

Article

VFD-YOLO: A Novel Method for Vehicle Detection from a Drone Perspective

Zongnan Liu ^{1,2}, Yong Xu ^{1,2}, Haoyi Xie ³ and Zepeng He ^{1,2,*} 

¹ School of Aerospace Engineering, Beijing Institute of Technology, Beijing 100081, China; lzn2007@163.com (Z.L.); xyxuyong@bit.edu.cn (Y.X.)

² Beijing Institute of Technology Chongqing Innovation Center, Chongqing 401100, China

³ Yuxi Industries Group Co., Ltd., Nanyang 200126, China; xhyxht002@163.com

* Correspondence: hezepeng0315@163.com

Abstract

Vehicle detection using drone-based imagery holds significant practical value in applications such as intelligent traffic management, urban situational awareness, and emergency response coordination. It is among the core technologies for achieving coordinated air-ground intelligent monitoring. However, when viewed from the high-altitude perspective of a drone, vehicle images suffer from issues such as complex backgrounds, blurriness, and low light, severely limiting the accuracy of model detection. Therefore, we propose a vehicle detection model for unmanned aerial vehicles based on YOLOv13 (VFD-YOLO). First, to address the issues of blurriness and low illumination in drone images, we designed a convolutional structure VFD-C3k2 specifically for vehicle feature extraction. It is a multi-scale feature-embedding structure based on the existing HLF structure. We decomposed the vehicle image features into high-frequency details such as edges and textures, as well as low-frequency structural information such as global contours. Through differentiated processing, we enhanced the image restoration and detail extraction capabilities, and improved the adaptability of the model to different types of images. Afterward, to address the strong background interference from vehicle targets in drone imagery, we employed the CASAB channel and spatial attention module. This module enhances the weights of key vehicle feature channels through channel attention, focuses on the target area of the vehicle through spatial attention, effectively suppresses background noise, and strengthens the model's ability to focus on and extract significant target features. The experimental results show that on the DroneVehicle dataset, the precision, recall rate, mAP@0.5 and mAP@0.5:0.95 of our proposed VFD-YOLO model reach 75.0%, 75.6%, 79%, and 54.7%, respectively, which represent improvements of 1.6%, 1.9%, 1.7%, and 2.4% compared with those of the baseline YOLOv13 model. In summary, the model we propose exhibits superior detection performance and can better meet the needs of practical scenarios such as intelligent transportation and urban surveillance.



Academic Editor: Bosko Rasuo

Received: 12 July 2026

Revised: 15 August 2026

Accepted: 18 August 2026

Published: 20 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: vehicle detection; YOLOv13; image blurring; VFD-C3k2; CASAB

1. Introduction

In recent years, the construction of smart transportation and urban three-dimensional safety management and control has been continuously promoted. Drones, with their advantages of maneuverability, flexibility, wide monitoring range, and low deployment cost, have become key equipment in areas such as traffic flow statistics, inspection of offending

vehicles, emergency rescue surveys, and security in smart cities [1,2]. In an unmanned aerial vehicle–ground cooperative perception system, vehicle target detection is the core technology that underpins the entire system. Through precise detection, key information such as vehicle positions, quantities, and movement states can be obtained in real time. This provides data support for intelligent traffic scheduling, emergency response, and road situation analysis. It is highly important for promoting digital and intelligent upgrades to traffic management [3,4]. However, traditional ground cameras have limitations in terms of field of view and coverage range, and thus cannot meet the requirements for large-scale dynamic monitoring. Traditional image processing methods based on manually designed features have drawbacks such as cumbersome feature extraction, weak generalization ability, and poor robustness. They have difficulty adapting to the complex aerial shooting scenarios of unmanned aerial vehicles [5]. Therefore, achieving fast, accurate, and robust vehicle target detection from a drone perspective has become a research hotspot in the fields of intelligent transportation and air–ground intelligent surveillance [6,7].

With its powerful ability to learn autonomous features, deep learning, as a cutting-edge technology in the field of artificial intelligence, has made target detection technology based on deep learning the current mainstream research direction [8,9]. According to the detection architecture, convolutional neural network (CNN) object detection models can be divided into two-stage and single-stage models [10]. The two-stage detection algorithm involves generating candidate regions first, followed by classification and position regression of targets within these regions. Typical models include R-CNN [11], Fast R-CNN [12], SPP-Net [13], etc. Although these algorithms achieve high detection accuracy, their complex workflow and slow inference speed make them unsuitable for meeting the real-time detection requirements of drones [14]. The single-stage detection algorithm eliminates the step of generating candidate regions and directly achieves target localization and classification through convolutional networks. It uses the dense anchor box strategy to complete end-to-end detection, featuring high speed and high degree of lightweighting. It is more suitable for real-time monitoring scenarios, and representative algorithms include the SSD series [15] and the YOLO series [16], etc. With the continuous iterative optimization of the YOLO series of models, their detection accuracy and speed have continued to improve. YOLOv13, as the latest version of this series, has achieved the optimal balance between detection accuracy and inference efficiency on multiple public datasets. It provides an ideal basic model for real-time vehicle detection from the perspective of unmanned aerial vehicles [17,18].

Owing to the excellent performance of the YOLO series of models, scholars both domestically and internationally have widely applied them to drone and vehicle detection tasks, and have conducted extensive research on their improvement [19]. Li et al. [20] addressed the issues of low detection efficiency and accuracy in drone vehicle detection, as well as the boundary discontinuity of rotating detectors. They proposed a drone vehicle detection method based on YOLOv5-OB, achieving high-precision drone vehicle detection with a detection accuracy of 98.5%. Ain et al. [21] reported that the performance of small target detection in unmanned aerial vehicle (UAV) remote sensing images was affected by background noise and poor data quality. By leveraging DCGAN data augmentation and transfer learning, they improved YOLOv8 and proposed the YDGAN-TL model. This model effectively enhanced the accuracy of small target detection on multiple UAV detection datasets. Alshehri et al. [22] addressed the issues of small target size and occlusion in vehicle detection using drone aerial photography. They constructed an end-to-end deep learning perception framework and employed YOLOv11 for vehicle detection, achieving excellent detection results.

The above methods have made certain progress in the detection of unmanned aerial vehicles (UAVs), but some shortcomings still exist. To address the poor performance of traditional detectors on long-distance tiny UAV targets, Liu et al. [23] proposed the GL-YOMO algorithm based on YOLOv5 with multi-scale feature fusion and multiframe motion detection. Tests on a self-built Fixed-Wing UAV dataset prove GL-YOMO boosts small-UAV detection accuracy and stability while retaining high efficiency. However, the framework relies only on spatial feature fusion and lacks joint channel-spatial attention to suppress complex background interference simultaneously. To tackle the large parameter volume and weak small-object capability of vanilla DETR for high-altitude UAV images, Liao et al. [24] proposed UAV-DETR with inverted residual blocks, cascaded linear attention and Wasserstein-distance-based loss. When evaluated on VisDrone2021, UAV-DETR cuts the model parameters by 44% and yields a 16.6% increase in accuracy over YOLOv8m. Nevertheless, this work does not design independent branches to separate high-frequency texture and low-frequency contour features; thus, it cannot effectively eliminate blurring and low-light degradation in aerial frames. Notably, high-altitude drones can easily cause image blurring, noise interference, uneven lighting and other degradation problems. There are also difficulties such as target similarity interference and chaotic backgrounds for vehicle targets. The existing models lack fine-grained processing of high- and low-frequency features of the images, have insufficient ability to jointly mine key features of channels and space, and are unable to simultaneously solve the problems of image quality degradation and complex scene feature extraction. Therefore, designing a detection model that adapts to the characteristics of UAV aerial photography and addresses issues such as poor image quality and strong background interference is crucial for improving the performance of UAV detection.

In this study, YOLOv13 is taken as the baseline model and addresses the issues of image degradation and background interference in vehicle detection from the perspective of unmanned aerial vehicles. A new unmanned aerial vehicle detection model VFD-YOLO is proposed. The model performs high-low frequency differentiated processing on features through VFD-C3k2, effectively improving image blurriness and noise problems, and enhancing the extraction of vehicle details. Moreover, the CASAB channel and spatial joint attention module are adopted, through channel weight calibration and spatial region focusing, to suppress complex background interference and highlight the key features of the target. On the basis of the drone vehicle detection dataset, comparative experiments and ablation experiments are conducted in this study to verify the effectiveness of the proposed model. The main contributions of this study are as follows:

- (1) The VFD-YOLO algorithm we proposed can be applied to the vehicle detection task from the perspective of unmanned aerial vehicles, which is in line with the actual application scenarios of intelligent transportation's air-ground collaborative monitoring. It has significant theoretical value and practical engineering significance.
- (2) To address the challenges of image degradation in drone aerial images and vehicle detection, we have designed a multi-scale feature-embedding structure VFD-C3k2 based on the existing HLFDA architecture, and integrated the CASAB joint attention module to optimize the model's feature extraction and anti-interference capabilities.
- (3) We preprocessed the DroneVehicle dataset, highlighting the data with blurred, low-light conditions, and strong background interference in vehicle images, and deleting the data unrelated to the research content.
- (4) We conducted sufficient comparison and ablation experiments to verify that the proposed model can effectively adapt to complex aerial shooting scenarios involving unmanned aerial vehicles. It has better detection accuracy in complex vehicle

backgrounds and degraded images, and better meets the requirements of practical engineering applications.

2. Experimental Data and Methods

2.1. Data Preprocessing

In this study, experiments were conducted using the DroneVehicle dataset [25], which is designed for drone target detection. The dataset covers various real aerial photography scenarios, such as urban roads and suburban road networks, as shown in Figure 1. The samples generally have problems such as motion blur, low-light imaging, and complex background interference. They also contain many irrelevant images and invalid samples. Direct training interferes with feature extraction and reduces model convergence and detection accuracy. Therefore, we carried out a systematic frame-by-frame preprocessing and screening process on the original data. First, all infrared images were discarded, and only single-channel visible light RGB aerial images were retained. Afterward, blank road images without vehicle annotations and nontraffic-related irrelevant scene images, such as wasteland and construction sites, were eliminated. High-difficulty degraded samples such as distant small vehicles, occluded vehicles, low-light conditions, motion blur, vegetation and building shadows were completely retained without any deletion. After the screening, the scenes were divided into isolated datasets according to the aerial photography flight path. The same continuous flight-shot images were uniformly classified into the training set or test set to prevent data leakage of the same scene. The statistical information of the filtered samples is shown in Table 1.

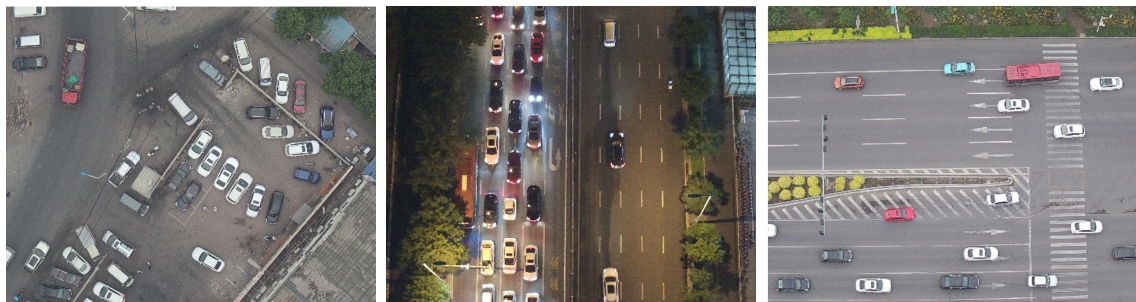


Figure 1. An example image displays data from different scenes.

Table 1. The number of images in the training and test datasets.

Dataset	Number of Images	Number of Labeled Fish Samples
Train	8442	106,115
Test	1466	23,205

The DroneVehicle dataset provides native annotations of five types of vehicle rotated bounding boxes. Considering that the core research objective of this paper is to verify the feature enhancement effect of the frequency-domain decomposition and attention module on degraded aerial images, vehicle angle regression-related optimizations are not carried out. In this study, all rotated annotations were uniformly standardized. The program reads the coordinates of the four vertices of each target rotated box and solves the corresponding minimum enclosing horizontal axis-aligned rectangle as the new annotation, which completely covers the vehicle entity and does not have the problem of target truncation. All training and testing processes are uniformly carried out using the converted horizontal annotations.

2.2. Proposed Vehicle Detection Method

To address issues such as image blurring, degradation under low-light conditions, and complex backgrounds in vehicle detection from a drone's perspective, this study proposes VFD-YOLO, a vehicle detection model tailored for drone aerial photography scenarios, using YOLOv13 as the baseline model. In this model, we design a multi-scale feature-embedding structure VFD-C3k2 on the basis of the existing HLF D architecture, specifically for vehicle feature extraction. This architecture employs an HLF D (high–low frequency decomposition) module to achieve differential enhancement of image features in the frequency domain. Furthermore, we incorporate the CASAB module into the head architecture to achieve frequency-domain differential enhancement of image features. This improves the detection performance of vehicle targets in complex aerial photography scenarios.

2.2.1. YOLOv13 Architecture

YOLOv13 is a real-time object detector based on hypergraph-enhanced adaptive visual perception. It overcomes the limitations of the traditional YOLO series, which are confined to local information aggregation and pairwise correlation modeling [17]. As shown in Figure 2, high-order correlation modeling is introduced into a real-time detection framework for the first time, achieving a better balance between accuracy, speed, and lightweight performance. Its core innovations include the HyperACE adaptive correlation enhancement mechanism, the FullPAD end-to-end aggregation and distribution paradigm, and deep separable lightweight modules. It efficiently performs global, cross-location, and cross-scale feature fusion and enhancement, making it better suited for complex aerial photography scenarios involving drones. The model achieves end-to-end information flow through FullPAD. The input layer adopts classic preprocessing strategies from the YOLO series, including mosaic enhancement, random scaling, mixup, and letterbox scaling, with a unified input size of 640×640 to enhance adaptability to multi-scale vehicle targets. The backbone network employs a lightweight DS-C3k2 module to replace traditional large-kernel convolutions, significantly reducing the number of parameters and computational load while retaining feature representation capabilities, and outputs three-level multi-scale feature maps (B_3 , B_4 , B_5).

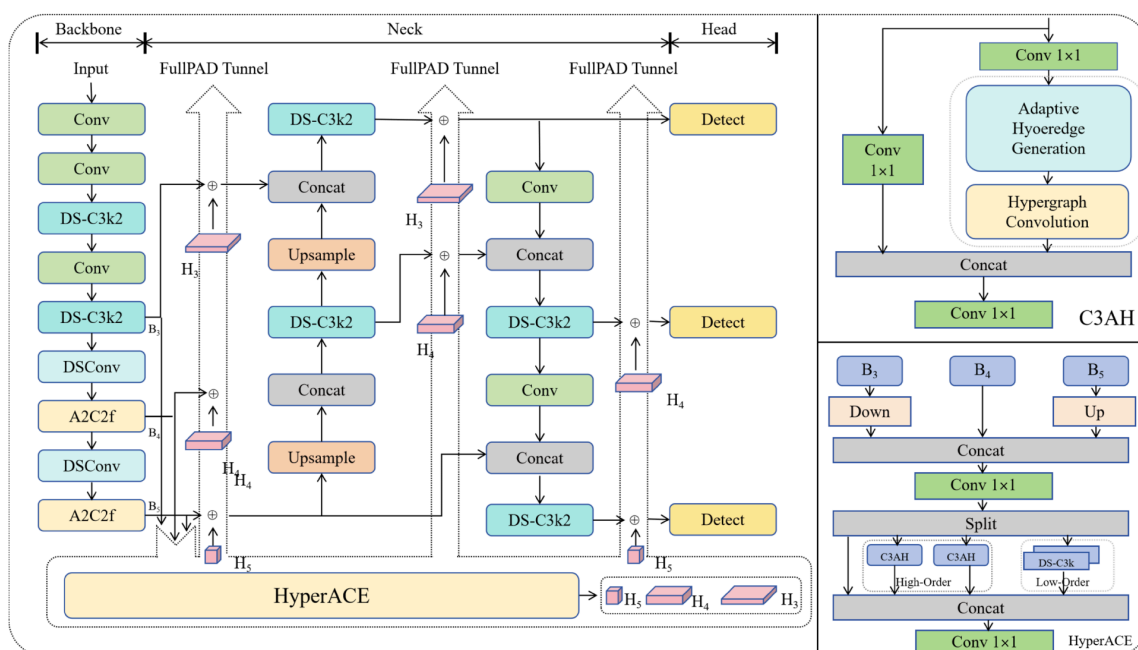


Figure 2. Schematic diagram of the YOLOv13 network structure.

The HyperACE hypergraph enhancement module takes these three-level features as input. It models global high-order semantic correlations through adaptive hyperedge generation and hypergraph convolution, and achieves feature complementarity via the high-order perception branch (C3AH) and low-order perception branch (DS-C3k) to output enhanced features. FullPAD distributes these enhanced features through three channels to the connections between the backbone and the neck, and between the neck and the detection head, achieving fine-grained information flow and representation synergy to improve gradient propagation. The neck network is based on an improved PAN-FPN architecture, fusing multi-level features and performing upsampling, downsampling, and cross-layer connections to enhance the feature representation of small objects and occluded vehicles. The detection head adopts a decoupled head architecture, separating the classification and regression tasks. It employs an anchor-free detection paradigm, combining distributed focus loss (DFL) with complete intersection-over-union (CIoU) to improve the accuracy of vehicle detection box regression and class identification.

The original YOLOv13 detection head is an anchor-free decoupling structure, only supporting horizontal bounding box coordinates and category regression, and no rotation angle prediction branch is designed. To ensure the single controllability of the ablation experiments and the horizontal comparison experiments variables, and to avoid additional interference caused by the addition of angle loss and hyperparameters, this study fully adopts the native detection head structure, and all the comparison models simultaneously use the converted horizontal annotations for training and evaluation.

2.2.2. Proposed VFD-YOLO Detection Model

In this study, we mainly improve the YOLOv13 model for the problem of vehicle detection in aerial drone imagery. The proposed model is named VFD-YOLO, as shown in Figure 3. The specific improvements are as follows:

- (1) A multi-scale feature-embedding structure VFD-C3k2 based on the existing HLFD architecture suitable for vehicle feature extraction was proposed, enhancing the high-frequency details such as vehicle edges and textures as well as the low-frequency information of the global structure.
- (2) In the head structure, the CASAB channel and the spatial joint attention module were integrated to precisely highlight the vehicle target features.

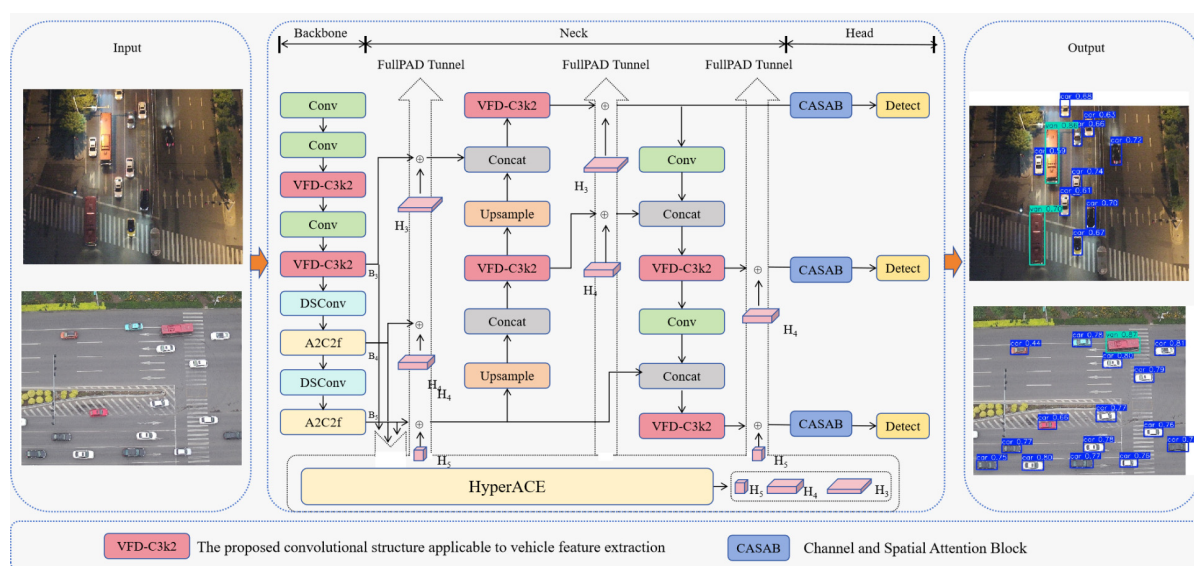


Figure 3. Network structure diagram of the proposed algorithm VFD-YOLO.

The Designed VFD-C3k2 Convolutional Structure

(1) Structural Analysis

The influence of factors such as motion blur, low light, and atmospheric scattering on the drone aerial photography images, lead to blurred vehicle edges, texture loss, and weakened structural information, etc. To further enhance the feature expression ability of the backbone network, on the basis of the original DSC3k2 module of YOLOv13, this study proposes a lightweight convolution structure VFD-C3k2 that integrates high and low frequency decomposition enhancement. As shown in Figure 4, this structure is based on the C2f skeleton of the original DSC3k2. The core improvement lies in the optimization of the feature extraction units within the branches, by integrating the high and low frequency decomposition enhancement module (HLFD). This achieves the separation enhancement and collaborative fusion of high-frequency detail features and low-frequency semantic features of the vehicle target. While ensuring the lightweight nature of the model and meeting the real-time detection requirements of drones, it effectively improves the feature extraction accuracy of vehicles. After the overall input features of VFD-C3k2 are compressed by a 1×1 convolution and undergo a split operation, they are divided into n parallel branches. Each branch adopts the VFD-C3k module instead of the original DS-bottleneck module. After internal feature fusion optimization, the features of each branch are fused through Concat across branches. Finally, through a 1×1 convolution adjustment, the output channels are strengthened, enhancing the ability to capture vehicle features.

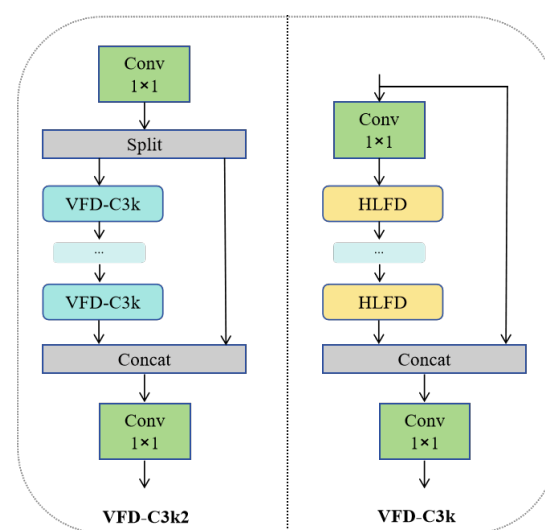


Figure 4. Network structure diagram of VFD-C3k2.

In the VFD-C3k2 structure, the HLFD module is the core for accurately extracting vehicle features. Its design aims to address the problem that traditional convolution modules have difficulty simultaneously considering both high-frequency details and low-frequency semantic information in vehicle detection. From the perspective of the drone, the license plate outline and body texture of the vehicle target, which are high-frequency details, are easily interfered by background noise. Low-frequency semantic features such as the overall vehicle outline are easily blurred because of scale changes and occlusion. It is difficult for a single convolution operation to achieve effective separation and enhancement of the two types of features. Therefore, the HLFD module is introduced to specifically handle the separation, independent enhancement, and collaborative fusion of high and low-frequency features, providing more discriminative feature support for the subsequent classification and positioning of vehicle targets.

The structure of the HLF module [26] is shown in Figure 5, which separates low-frequency and high-frequency features through 2×2 average pooling and difference operations. The low-frequency features are enhanced by UNet to alleviate blurring caused by scale variation and occlusion, while the high-frequency features are strengthened by the dense module to enhance the recognition of small targets. After being concatenated, the two features undergo 1×1 convolution for alignment, CALayer attention calibration, and then 3×3 convolution for optimized fusion. Finally, the residual connection is input into the features to ensure that they possess both detailed and semantic information.

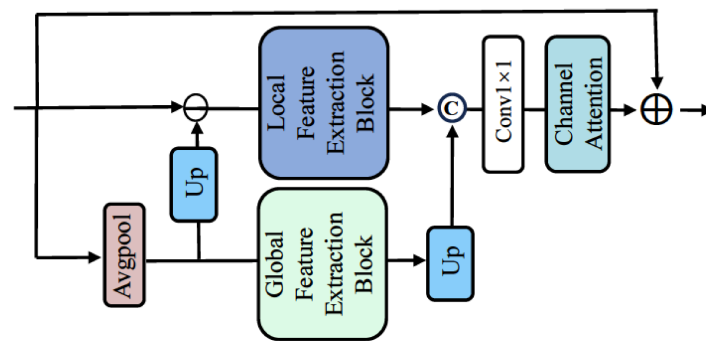


Figure 5. Structural diagram of the HLF.

The calculation process of the HLF is as follows:

$$F_{low} = \text{AvgPool}(F) \quad (1)$$

$$F' = \text{Upsample}(F_{low}) \quad (2)$$

$$F_{high} = F - F' \quad (3)$$

In the formula, F represents the input features, $\text{AvgPool}(\cdot)$ is the average pooling, which extracts the low-frequency structures such as the global outline of the vehicle and the road background, $\text{Upsample}(\cdot)$ is the bilinear upsampling, which ensures the alignment of the scales of high and low frequency features, F_{low} is the low-frequency features, and F_{high} is the high-frequency features.

Considering the differences between high-frequency features and low-frequency features, the HLF module is designed with a local feature extraction branch and a global feature extraction branch to handle them separately.

In the aspect of enhancing high-frequency local features, high-frequency features carry details such as vehicle edges, corners, and textures, and are crucial for restoring blurriness and noise. This branch employs small convolution kernels and a dense connection structure to focus on local details. The calculation formula is as follows:

$$F'_{high} = \text{LocalBlock}(F_{high}) \quad (4)$$

In the formula, $\text{LocalBlock}(\cdot)$ represents the local feature extraction block. The Local Feature Extraction Block processes high-frequency information such as vehicle edges and surface textures. Its structure contains three stacked residual dense units built with three by three depthwise convolution layers, paired with one by one projection convolution layers at the input and output terminals. LeakyReLU activation and batch normalization are adopted throughout the block. Small convolution kernels retain fine details of tiny blurred vehicles and suppress irrelevant background texture noise.

In terms of extracting low-frequency global features, low-frequency features carry global information such as the overall shape and scene structure of the vehicle. This branch

introduces the channel self-attention transformer and the multi-scale wavelet fusion module to capture long-range dependencies and avoid the loss of vehicle structure information caused by downsampling. The calculation formula is as follows:

$$F'_{low} = \text{GlobalBlock}(F_{low}) \quad (5)$$

In the formula, $\text{GlobalBlock}(\cdot)$ represents the global feature extraction block. The Global Feature Extraction Block focuses on low-frequency structural information covering integral vehicle outlines and overall road scene layout. The module starts with two by two average pooling for feature downsampling, followed by four head channel self attention and a three scale wavelet fusion unit. Bilinear upsampling is applied at the final stage to restore original feature resolution. This design captures long range spatial dependencies and alleviates vehicle contour blurring caused by high altitude drone photography.

Ultimately, the HLFD module combines the enhanced high-frequency and low-frequency features element-wise and outputs a high-quality feature map:

$$F_{out} = F'_{high} + F'_{low} \quad (6)$$

The HLFD module can effectively improve image degradation, providing clearer and more complete vehicle features for the subsequent HyperACE high-order correlation modeling. It significantly enhances the vehicle detection accuracy in blurry and low-light scenarios.

(2) Comparative Analysis with Existing Frequency Decomposition Approaches

Three typical frequency decomposition methods are compared with our VFD-C3k2. Wavelet transform decomposition incurs heavy computational overhead and enhances background noise indiscriminately, which prevents it from satisfying the real-time detection requirements for UAV platforms. Our lightweight pooling-difference HLFD lowers computational complexity and achieves separate enhancement for high-frequency texture and low-frequency contour features of vehicles. The original HLFD as a standalone serial block designed for general image restoration. Our upgraded VFD-C3k2 adopts a multi-branch parallel structure to extract multi-scale vehicle features and works together with CASAB to suppress background interference. Quantitative ablation results confirm steady performance gains across all evaluation metrics. Fourier transform filtering conducts global frequency conversion that destroys local spatial correlation and consumes substantial memory resources. Our spatial-domain frequency separation preserves the spatial context of vehicle targets and fits edge deployment on UAV devices.

Overview of the Channel and Spatial Attention Module

In the VFD-YOLO model we constructed, we integrated the CASAB (Channel and Spatial Attention Block) [27] at the key position of the model's head. Specifically, we inserted this module at the location where the refined features from each scale (P3/P4/P5) are aggregated through feature refinement and FullPAD_Tunnel, and before the final input to the detect detection head. This allows it to independently act on the three-level multi-scale feature maps of P3, P4, and P5. High-resolution P3 features are used for small vehicle target perception from the drone's perspective, medium-resolution P4 features are suitable for medium-sized vehicle target detection, and the low-resolution P5 features are responsible for extracting large vehicle targets. We utilize the CASAB module to achieve differentiation enhancement on the basis of the distribution characteristics of features of different resolutions without disrupting the original multi-scale detection structure, and

precisely suppress complex background interference and strengthen the response of key vehicle target features.

The CASAB module adopts a dual-branch parallel structure, synchronously completing the weight calibration in the channel dimension and the target focusing in the spatial dimension, as shown in Figure 6. Here, GMP represents global max pooling, GAP represents global average pooling, MP represents max pooling, MP_α represents mean pooling, MP_β represents min pooling, and SP represents sum pooling.

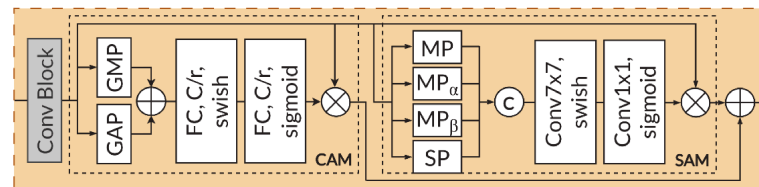


Figure 6. CASAB structure schematic diagram.

The specific calculation process is as follows:

(1) Channel Attention Branch (CAM)

The channel attention branch integrates the global statistical information and key activation information of the features through global average pooling (GAP) and global max pooling (GMP). It adaptively learns the importance weights of different channels, strengthens the feature channels related to the vehicle target, and suppresses the interference from redundant background channels. The calculation formula is as follows:

$$CAM(F) = F \cdot \sigma(FC(\text{Swish}(FC(GAP(F) + GMP(F))))) \quad (7)$$

In the formula, $GAP(\cdot)$ represents global average pooling, $GMP(\cdot)$ represents global maximum pooling, $FC(\cdot)$ represents the fully connected layer, $\text{Swish}(\cdot)$ represents the smooth activation function, and $\sigma(\cdot)$ represents the sigmoid activation function.

(2) Spatial Attention Module (SAM)

The spatial attention branch comprehensively captures the spatial distribution information of the feature map through mean pooling, max pooling, min pooling, and sum pooling. It then models the large-scale spatial context through 7×7 deep convolution to generate a spatial attention map, which focuses on the area where the vehicle target is located and weakens the background noise. The calculation formula is as follows:

$$F_{pool} = \text{Concat}(MP(F), MP_\alpha(F), MP_\beta(F), SP(F)) \quad (8)$$

$$SAM(F) = F \cdot \sigma\left(\text{Conv}_{1 \times 1}\left(\text{Swish}\left(\text{Conv}_{7 \times 7}(F_{pool})\right)\right)\right) \quad (9)$$

In the formula, $MP(\cdot)$ represents mean pooling, $MP_\alpha(\cdot)$ represents max pooling, $MP_\beta(\cdot)$ represents minimum pooling, $SP(\cdot)$ represents summation pooling, $\text{Conv}_{7 \times 7}(\cdot)$ represents 7×7 deep convolution, and $\text{Concat}(\cdot)$ represents the feature concatenation operation.

(3) Feature Fusion and Output

The CASAB module integrates the output features of channel attention and spatial attention element-wise addition, achieving joint enhancement in both channel and spatial dimensions. The final output is as follows:

$$CASAB(F) = CAM(F) + SAM(F) \quad (10)$$

By introducing the CASAB module at the end of the P3/P4/P5 multi-scale features, the VFD-YOLO model can enhance the features in a refined manner before the final detection. It not only retains the advantages of the original YOLOv13's multi-scale detection and high-order correlation modeling, but also specifically addresses the issues of cluttered backgrounds and weak target features in drone aerial photography scenarios. This approach significantly improves the detection accuracy and robustness of vehicles of different sizes.

3. Experiments and Results

3.1. Experimental Configuration

In the experiments, we used the images preprocessed from the DroneVehicle public unmanned aerial vehicle detection dataset for training and testing. The loss function of the proposed VFD-YOLO algorithm is consistent with that of YOLOv13, and it is composed of classification loss, localization loss and distribution focus loss as the optimization objective. Moreover, all our experiments were completed on Ubuntu operating system servers, and the code for building and training the improved vehicle detection algorithm was implemented on the basis of the PyTorch 2.4.1 deep learning framework. The initial learning rate was set to 1×10^{-2} , the weight decay coefficient was set to 5×10^{-4} , and the optimizer momentum was set to 0.937. In this experiment, two RTX 2080Ti graphics cards were used for calculations. Considering the memory usage limit during the training process, the batch size was set to 32, and the number of training iteration rounds was set to 100. The specific experimental hardware and software configuration environment is shown in Table 2.

Table 2. Experimental parameter configuration.

Parameter/Configuration	Result Value
GPU	NVIDIA GeForce RTX 2080Ti
epoch	100
momentum	0.937
batch size	32
the weight decay	0.0005
the initial learning rate	0.01

3.2. Evaluation Metrics

Evaluation indicators are the core basis for measuring the performance of target detection algorithms. To accurately assess the performance of the vehicle detection model from the perspective of unmanned aerial vehicles. In this study, the precision, recall rate, average precision, and mean average precision are selected as the core evaluation indicators. The value range of all the indicators is [0,1]; the higher the value is, the better the detection performance of the model.

(1) Precision (P)

The precision rate is the probability that the model predicts a sample as a vehicle target, among all samples that are actually vehicle targets. It reflects the accuracy of the model's detection results and can be used to effectively evaluate the model's ability to avoid misclassifying background, vegetation, and other non-vehicle targets as vehicles. Its expression is as follows:

$$precision = \frac{TP}{TP + FP} \quad (11)$$

Among them, TP represents the number of samples that were correctly identified as vehicles, while FP represents the number of background or other target samples that were wrongly detected as vehicles.

(2) Recall Rate (R)

The recall rate is the probability that all actual vehicle targets are successfully detected among all existing vehicle targets. It reflects the completeness of the model's identification of vehicle targets and can be used to effectively evaluate the model's ability to avoid missing small, occluded or distant vehicle targets. Its expression is as follows:

$$recall = \frac{TP}{TP + FN} \quad (12)$$

Among them, FN represents the number of actual vehicle targets that were not detected.

(3) Average Precision (AP)

The average precision represents the average classification performance of individual vehicle categories such as cars, trucks, vans, and buses. It is the area under the PR curve for all images in that category. The higher the AP value, the better the model's detection accuracy for that type of vehicle. Its formula is as follows:

$$AP = \int_0^1 P(R) dR \quad (13)$$

Among them, $P(R)$ represents the precision rate as a function of recall rate.

(4) Mean Average Precision (mAP)

The mean Average Precision (mAP) is a key metric for evaluating the performance of multi-class vehicle detection, which measures the comprehensive detection capability of the model across all vehicle classes. It is the arithmetic mean of the AP values for each class. Its expression is as follows:

$$mAP = \frac{\sum_{i=1}^C AP_i}{C} \quad (14)$$

Among them, AP_i represents the AP value for the i -th type of vehicle, and C represents the total number of vehicle categories in the dataset.

In addition, in this study, $mAP@0.5$ and $mAP@0.5:0.95$ are used as supplementary evaluation metrics to more comprehensively assess the model performance. $mAP@0.5$ represents the average mAP of all vehicle categories when the intersection-over-union (IoU) threshold is set to 0.5. It focuses on evaluating the model's loose positioning accuracy for vehicle targets. $mAP@0.5:0.95$ represents the average mAP of all vehicle categories when the IoU threshold is taken at values ranging from 0.5 to 0.95 with a step size of 0.05. It more strictly assesses the model's comprehensive detection capability under different positioning accuracy requirements. Among them, the intersection-over-union (IoU) is the core indicator for measuring the degree of overlap between the predicted box and the real box. It is defined as the ratio of the intersection area to the union area of the predicted box and the real box. Its expression is as follows:

$$IoU = \frac{Area(\text{Box}_{pred} \cap \text{Box}_{gt})}{Area(\text{Box}_{pred} \cup \text{Box}_{gt})} \quad (15)$$

Among them, Box_{pred} represents the vehicle bounding box predicted by the model, while Box_{gt} represents the actual annotated vehicle bounding box. The higher the IoU value, the more accurate the bounding box positioning is.

3.3. Ablation Study

To verify the effectiveness of the VFD-C3k2 convolution structure proposed in this study, the CASAB channel and spatial attention module are used in the vehicle detection task from the drone perspective. Using YOLOv13n as the baseline model, multiple sets of ablation comparison experiments were designed to evaluate the performance gains of individual modules and combined modules. The experimental results are shown in Table 3.

Table 3. Evaluation results of ablation experiments.

YOLOv13	VFD-C3k2	CASAB	Precision	Recall	mAP@0.5	mAP@0.5:0.95
✓			0.734	0.737	0.773	0.523
✓	✓		0.752	0.746	0.787	0.531
✓		✓	0.739	0.741	0.783	0.529
✓	✓	✓	0.75	0.756	0.79	0.547

Note: The bolded items in the table indicate the best effects.

3.3.1. Performance of the Baseline Model

The YOLOv13n benchmark model achieved a basic accuracy of 0.734, a recall rate of 0.737, a mAP50 of 0.773, and a mAP50-95 of 0.523 on the DroneVehicle dataset. This demonstrated its excellent adaptability in the drone vehicle detection task and provided a reference benchmark for subsequent module improvements.

3.3.2. Performance Analysis of a Single Module

Validation of the effectiveness of the VFD-C3k2 module. Compared with the baseline model, after the VFD-C3k2 module was designed, the model accuracy increased to 0.752, the recall rate increased to 0.746, and the mAP@0.5 and mAP@0.5:0.95 respectively increased to 0.787 and 0.531, respectively. The results show that the VFD-C3k2 module effectively repairs degradation problems such as blurriness and noise in drone aerial images by integrating the high–low frequency decomposition mechanism of HLFDD. It enhances the edge and texture detail features of vehicles. It improves the recall ability and comprehensive detection accuracy of the model for vehicle targets. In particular, in the detection of small target vehicles in low-light and motion-blurred scenes, it performs better.

Validation of the effectiveness of the CASAB module. Compared with the baseline model, after the CASAB module was introduced, the model accuracy slightly increased to 0.739, the recall rate increased to 0.741, and the mAP@0.5 and mAP@0.5:0.95 respectively increased to 0.783 and 0.529, respectively. The results show that the CASAB module precisely suppresses the interference of complex backgrounds such as vegetation, buildings, and shadows through the channel and spatial joint attention mechanism, focusing on the vehicle target area. Although it slightly decreases the detection accuracy, it effectively improves the recognition integrity of the model for the target. The detection robustness in scenes with cluttered backgrounds and occluded targets is enhanced.

3.3.3. Performance Analysis of the Combined Modules

After the VFD-C3k2 module was combined with the CASAB module, the model performance improved relatively significantly. Compared with the baseline model, the accuracy increased to 0.75, the recall rate increased to 0.756, the mAP@0.5 increased to 0.790, and the mAP@0.5:0.95 increased to 0.547, which was the best performance among all the experimental groups.

Compared with a single module, the combined module achieved greater gains in mAP@0.5 and mAP@0.5:0.95. These findings indicate that the VFD-C3k2 module and the CASAB module have good complementarity. The former optimizes feature input from

the image quality perspective, whereas the latter strengthens target expression from the feature focus perspective. The synergy of the two works together, retaining the efficient detection capability of YOLOv13n while specifically addressing the image degradation and background interference problems in the drone aerial photography scenario. This has comprehensively improved the accuracy and robustness of vehicle detection.

The ablation experiment results fully verified the effectiveness of the two core innovations of this study. The VFD-C3k2 module can effectively improve image degradation and enhance detailed features. The CASAB module can precisely suppress background interference and focus on the target area. After combining them, the model achieved better comprehensive performance in the drone perspective vehicle detection task, proving the rationality and superiority of the VFD-YOLO model.

In addition, we further supplement a control group equipped with the original HLFD module for fair comparison. As shown in Table 4, the standalone original HLFD results in only a limited improvement in performance over the baseline, and all the indicators are inferior to those of our proposed VFD-C3k2. These findings demonstrate that our multi-branch improved frequency decomposition pipeline is more suitable for blurry and low-light UAV detection scenarios.

Table 4. Compared with the frequency decomposition method.

Model Configuration	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv13 + Original HLFD	0.741	0.736	0.775	0.526
YOLOv13 + VFD-C3k2	0.752	0.746	0.787	0.531

To verify the real-time detection performance of unmanned aerial vehicles, we conducted relevant performance tests; the results are shown in Table 5. The detection speed of the proposed algorithm VFD-YOLO reached 168 frames per second, which can achieve real-time detection.

Table 5. Real-time performance indicators of the model.

Model	Parameters	GFLOPs	FPS
YOLOv13 (Baseline)	2,448,870	6.2	303
VFD-YOLO (Ours)	5,584,420	12.6	168

3.3.4. Visualization Results of the Ablation Experiment

To more intuitively verify the improvement effect of the VFD-YOLO model proposed in this study, we selected typical drone aerial photography scenarios such as urban roads under normal daytime lighting, parking lots under low nighttime lighting, and expressway sections under low nighttime lighting. The detection results of the original images, the YOLOv13 benchmark model, and the VFD-YOLO model were visually compared. The results are shown in Figure 7. We have marked the missed detection targets generated during the model's detection process with red boxes. This is done to provide a more intuitive assessment of the model's performance.

In the daytime normal lighting urban road scenario, the original image has a dense distribution of vehicles and clear features. The YOLOv13 model can complete the detection of most vehicles, but some vehicles are missed. However, the VFD-YOLO model not only completely detects all vehicle targets but also more accurately detects box positioning. The detection confidence of various vehicle categories has significantly improved, demonstrating better detection completeness and confidence performance.

In the nighttime low-light parking lot scenario, the original image has problems such as insufficient lighting and messy backgrounds. The YOLOv13 model has obvious false

detections, and the confidence of some target detections is relatively low. The VFD-YOLO model effectively suppresses the background light interference and successfully detects all small vehicle targets. Moreover, it improves the detection confidence of each target and verifies the degradation repair ability of the VFD-C3k2 module for low-light images and the suppression effect of the CASAB module on background interference.

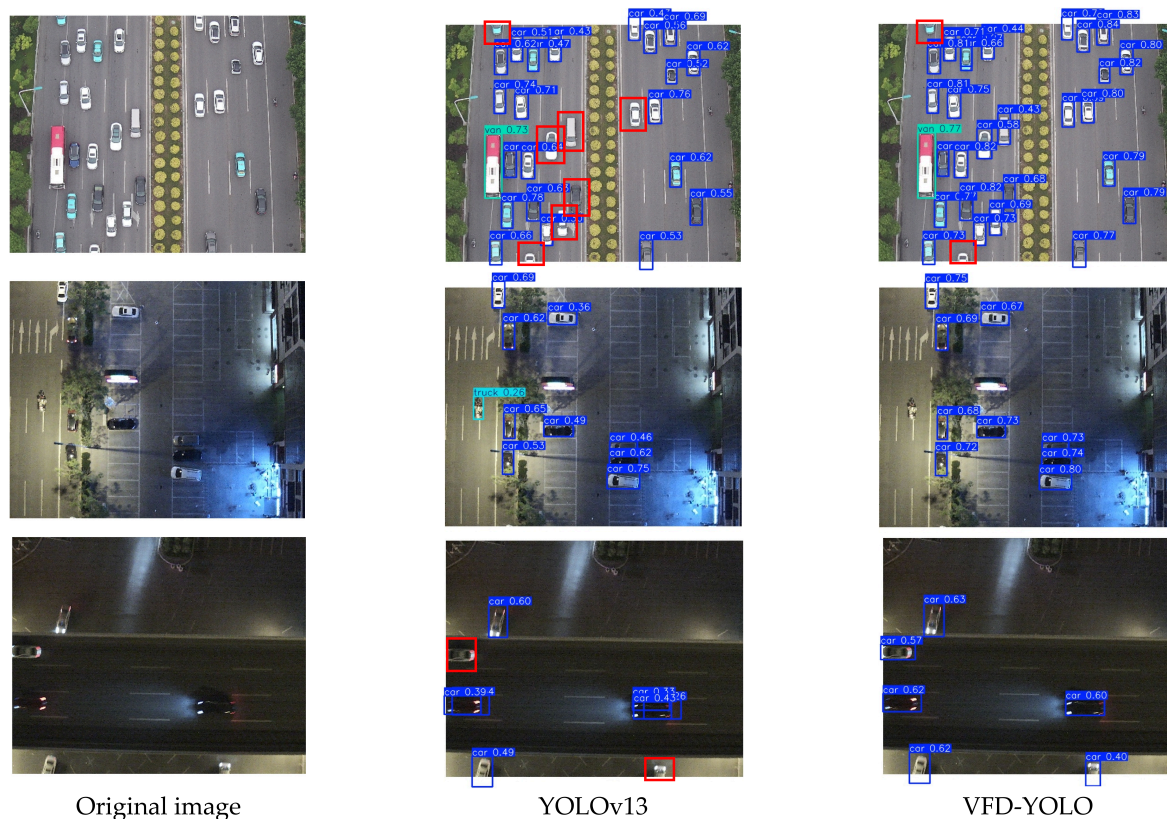


Figure 7. Visualization results of ablation experiment (Note: The model only outputs horizontally aligned detection boxes. The original dataset's native rotation-oriented annotations have been uniformly converted into the smallest enclosing horizontal boxes for model training and visualization plotting. The red boxes indicate the vehicles that were not inspected properly).

In the nighttime low-light expressway section scenario, the original image has problems such as low lighting and motion blur. The YOLOv13 model has obvious false detection situations, it misidentifies vehicle shadows as vehicles, and there is a slight missed detections occur. The VFD-YOLO model significantly improves the detection confidence of the targets, and accurately detects all motion-blurred vehicles, with precise positioning. This fully demonstrates the advantages of the improved model in scenarios with image blur, similar backgrounds, etc.

In conclusion, the visual results directly verify the superiority of the VFD-YOLO model. In different lighting conditions and different complexity drone aerial photography scenarios, compared with the YOLOv13 benchmark model, VFD-YOLO can not only improve the detection confidence and positioning accuracy but also effectively reduce missed detections and false detections. Specifically, in complex scenarios such as low lighting, image blur, and background interference, the collaborative effect of the improvement module is fully demonstrated. This further proves the effectiveness of the innovation points of this study.

3.4. Comparison with Other Advanced Detection Algorithms

To verify the superiority of the VFD-YOLO model proposed in this study in the vehicle detection task from the perspective of unmanned aerial vehicles, we selected the current

mainstream lightweight object detection algorithms for comparison, including the YOLO series of YOLOv5n [28], YOLOv8n [29], YOLOv10n [30], YOLOv11n [31], YOLOv12n [32], YOLO26 [33], YOLOv13n [17], and the transformer-based detector RT-DETR [34]. All the models were trained on the preprocessed DroneVehicle dataset with the same configuration. Specifically, the number of epochs was set to 100, the batch size was 32, and the optimizer and learning rate strategies were consistent. The comparison results are shown in Table 6.

Table 6. Comparison of target detection results.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5	0.682	0.725	0.741	0.506
YOLOv8	0.692	0.728	0.743	0.511
YOLOv10	0.746	0.703	0.742	0.514
YOLOv11	0.738	0.741	0.772	0.529
YOLOv12	0.713	0.729	0.763	0.522
YOLOv26	0.749	0.722	0.765	0.523
RT-DETR	0.769	0.753	0.751	0.502
YOLOv13	0.734	0.737	0.773	0.523
VFD-YOLO	0.75	0.756	0.79	0.547

Note: The bolded items in the table indicate the best effects.

It can be observed from the experimental results that the VFD-YOLO model proposed in this study achieved the best performance among all the comparison models in terms of core detection metrics such as the recall rate, mAP50, and mAP50-95. As shown in Figure 8, in terms of accuracy metrics, it is only slightly inferior to RT-DETR, but the overall detection performance is more balanced and superior. Specifically, the precision of VFD-YOLO reached 0.750, which was significantly greater than those of YOLOv5n (0.682), YOLOv8n (0.692), YOLOv12n (0.713), and the benchmark model YOLOv13n (0.734), which were only slightly lower than those of RT-DETR (0.769) and YOLOv26n (0.749). This finding indicates that the model performs exceptionally well in suppressing background false detections and ensuring detection accuracy. The recall rate reached 0.756, the highest among all the models. It improved by 0.019 compared with YOLOv13n, by 0.003 compared with RT-DETR, and by 0.031 and 0.028, respectively, compared with earlier YOLOv5n and YOLOv8n. This fully demonstrates that the model can effectively reduce the missed detection of small, occluded, and distant vehicle targets and has stronger integrity in target recognition. The mAP50 and mAP50-95 reached 0.790 and 0.547, respectively, which were 0.017 and 0.024 greater than those of YOLOv13n and 0.039 and 0.045 greater than those of RT-DETR. Under both loose and strict positioning accuracy requirements, the model has significant advantages in terms of comprehensive detection accuracy and bounding box positioning accuracy for multiple vehicle categories. It outperforms all the comparison YOLO series lightweight versions and the transformer detector RT-DETR. The results of the comparative experiments fully verified the effectiveness and synergy of the core innovation of this research. The VFD-C3k2 convolution structure, which integrates the high-low frequency decomposition mechanism of the HLFD, effectively improved the blurring and noise problems of unmanned aerial vehicle (UAV) aerial images, and enhanced the extraction of vehicle details. The CASAB channel and spatial attention module introduced before the detection head precisely suppressed the interference from complex backgrounds, focusing on the vehicle target area. The combination of these two elements enables VFD-YOLO to retain the efficient detection capability of YOLOv13 while significantly improving the robustness and detection performance of the model in complex aerial shooting scenarios. It is more suitable for the actual needs of vehicle detection from the perspective of the UAV.

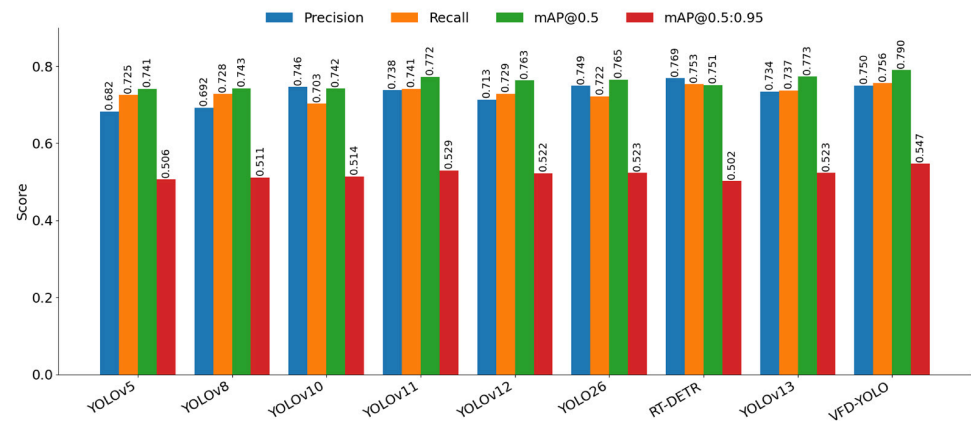


Figure 8. Bar chart comparison of target detection results.

The VFD-YOLO model proposed in this study demonstrates superior detection performance in various drone aerial photography scenarios, as shown in Figure 9. In terms of the missed detection rate, VFD-YOLO can completely detect all vehicle targets on dense urban roads during the day, low-light parking lots at night, and blurry highway sections at night. The missed detection rate is significantly lower than that of the comparison models, such as YOLOv5, YOLOv8, and YOLOv13, effectively avoiding the omission of small target vehicles and moving blurry vehicles in nighttime scenes. In terms of the false detection rate, VFD-YOLO has almost no background false detections. The CASAB attention module precisely suppresses complex background interference, such as vegetation and lights, whereas models such as YOLOv5 and DETR are prone to misclassifying non-vehicle interference objects as vehicles in nighttime scenes. In terms of detection confidence, the detection confidence of VFD-YOLO for vehicles such as cars and vans is higher than that of the comparison models, especially in low-light and blurry image scenarios, where the confidence of the van category in daytime urban roads increases from 0.48 in YOLOv5 and 0.73 in YOLOv13 to 0.77, and the confidence of the white vehicle in the lower left corner of the nighttime highway image increases from 0.49 in YOLOv13 to 0.62. This fully validates the image degradation repair ability and detail feature enhancement effect of the VFD-DSC3k2 module.

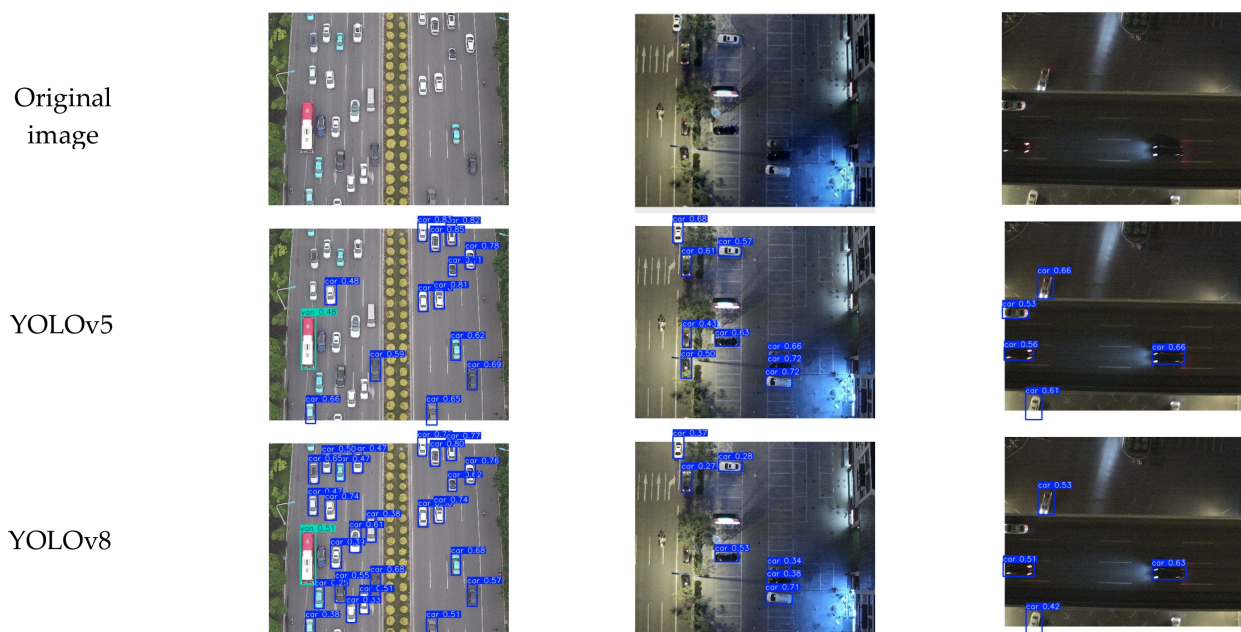


Figure 9. Cont.

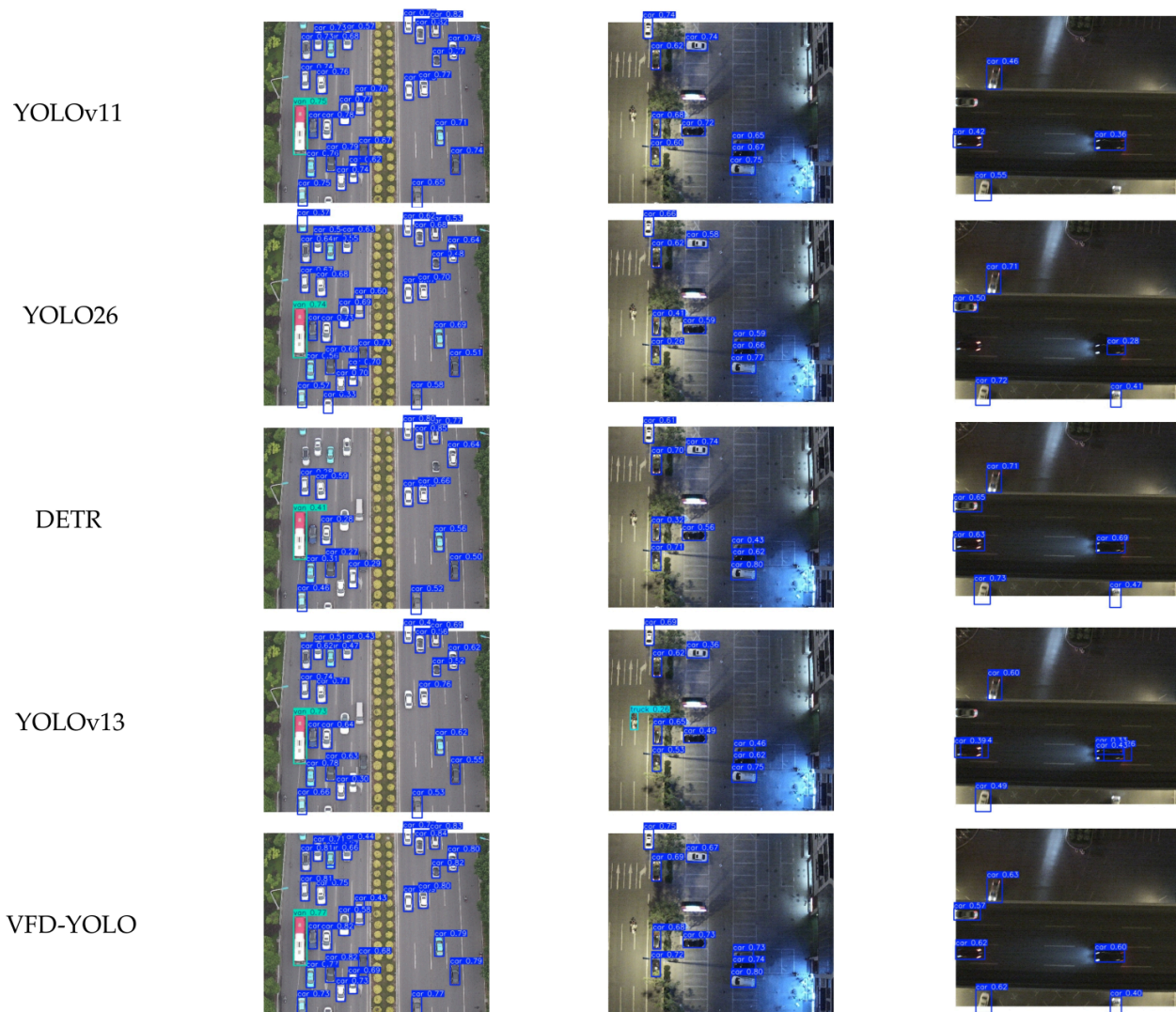


Figure 9. The performance of different algorithms in vehicle detection. (Note: The model only outputs horizontally aligned detection boxes. The original dataset's native rotation-oriented annotations have been uniformly converted into the smallest enclosing horizontal boxes for model training and visualization plotting).

4. Conclusions

From the perspective of unmanned aerial vehicles, vehicle detection is a core technology in intelligent transportation and urban security. However, problems such as image blurring, noise interference, and complex backgrounds in aerial photography scenarios severely restrict detection performance. To address these issues, this study uses YOLOv13 as the benchmark and proposes the VFD-YOLO vehicle detection model. First, a VFD-C3k2 convolution structure integrating the HLFM mechanism was designed. The image features were explicitly decomposed into high-frequency detail components and low-frequency structural components, and differentiated enhancement was implemented. This effectively repaired the degradation problem of unmanned aerial vehicle images and enhanced the ability to extract key details such as vehicle edges and textures, restoring image degradation and enhancing vehicle detail features. Second, the CASAB channel and spatial joint attention module were embedded in the detection head through channel dimension weight calibration and spatial dimension target area focusing, precisely suppressing background interference and focusing on the target area. Experiments on the DroneVehicle dataset reveal that compared with the YOLOv13n benchmark model, VFD-YOLO has a 1.9% in-

crease in the recall rate, a 1.7% increase in the mAP50, and a 2.4% increase in the mAP50-95. The core indicators are superior to those of current mainstream lightweight detection algorithms, demonstrating better accuracy and robustness in blurry, low-light, and complex background scenarios. In the future, the lightweight design of the model will be further optimized, and multimodal data fusion methods will be explored to better adapt to real-time detection on unmanned aerial vehicle ends and multimodal application requirements.

Author Contributions: Conceptualization, Methodology and Writing—Original Draft Preparation, Z.L.; Validation and Formal Analysis, Y.X.; Investigation and Supervision, H.X.; Writing—Reviewing, Editing and Methodology, Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was Supported by Jiangsu Province Key Laboratory of Aerospace Power System (No. CEPE2025003).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. We also thank Jiangsu Province Key Laboratory of Aerospace Power System for providing technical support in data collection and analysis.

Conflicts of Interest: Author Haoyi Xie was employed by the company “Yuxi Industries Group Co., Ltd.” The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ma, Z.; Wang, C.; Chen, C.; Chen, J.; Zheng, G. DDF-YOLO: A Small Target Detection Model Using Multi-Scale Dynamic Feature Fusion for UAV Aerial Photography. *Aerospace* **2025**, *12*, 920. [\[CrossRef\]](#)
2. Al-Iqubaydhi, N.; Alenezi, A.; Alanazi, T.; Senyor, A.; Alanezi, N.; Alotaibi, B.; Alotaibi, M.; Razaque, A.; Hariri, S. Deep learning for unmanned aerial vehicles detection: A review. *Comput. Sci. Rev.* **2024**, *51*, 100614. [\[CrossRef\]](#)
3. Kevin, P.; Kumar, C.S.; Naveen Raj, R.; Vani, V. Lightweight monocular vision pipeline for real-time safe landing zone detection in unmanned aerial vehicles. *Eng. Appl. Artif. Intell.* **2026**, *174*, 114561. [\[CrossRef\]](#)
4. Song, G.; Du, H.; Zhang, X.; Bao, F.; Zhang, Y. Small object detection in unmanned aerial vehicle images using multi-scale hybrid attention. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107455. [\[CrossRef\]](#)
5. Al-Kaff, A.; Martin, D.; Garcia, F.; de la Escalera, A.; Armingol, J.M. Survey of computer vision algorithms and applications for unmanned aerial vehicles. *Expert Syst. Appl.* **2018**, *92*, 447–463. [\[CrossRef\]](#)
6. Hu, S.; Xing, R.; Han, L.; Liu, T. EBAD-YOLO: Efficient bidirectional adaptive dense network for UAV small-object detection. *J. Real-Time Image Process.* **2026**, *23*, 40. [\[CrossRef\]](#)
7. Bakirci, M. Enhancing vehicle detection in intelligent transportation systems via autonomous UAV platform and YOLOv8 integration. *Appl. Soft Comput.* **2024**, *164*, 112015. [\[CrossRef\]](#)
8. Qiu, J.; Yu, F.; Xu, F.; Chen, X.; Wang, J. A Relative Attitude Detection Method for Unmanned Aerial Vehicles Based on You Only Look Once Framework. *Aerospace* **2025**, *12*, 191. [\[CrossRef\]](#)
9. Saiveena, K.; Praveen, P. Edge intelligence-enabled vision transformer with YOLOv5 for accurate vehicle detection and image segmentation using adaptive multiscale deep learning mechanism. *Comput. Electr. Eng.* **2026**, *134*, 111086. [\[CrossRef\]](#)
10. Wu, X.; Sahoo, D.; Hoi, C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [\[CrossRef\]](#)
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2014, Columbus, OH, USA, 23–28 June 2014*; IEEE: New York, NY, USA, 2014; pp. 580–587. [\[CrossRef\]](#)
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#) [\[PubMed\]](#)
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and Tracking Meet Drones Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [\[CrossRef\]](#) [\[PubMed\]](#)

15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV) 2016, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Germany, 2016; pp. 21–37. [\[CrossRef\]](#)
16. Chen, C.; Zheng, Z.; Xu, T.; Guo, S.; Feng, S.; Yao, W.; Lan, Y. YOLO-Based UAV Technology: A Review of the Research and Its Applications. *Drones* **2023**, *7*, 190. [\[CrossRef\]](#)
17. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733v2. [\[CrossRef\]](#)
18. Wang, D. Multi-scale and precise real-time detection method for weeds on sidewalks. *Eng. Appl. Artif. Intell.* **2026**, *170*, 114183. [\[CrossRef\]](#)
19. Tong, J.; Cao, H.; Lu, X.; Zhang, X.; Gao, B. DMR-YOLO: A Lightweight Visual Inspection Method for Surface Defect Detection of Aero-Engine Components. *Aerospace* **2026**, *13*, 360. [\[CrossRef\]](#)
20. Li, X.; Wu, J. Developing a More Reliable Framework for Extracting Traffic Data From a UAV Video. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 12272–12283. [\[CrossRef\]](#)
21. Ain, Q.; Xia, K.; Fatima, T. YDGAN-TL: Enhanced YOLOv8 With DCGAN and Transfer Learning for UAV Small Object Detection in Remote Sensing. *IEEE Aerosp. Electron. Syst. Mag.* **2025**, *40*, 20–37. [\[CrossRef\]](#)
22. Alshehri, M.; Xue, T.; Mujtaba, G.; AlQahtani, Y.; Almujaally, N.A.; Jalal, A.; Liu, H. Integrated neural network framework for multi-object detection and recognition using UAV imagery. *Front. Neurobotics* **2025**, *19*, 1643011. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Liu, J.; Plategher, L.; Roura, E.; de Souza Junior, C.; He, S. Real-Time Detection for Small UAVs: Combining YOLO and Multiframe Motion Analysis. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 13419–13433. [\[CrossRef\]](#)
24. Liao, N.; Zhang, Y.; Yu, Z.; Huang, J.; Zhu, M.; Peng, B. UAV-DETR: Few-Parameter DETR for Small Object Detection in High-Altitude UAV Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 2575–2587. [\[CrossRef\]](#)
25. Zhu, P.; Sun, Y.; Wen, L.; Feng, Y.; Hu, Q. Drone Based RGBT Vehicle Detection and Counting: A Challenge. *arXiv* **2020**, arXiv:2003.02437. [\[CrossRef\]](#)
26. Chen, G.; Dai, K.; Yang, K.; Hu, T.; Chen, X.; Yang, Y.; Dong, W.; Wu, P.; Zhang, Y.; Yan, Q. Bracketing Image Restoration and Enhancement with High- Low Frequency Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024, Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 6097–6107. [\[CrossRef\]](#)
27. Wazir, S.; Kim, D. Rethinking Decoder Design: Improving Biomarker Segmentation Using Depth-to-Space Restoration and Residual Linear Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025, Nashville, TN, USA, 11–15 June 2025*; IEEE: New York, NY, USA, 2025; pp. 30861–30871. [\[CrossRef\]](#)
28. Glenn, J. YOLOv5 by Ultralytics. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 15 June 2026).
29. Glenn, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 15 June 2026).
30. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2024; pp. 107984–108011.
31. Glenn, J. Ultralytics YOLO11. 2024. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 16 June 2026).
32. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
33. Glenn, J. Ultralytics YOLO26. 2026. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 16 June 2026).
34. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024, Seattle, WA, USA, 17–21 June 2024*; IEEE: New York, NY, USA, 2024; pp. 16965–16974. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.