.2. Segmentation Using Deeplabv3

UAV-based vehicle monitoring heavily depends on segmentation as it creates clear vehicle boundaries which allows for proper detection as well as classification precision. DeepLabV3+ serves as the state-of-the-art deep learning model for semantic segmentation to extract regions of vehicles accurately. An Atrous Spatial Pyramid Pooling (ASPP) module with multiple scale context capabilities exists in this model through the application of dilated convolutions at various rates [19]. The decoder proceeds to refine object boundaries, which makes it an ideal solution for UAV imagery because objects occur at different scales and experience obscured conditions. Given an input preprocessed UAV image I_{HE} (x, y) the segmented output S (x, y) is obtained as:

$$S(x,y) = \sigma \sum_{i=1}^{\infty} wi \,.\, fASPP\,(I_{HE}(x,y), ri) + b) \qquad (2)$$

Whereas f *ASPP* represent the multi-scale atrous convolutional operation applied at different dilation rate ri . wi are the learned weight b is the bias term and σ denotes the softmax activation function for pixel-wise classification. Unlike conventional segmentation methods, this adaptive feature extraction mechanism enables DeepLabV3+ to effectively differentiate vehicles from shadows, roads, and other objects in UAV imagery [15]. This segmentation step significantly enhances detection accuracy in the next stage, where YOLOv10 is utilized for precise vehicle localization. Figure 3 illustrates the segmented vehicle regions obtained using the proposed methodology, demonstrating the model's effectiveness in distinguishing vehicles from the surroundings with high precision whereas Figure 4 shows the architecture of Deeplabv3. This precise segmentation output serves as a critical foundation for the subsequent vehicle detection and classification stages, ensuring higher accuracy and reliability in UAV-based surveillance scenarios.

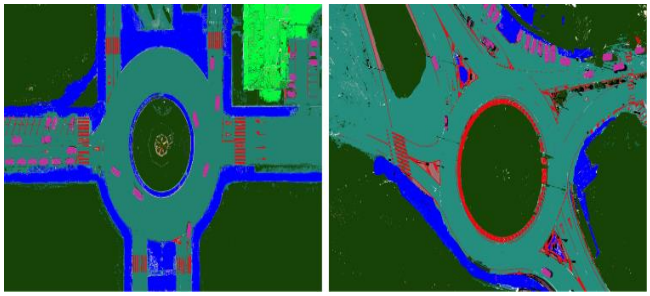

Figure 3. Results of road segmentation.

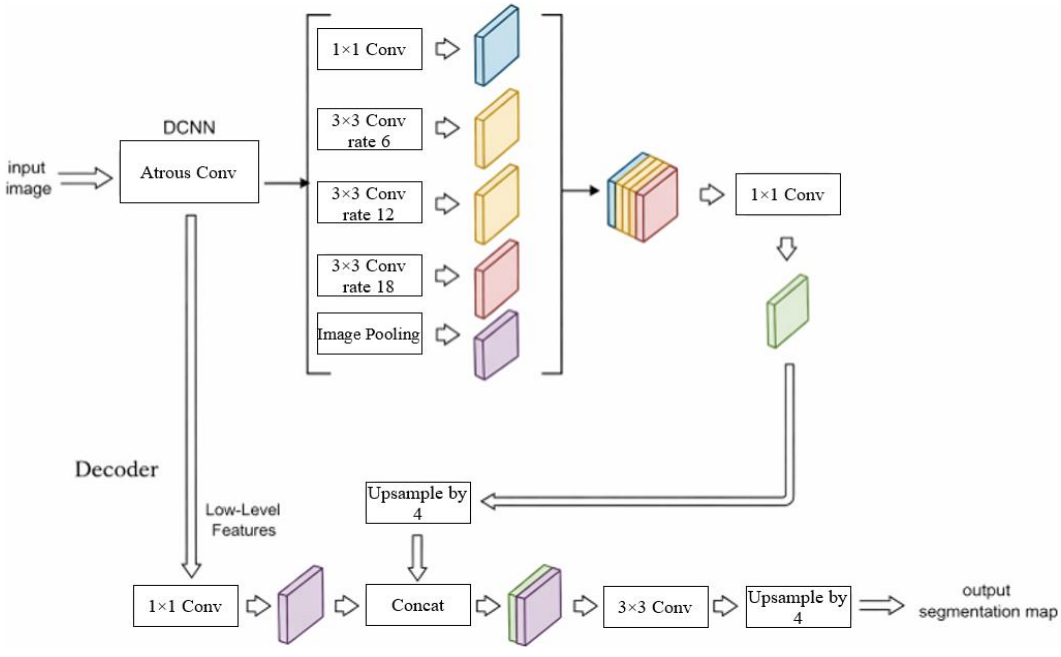

Figure 4. Architecture of deeplabv3.

3.3. Vehicle Detection Via YOLOv10

After achieving accurate segmentation using DeepLabV3+, the next crucial step is vehicle detection, where we localize vehicles within [5] the segmented UAV images. For this, we employ YOLOv10, a cutting-edge real-time object detection model known for its speed and precision [20]. YOLOv10 efficiently detects vehicles by leveraging an enhanced CSP-based backbone for feature extraction, a Path Aggregation Network (PAN) Neck for multi-scale feature fusion, and an optimized Detection Head for accurate bounding box regression and classification. These advancements enable YOLOv10 to detect vehicles with high accuracy, even in aerial images with varying scales, occlusions, and complex backgrounds. Given a Segmented UAV image $S(x, y)$ the detected vehicle bounding boxes B_i are determined as:

$$B_i = arg\ max_{(x,y,w,h)} \sigma\,(f\ YOLO\ (S(x,y), 0 \qquad (3)$$

Where f *YOLO* represent the *YOLOv*10 detection model θ denotes the learned model parameters, and $max_{(x,\,y,\,w,\,h)}$

defines the bounding box coordinates with width and height. The sigmoid activation function σ ensures probabilistic confidence scores for vehicle detection. The integration of DeepLabV3+ segmentation with *YOLOv*10 reduces false detections, enhancing model reliability. *YOLOv*10's multi-scale feature extraction and anchor-free detection strategy allow it to outperform previous versions, making it highly effective for UAV-based vehicle monitoring applications. Figure 5 shows the detected vehicles with bounding boxes, while Figure 6 illustrates the *YOLOv*10 architecture used in this study. These results validate the robustness of our detection pipeline, ensuring high precision and recall.

Figure 5. YOLOv10 based vehicle detection.

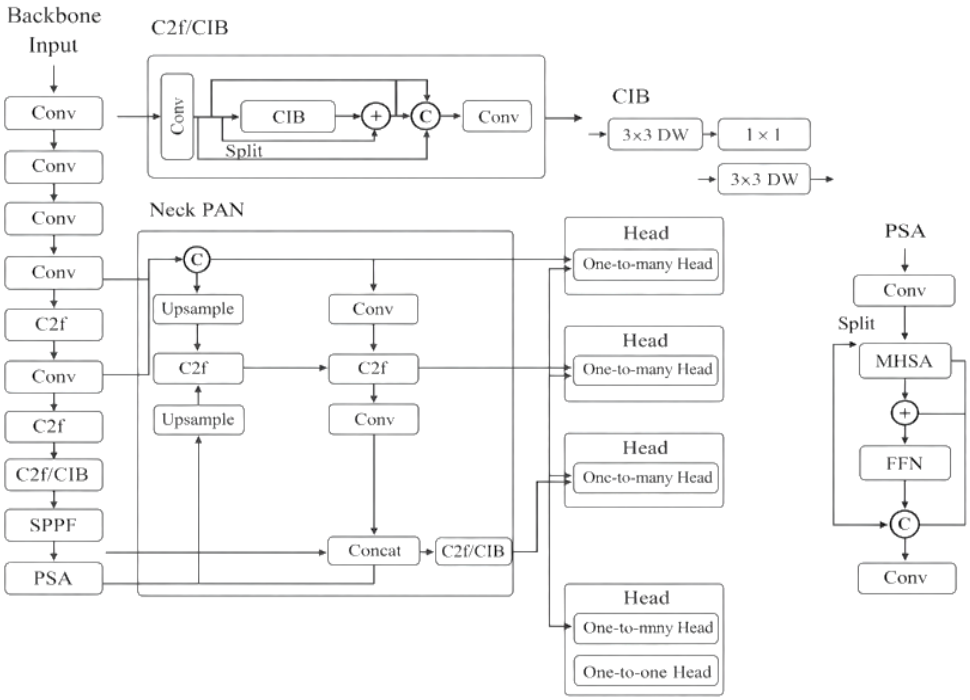

Figure 6. Overview of the yolo object detection algorithm.

3.4. Feature Extraction

UAV-based vehicle monitoring requires an essential process called feature extraction, which converts plain data images into object-defining characteristics for precise detection alongside classification. Through the methodology, the essential characteristics of vehicle form-shape and surface texture and structural elements are detected against background elements. The combination of Wavelet Transform and Gabor Filters and SURF processes strengthens features because these techniques defend spatial data and frequency data simultaneously. The recognition performance benefits from effective feature extraction, which leads to enhanced efficiency of optimization and classification processes.

3.4.1. Wavelet Transform Feature Extraction

We employ Wavelet Transform Features, which provide multi-resolution analysis by decomposing an image into different frequency components. This allows for better representation of vehicle shapes, edges, and textures, particularly in aerial images where vehicles appear at different scales and orientations [21]. Wavelet Transform decomposes the detected vehicle region $V(x, y)$ into multiple sub-bands using a series of low-pass and high-pass filters. The transformed feature set $Wf(x,y)$ can be expressed as:

$$Wf(x,y) = \sum_{m=0}^{M-1\,N-1} V(x-m, y-n) \cdot \Psi l, h(m,n) \quad (4)$$

Where $\Psi i, h(m, n)$ represents the wavelet basis function, with l and h denoting the low-pass and high-pass filters. The decomposition process generates four sub-bands: LL (approximation), LH (horizontal details), HL (vertical details), and HH (diagonal details), capturing fine details crucial for vehicle recognition. By extracting statistical and energy-based wavelet coefficients from these sub-bands, we obtain a compact yet discriminative feature representation. This improves the robustness of the model by preserving crucial textural patterns while reducing redundancy [6]. Figure 7 illustrates the extracted wavelet features, which serve as a critical input for the subsequent feature optimization and

classification stages, ensuring enhanced vehicle recognition performance.

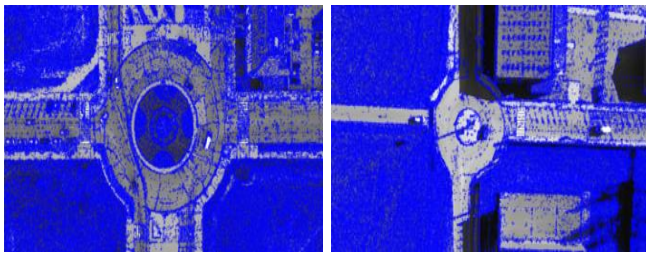

Figure 7. Feature extraction via wavelet.

3.4.2. Gabor Filter

After wavelet feature extraction, we further enhance texture representation using Gabor Filters, which are highly effective in capturing spatial frequency, orientation, and edge information. Gabor filters are particularly useful for UAV-based vehicle monitoring, as they mimic the human visual system in detecting directional patterns and textures [10]. A Gabor Filter is defined as:

$$G(x,y;\theta,\lambda,\sigma,\gamma) = exp\left(-\frac{x^2+\gamma^2 y_2}{2\sigma^2}\right) Cos\ (\frac{2\pi x}{\lambda}+\phi) \qquad (5)$$

Where $x'=x\ cos0+y\ sin0$ and $y'=-xsin0+ycos$ represent the rotated coordinates 0 is the filter orientation, λ is the wavelength of the sinusoidal component, σ is the standard deviation of the Gaussian envelope, γ is the spatial aspect ratio, and ϕ is the phase offset. We extract rich texture-based features that enhance vehicle classification by convolving the detected vehicle regions with multiple Gabor kernels at different orientations and scales [1]. These features and wavelet coefficients contribute to a more robust and discriminative representation, further improving classification accuracy in the later stages. Figure 8 illustrates the extracted Gabor filter responses, highlighting the texture variations and edge details captured for enhanced vehicle representation

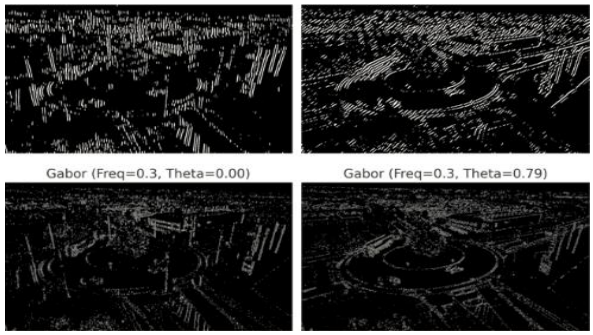

Figure 8. Feature extraction via gabor filter.

3.4.3. Feature Extraction Via SURF

To further enhance feature representation, we employ SURF, a robust key point detection and descriptor extraction technique. SURF efficiently detects distinctive vehicle features by leveraging an integral image-based Hessian matrix approximation, making it highly suitable for UAV-based vehicle monitoring where scale and rotation invariance are crucial [2]. The Detected key points and descriptors S^f for a given vehicle region $V(x, y)$ are computed as:

$$(S_f) = \sum_{i=1}^{n} H(x_i\ y_i).D(x_i\ y_i) \qquad (6)$$

Where $H=(xi, yi)$ represents the Hessian determinant response at key point (xi, yi) and $D(xi, yi)$ denotes the descriptor vector capturing local gradient information. The combination of key points and descriptors forms a highly discriminative feature set, enabling precise vehicle recognition [3]. Figure 9 presents the detected SURF key points, showcasing their robustness in capturing vehicle-specific patterns and structures.

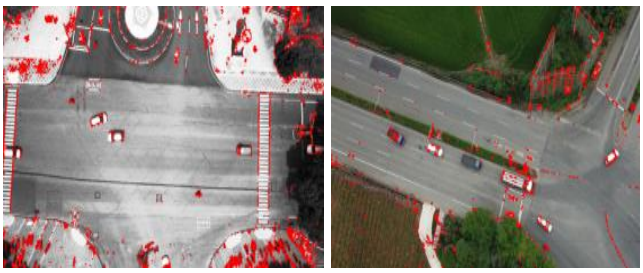

Figure 9. SURF based feature extraction.

3.5. Feature Optimization via Genetic Algorithm

After extracting multi-scale features from Wavelet Transform, Gabor Filters, and SURF, the next step is feature optimization to enhance classification accuracy and computational efficiency [10]. We employ a GA, a powerful evolutionary optimization technique, to select the most discriminative features while eliminating redundant and irrelevant ones [9]. GA mimics the natural selection process by iteratively evolving a population of feature subsets through selection, crossover, and mutation operations. The optimal feature subset F is obtained using the following equation:

$$F = \arg\max(\frac{\sum_{j=1}^{N} wj.fi}{Fi}), wj = \frac{1}{1+e^{-bj(Fi)}} \qquad (7)$$

Where F represents the full feature set, Fi is a candidate subset, fj are individual features, wj are adaptive weights computed using a sigmoid-based fitness function, β controls selection pressure, and $J(Fi)$ represents the classification accuracy achieved using the subset Fi. This approach ensures that only the most informative features contribute to the final classification stage. In our implementation, the Genetic Algorithm was configured with a population size of 40, running for 60 generations. A crossover rate of 0.8 and a mutation rate of 0.05 were selected based on preliminary grid search experiments. The algorithm was terminated once either the classification accuracy plateaued over 10 consecutive generations or the maximum generation limit was

reached. This setup ensured a balance between convergence stability and computational efficiency during feature subset selection. By optimizing the feature set using GA, we achieve a compact and highly discriminative representation, reducing computational complexity while maintaining superior classification performance.

3.6. Vehicle Classification Using Swin Transformer

The classification of detected vehicles into specific categories constitutes the last step after the Genetic Algorithm optimization of extracted features in our pipeline. This work uses the Swin Transformer as its vehicle classification model because of its innovative shifted window-based attention system, which builds upon traditional Vision Transformers (ViTs). The Swin Transformer provides better performance than traditional CNNs because it successfully analyzes vehicle images through both short-range and distant relationships and handles different image scaling dimensions [4]. The Swin Transformer processes input features through a hierarchical architecture, progressively increasing the receptive field while maintaining computational efficiency. The model starts by embedding the optimized feature set $F*$ into a series of patch embeddings, which are then passed through multiple swin transformer blocks. Each block consists of Shifted Window Multi-Head Self-Attention (SW-MSA) layers, which allow for more effective feature interaction across different spatial locations. This is particularly useful in UAV imagery, where vehicles may appear at various scales, orientations, and lighting conditions. Mathematically, the Classification Output C is computed as follows:

$$C = Softmax\ (W_s. \sum_{i=0}^{n} A(W_t F_i) \quad (8)$$

Where W_t represents the trainable weights of the Transformer Encoder, $A(\cdot)$ is the self-attention function, which enhances feature representations by learning spatial dependencies, and W_S is the classification head that maps the learned representations to class labels. Softmax executes to find final classification probabilities that determine confidence scores for vehicle categories. Through its shifted window method, the Swin Transformer excel at processing large aerial images with high resolution efficiently [11].

The model processes aerial image sections rather than entire pictures to enhance feature collaboration with smaller processing requirements. For different types of UAV images, our method improves both the accuracy and flexibility of model applications. Training progress stops when both the adaptive learning rate and cross-entropy loss achieve optimal performance to minimize misclassifications. Swin Transformer model displays its full design elements and explains its techniques for feature extraction and attention linking in Figure 10.

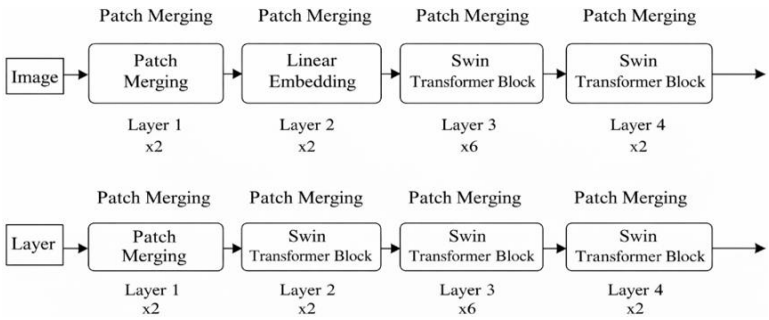

Figure 10. Swin transformer architecture.

4. Result and Analysis

The testing took place on a system that featured an Intel Core i7-12700H (2.70 GHz) CPU with 32 GB RAM and used a NVIDIA RTX 3060 GPU that provided 6 GB VRAM. The created system combined PyTorch with TensorFlow in addition to CUDA acceleration which boosted deep learning processing speed. Our deep learning model needed perfect settings that we gained from extensive hyperparameter adjustments, including weight rules during learning rate selection and selection of the best batch size

4.1. Dataset Description

4.1.1. Roundabout Aerial Image

The roundabout dataset utilizes UAV imagery to classify vehicle types even in busy traffic scenes. The Roundabout Aerial Image Dataset contains both aerial images and video sequences that capture different vehicle types and show how cars interact through multiple lanes even when they partly disappear from view [18]. The dataset includes two types of labeling to assist traffic research. It shows boundaries between objects and tags traffic types for deep learning studies.

4.1.2. VAID Dataset

The VAID dataset offers testing tools to detect objects with air vehicle identification followed by tracking their movements [16]. The VAID dataset consists of 100 video sequences amounting to 80,000 frames, which were acquired through a UAV platform. The video material crosses 10 hours of continuous footage that

displays different city settings. Each JPG image has a 1080×540-pixel resolution at 30 frames per second as its capture parameters. T-junctions and arterial routes, together with highways and squares as well as crossings, represent the various road configurations found in this dataset.

4.2. Performance Evaluation

The system evaluation occurred through rigorous testing with Roundabout Aerial Image and VAID datasets to validate its effectiveness as a UAV-based vehicle monitoring solution. The accuracy of the performance assessment relied on five independent tests where researchers recorded the mean results. Precision and recall along with F1-score served as the key evaluation metrics to measure both the algorithmic accuracy and robustness during vehicle detection operations. The UAV-based vehicle surveillance model proposed has a high detection and classification rate in benchmark datasets. Tables 1 and 2 indicate class-wise precision, recall, and F1-score on the Roundabout and VAID datasets, with average F1-scores of 0.97 and 0.96.

Table 1. Vehicle detection accuracy, precision, recall, and f1-score evaluation of roundabout dataset.

Classes	Precision	Recall	F1-score
C	0.97	0.96	0.96
Tru	0.98	0.97	0.97
B	0.97	0.95	0.96
Cy	0.98	0.95	0.96
V	0.97	0.95	0.96
MB	0.99	0.97	0.98
Tra	0.99	0.98	0.98
Mean	**0.98**	**0.96**	**0.97**

Table 2. Vehicle detection accuracy, precision, recall, and f1-score evaluation of VAID dataset.

Classes	Precision	Recall	F1-score
C	0.96	0.95	0.95
Tru	0.97	0.96	0.96
B	0.96	0.94	0.95
Cy	0.97	0.93	0.95
V	0.96	0.94	0.95
MB	0.98	0.96	0.97
Tra	0.98	0.97	0.97
Mean	**0.97**	**0..95**	**0.96**

Table 3 shows that the method outperforms state-of-the-art models, including SSD, RetinaNet, YOLOv5, and EfficientNet.

Table 3. Comparison of model detection rate with other state-of-the-art methods.

Datasets	Models	Precision
Roundabout dataset	SSD	0.76
	RetinaNet	0.68
	Blob detection	0.73
	Our method	**0.97**
VAID dataset	Yolov6	0.84
	Yolov5	0.69
	EfficientNet	0.83
	Our method	**0.96**

Tables 4 and 5 show the confusion matrices having an overall classification accuracy of 97.71% and 98.57% respectively which means that the inter-class misclassification is not much.

Table 4. Confusion matrix for vehicle classification over the VAID dataset.

Classes	C	Tru	B	Cy	V	MB	Tra
C	**99**	0	1	0	0	0	0
Tru	0	**98**	1	0	0	1	0
B	0	0	**97**	1	1	1	0
Cy	0	0	0	**98**	0	1	1
V	0	0	1	0	**98**	0	1
MB	0	0	1	0	1	**97**	1
Tra	0	0	1	1	0	1	**97**
Mean: 97.71%							

Table 5. Confusion matrix for vehicle classification over the roundabout aerial dataset.

Classes	C	TR	B	SD	V	MB	Tra
C	**98**	0	1	0	0	0	1
TR	1	**99**	0	0	0	0	0
B	1	0	**98**	0	1	0	0
SD	0	1	0	**99**	0	0	0
V	0	0	1	0	**98**	1	0
MB	0	0	0	0	1	**99**	0
Tra	0	0	0	1	0	0	**99**
Mean: 98.57%							

*Mn=Minibus, TR=Truck, PT=Pickup Truck, B=Bus, SD=Sedan, C=Car, CT=Cement Truck, Tra=Trailer.

Table 6 highlights superior or competitive performance compared to existing studies, confirming the framework's robustness, reliability, and effectiveness for UAV-based traffic surveillance in diverse scenarios.

Table 6. Classification Comparison with other State-of-the-art Models.

Method	Roundabout	VAID
Lin *et al.* [16]	--	91.3%
Hussein *et al.* [10]	--	95.50%
Kumar *et al.* [13]	86.7%	--
Gopta *et al.* [7]	89.5%	--
Proposed method	97.71%	95.50%

The proposed model shows remarkable accuracy together with strong robustness levels for UAV-based vehicle detection yet it requires certain specified limitations to be examined. The detection system faces important difficulties from environmental objects that partially or fully hide vehicles because this condition leads to classification errors. The detection accuracy and feature extraction process get impaired through weather elements that combine rain and fog with decreased visibility when observing images. The model operates best during daylight conditions since it has not been optimized to perform effectively under nighttime conditions where illumination problems and sensor noise function as significant performance decreases. Furthermore, while the model performs well in structured environments, highly congested and unstructured traffic scenarios may introduce complexities due to overlapping vehicles and varying perspectives. Future work should explore the integration of multi-spectral imaging, low-light enhancement techniques, and domain adaptation strategies to improve

the model's robustness under diverse conditions.

5. Conclusions

The research introduced an effective UAV-based vehicle detection and classification structure that applied state-of-the-art deep learning methodologies along with optimized feature extraction methods. The proposed method unites histogram equalization image enhancement with DeepLabV3+ segmentation and YOLOv10 vehicle detection and a feature extraction system that utilizes Wavelet Transform and Gabor Filters and SURF to analyze spatial and frequency domain signal information. A Genetic Algorithm served to optimize feature selection through an improvement of both computational efficiency and classification accuracy. Vehicle classification was performed by using the Swin Transformer, which demonstrated exceptional capabilities for detecting fine-scale information. The proposed framework was tested on two benchmark UAV datasets, achieving a classification accuracy of 97.71% on the Roundabout Aerial Images dataset and 98.57% on the VAID dataset, demonstrating its strong performance. This work contributes significantly to the advancement of autonomous aerial surveillance, offering a scalable and high-performance solution for intelligent traffic analysis.

Acknowledgment

The authors acknowledge Princess Nourah Bint Abdulrahman University Researchers supporting Project number (PNURSP2025R410), Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia.

References

[1] Al Mokhtar Z. and Dawwd S., "3D VAE Video Prediction Model with Kullback Leibler Loss Enhancement," *The International Arab Journal of Information Technology*, vol. 21, no. 5, pp. 879-888, 2024. DOI:10.34028/iajit/21/5/9

[2] Bhosale B., Kayastha V., and Harpale V., "Feature Extraction Using SURF Algorithm for Object Recognition," *International Journal of Technical Research and Applications*, vol. 2, no. 4, pp. 197-199, 2014. https://www.ijtra.com/view.php-paper-feature-extraction-using-surf-algorithm-for-object-recognition.pdf

[3] Bilik S. and Horak K., "SIFT and SURF-Based Feature Extraction for Anomaly Detection," *arXiv Preprint*, vol. arXiv:2203.13068, pp. 1-7, 2022. https://arxiv.org/pdf/2203.13068

[4] Chen Y., Feng J., Liu J., Pang B., and et al., "Detection and Classification of Lung Cancer Cells Using Swin Transformer," *Journal of Cancer Therapy*, vol. 13, no. 7, pp. 464-475, 2022, https://www.scirp.org/journal/paperinformation?paperid=118642

[5] Cheng X. and Zhang P., "Enhanced Soccer Training Simulation Using Progressive Wasserstein GAN and Termite Life Cycle Optimization in Virtual Reality," *The International Arab Journal of Information Technology*, vol. 21, no. 4, pp. 549-559, 2024. DOI: 10.34028/iajit/21/4/1

[6] Ghazali K., Mansor M., Mustafa M., and Hussain A., "Feature Extraction Technique Using Discrete Wavelet Transform for Image Classification," *in Proceedings of the 5th Student Conference on Research and Development*, Selangor, pp. 1-4, 2007. https://doi.org/10.1109/SCORED.2007.4451366

[7] Gupta P., Pareek B., Singal G., and Rao D., "Edge Device-based Military Vehicle Detection and Classification from UAV," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19813-19834, 2022. DOI: 10.1007/s11042-021-11242-y

[8] Hamzenejadi M. and Mohseni H., "Real-Time Vehicle Detection and Classification in UAV Imagery Using Improved YOLOv5," *in Proceedings of the 12th International Conference on Computer and Knowledge Engineering*, Mashhad, pp. 231-236, 2022. DOI: 10.1109/ICCKE57176.2022.9960099

[9] Homaifar A., Qi C., and Lai S., "Constrained Optimization via Genetic Algorithms," *Simulation*, vol. 62, no. 4, pp. 242-253, 1994. https://doi.org/10.1177/003754979406200405

[10] Hussein F., Kharma N., and Ward R., "Genetic Algorithms for Feature Selection and Weighting: A Review and Study," *in Proceedings of the 6th International Conference on Document Analysis and Recognition*, Seattle, pp. 1240-1244, 2001. https://doi.org/10.1109/ICDAR.2001.953980

[11] Jamali A. and Mahdianpari M., "Swin Transformer and Deep Convolutional Neural Networks for Coastal Wetland Classification using Sentinel-1, Sentinel-2, and LiDAR Data," *Remote Sensing*, vol. 14, no. 2, pp. 359, 2022. https://doi.org/10.3390/rs14020359

[12] Kanistras K., Martins G., Rutherford M., and Valavanis P., "A Survey of Unmanned Aerial Vehicles (UAVs) for Traffic Monitoring," *in Proceedings of the International Conference on Unmanned Aircraft Systems*, Atlanta, pp. 221-234, 2013. DOI: 10.1109/ICUAS.2013.6564694

[13] Kumar B., Kumar A., and Pandey R., "MF-MSCNN: Multi-Feature based Multi-Scale Convolutional Neural Network for Image Dehazing via Input Transformation," *IETE Journal of Research*, vol. 71, no. 5, pp. 1527-1546, 2025. DOI: 10.1080/03772063.2025.2462789

[14] Li S., Yang X., Lin X., Zhang Y., and Wu J., "Real-Time Vehicle Detection from UAV Aerial Images

Based on Improved YOLOv5," *Sensors*, vol. 23, no. 12, pp. 1-18, 2023. DOI: 10.3390/s23125634

[15] Li W., Mao K., Zhang H., and Chai T., "Selection of Gabor Filters for Improved Texture Feature Extraction," *in Proceedings of the IEEE International Conference on Image Processing*, Hong Kong, pp. 361-364, 2010. https://doi.org/10.1109/ICIP.2010.5653278

[16] Lin H., Tu K., and Li C., "VAID: An Aerial Image Dataset for Vehicle Detection and Classification," *IEEE Access*, vol. 8, pp. 212209-212219, 2020. DOI: 10.1109/ACCESS.2020.3039798

[17] Mustafa W. and Abdul Kader M., "A Review of Histogram Equalization Techniques in Image Enhancement Application," *Journal of Physics: Conference Series*, vol. 1019, pp. 1-8, 2018. DOI: 10.1088/1742-6596/1019/1/012026

[18] Puertas E., Heras G., Andres J., and Soriano J., "Dataset: Roundabout Aerial Images for Vehicle Detection," *Data*, vol. 7, no. 4, pp. 1-11, 2022. DOI: 10.3390/data7040047

[19] Yurtkulu S., Sahin Y., and Unal G., "Semantic Segmentation with Extended DeepLabv3 Architecture," *in Proceedings of the 27th Signal Processing and Communications Applications Conference*, Sivas, pp. 1-4, 2019. DOI: 10.1109/SIU.2019.8806539

[20] Zhang L., "Enterprise Employee Work Behavior Recognition Method Based on Faster Region-Convolutional Neural Network," *The International Arab Journal of Information Technology*, vol. 22, no. 2, pp. 291-302, 2025. https://doi.org/10.34028/iajit/22/2/7

[21] Zhao Q. and Zhang L., "ECG Feature Extraction and Classification Using Wavelet Transform and Support Vector Machines," *in Proceedings of the International Conference on Neural Networks and Brain*, Beijing, pp. 1089-1092, 2005. https://doi.org/10.1109/ICNNB.2005.1614807

Nouf Abdullah Almujally received the Ph.D. in Computer Science from the University of Warwick, UK. She is currently an Assistant Professor in Computer Science at the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. Her research interests include Human-Computer Interaction (HCI), Artificial Intelligence (AI), Machine Learning, Deep Learning, and Computer Based Applications.

Ghulam Mujtaba received his M.S. degree in Computer Science from Air University, Islamabad. He is currently a Ph.D. candidate and a Research Assistant at the Intelligent Media Centre. His research focuses on Machine Learning, Deep Learning, Artificial Intelligence, and Human-Computer Interaction, with a particular emphasis on Camera and Sensor-based Gesture Recognition and Virtual Reality.

Shuaa S. Alharbi is a member in IEEE, she received the B.Sc. and M.Sc. degrees in computer science from Qassim University, Saudi Arabia, and the Ph.D. degree from Durham University, U.K. She is currently an Accomplished Researcher specializing in video and image analysis using deep learning. She is also an Assistant Professor with the Computer College, Qassim University. Her interdisciplinary research interests include machine learning and image processing, with a strong focus on the biology and medical domains. She is dedicated to employing deep neural networks to uncover patterns in image data and enhance diagnostic accuracy. Her innovative work includes developing advanced deep learning architectures and data-driven methodologies to improve the precision of medical image analysis. She is also exploring cutting-edge techniques for image processing and analysis across diverse applications.

Noif Alshammari is a Postdoc degree from Department of Cyber Security, College of Humanities, POSTECH University. His research interests include Machine Learning and Deep Learning.

Ahmad Jalal is currently a Professor from Department of Computer Science and Engineering, Air University, Pakistan. He received his Ph.D. degree in the Department of Biomedical Engineering at Kyung Hee University, Republic of Korea. Now, he was working as Postdoctoral Research fellowship at POSTECH. His research interest includes Multimedia Contents, Artificial Intelligence and Machine Learning.