Mathematics

AIMS Mathematics, 11(4): 12043–12063.
DOI: 10.3934/math.2026494
Received: 31 January 2026
Revised: 07 March 2026
Accepted: 17 March 2026
Published: 29 April 2026

https://www.aimspress.com/journal/Math

Research article

DV-YOLO: a deep learning framework for small-object detection in UAV-based remote sensing imagery with applications to smart logistics

Ahmed A. Alsheikhy[1], Mohammad Barr[2], Sahbi Boubaker[1,*] and Yahia Said[3,*]

[1] Department of Computer and Network Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 21959, Saudi Arabia
[2] Department of Electrical Engineering, College of Engineering, Northern Border University, Arar 91431, Saudi Arabia
[3] Center for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia

* **Correspondence:** Email: sboubaker@uj.edu.sa, yahia.said@nbu.edu.sa.

Abstract: Unmanned aerial vehicles (UAVs) are being increasingly adopted as flexible remote sensing platforms for smart logistics applications, including warehouse inventory, last-mile delivery supervision, traffic flow analysis, port operations, and infrastructure inspection. Despite their advantages, reliable object detection in UAV-based remote sensing imagery remains challenging due to small object sizes, dense object distributions, arbitrary orientations, and complex backgrounds commonly encountered in logistics environments. Although recent YOLO-based detectors have shown promising performance, their effectiveness is often limited in high-resolution aerial scenes and under practical computational constraints imposed by UAV platforms. To address these challenges, this paper proposes DV-YOLO, an enhanced deep learning framework tailored for object detection in UAV-based remote sensing imagery for logistics-oriented applications. The proposed model extends YOLOv9 through a deeper and wider backbone architecture coupled with optimized feature fusion strategies that jointly exploit spatial and semantic representations. A novel cross-path fusion network at deep feature map (CPFNDFM) is introduced to improve the detection of small and densely distributed logistics-related objects such as vehicles, containers, and infrastructure elements. In addition, a lightweight connection aggregation (CA) module, inspired by VoVNet and ShuffleNetV2, is integrated to enhance feature reuse while maintaining computational efficiency suitable for real-time UAV deployment. Furthermore, a challenging benchmark dataset, termed harder vision drone, is constructed by combining and refining samples from VisDrone and DOTA to better reflect real-world UAV remote

sensing scenarios in logistics environments. Extensive experimental evaluations conducted on VisDrone 2021, DOTA v2, and the proposed dataset demonstrate that DV-YOLO consistently outperforms state-of-the-art detectors, achieving up to 3.5% improvement in mean average precision (mAP) compared with YOLOv9. These results highlight the potential of the proposed framework to support robust, accurate, and efficient aerial perception for smart logistics and UAV-based remote sensing applications.

Keywords: UAV remote sensing; smart logistics; aerial small-object detection; deep learning; YOLO-based models; logistics monitoring; drone-based perception
Mathematics Subject Classification: 68T07, 68U10

1. Introduction

Unmanned aerial vehicles (UAVs), commonly referred to as drones, are increasingly deployed in logistics-oriented applications such as last-mile delivery supervision, warehouse and port monitoring, traffic flow management, and infrastructure inspection. In these scenarios, UAVs function as agile remote sensing platforms capable of acquiring high-resolution aerial imagery over dynamic and complex logistics environments. While the acquisition of aerial data is a fundamental step, the true value of UAV-based logistics systems lies in the effective interpretation and analysis of remote sensing imagery. However, logistics environments present significant challenges for aerial image analysis, including small and densely distributed objects, occlusions, scale variations, and cluttered backgrounds. Addressing these challenges is essential to transform raw UAV imagery into actionable information that supports intelligent logistics monitoring and decision-making. Owing to the importance of businesses built around UAVs, artificial intelligence (AI) is taking place as an enabler by opening new avenues for growth and development. Therefore, more precise target locations from high altitudes, reduced labor cost of labeling and filtering, and a higher degree of aerial image processing become possible with a combination of UAVs and computer vision algorithms integrated into expert systems [1]. Furthermore, UAVs equipped with an effective object detection algorithm have several potential applications, including the inspection of power infrastructure via reliable communication between drones that are capturing images of the power lines and base stations, aiming to process the received information [2]. In addition, wildfires in forests are a natural threat that becomes more and more frequent due to several unpredictable sources. To contribute to early detection and mitigate the effects of wildfires, drone imagery combined with internet of things (IoT) devices embedded on drones is among the state-of-the-art applications of UAVs [3]. Because of the device's power consumption and the larger ground object range after adjusting the UAV flight level, it is still a challenging task to deploy object detection algorithms on UAV platforms while maintaining good multi-scale detection accuracy and real-time performance.

As drones are nowadays widely used in logistics and shipment delivery, the development of reliable vision systems for these drones may address the critical need for efficient navigation and object recognition capabilities in autonomous UAVs when they are accurately locating delivery places or even the persons to whom they may deliver. As the demand for unmanned aerial delivery services continues to rise, particularly in sectors such as e-commerce, logistics, and medical emergency response, there is a pressing need for advanced vision systems that can enable drones to navigate safely, identify obstacles, and locate delivery destinations accurately. Traditional methods of drone navigation

often rely on the global positioning system and pre-programmed multi-pathway flights, which can be limited in accuracy and effectiveness, especially in urban environments with dense buildings or challenging weather conditions. Vision systems offer a more robust solution by allowing drones to perceive and interpret their surroundings in real-time, enabling them to adapt to dynamic environments and navigate with greater precision. Research in the field of vision systems for logistics services and delivery drones aims to overcome the technical challenges associated with autonomous navigation and enable the widespread adoption of drone delivery services across various industries. By enhancing drones' perception and decision-making capabilities, vision systems play a crucial role in unlocking the full potential of unmanned aerial delivery as a safe, efficient, and sustainable transportation solution.

Due to their ability to abstract from input data attributes, several variants of deep learning algorithms [4] using nonlinear models have found extensive usage in object detection. With its robust representation and learning capabilities, DL algorithms can automatically find the characteristics required for detection or classification tasks. A residual network, ResNet [5], with no impact on error while expanding the network's depth, has been suggested via a cross-layer link, which increased the network's performance. Using the proposal box extraction approach, the R-CNN algorithm developed by Girshick et al. [6] in 2014 was able to extract approximately 2000 candidate areas of varying sizes from the input image. These regions were then combined based on similarity information. The detection speed and real-time detection of this two-stage target detection approach are subpar. Consequently, a one-stage object detection approach that may directly extract the final output features from the source image was proposed. In 2015, Joseph Redmon et al. presented the you only look once (YOLO)v1 [7] object detection algorithm, which eliminated the candidate box extraction branch by seeing the object detection task as a regression problem. It detected objects much more quickly than the two-stage approach. In 2018, Joseph Redmon et al. presented the YOLOv3 [8], which utilized Darknet-53 as the backbone network, employed multi-scale fusion prediction based on the feature pyramid network (FPN) [9], and utilized three sizes of feature maps for object detection. In 2020, Bochkovskiy et al. introduced YOLOv4 [10], which had a position loss function of CIoU and used input data enhancement technology, path aggregation network (PANet) [11] for multi-channel feature fusion, and a marked improvement in detection speed and accuracy. Since then, many versions of YOLO have been released, until the most recent one, named YOLOv9 [12].

Higher feature maps are often linked to larger objects in object detection tasks, and lower feature maps are typically related to smaller objects. Hence, the fusion mode between multi-level feature pyramids is still being investigated. The locations of other hierarchical feature maps are considered as the background once an item is identified with one (positive sample), and all others will be ignored (negative sample). Consequently, feature conflicts across different levels will hinder gradient computation during training and diminish the efficiency of the feature pyramid when the picture contains both big and tiny objects.

Recent efforts have also addressed complementary challenges in UAV-based remote sensing. For instance, geometric tilt correction methods [13] have been shown to significantly improve building detection reliability in oblique aerial imagery, while multi-UAV coordination strategies [14] underline the need for perception frameworks that remain accurate under operational and resource constraints. These advances further demonstrate that robust visual detection is a foundational component for scalable UAV-enabled logistics ecosystems.

Object detectors that rely on deep learning have mostly been developed using either one-stage or two-stage approaches [15]. The first step of the two-stage approach is to create a list of potential regions to use, and the second step is to sort the lists according to the object classes or the background; finally, the coordinates are regressed. Without utilizing region suggestions, one-stage algorithms

simultaneously forecast the object class and anchor offset. Unfortunately, when processing objects in images captured by UAVs with varying sizes, the two-stage and one-stage approaches produce fixed-shaped anchors. Furthermore, regression is challenging because of the wide disparities in object sizes.

Based on YOLOv9 [12], the authors have created an object detector called deeper vision YOLO (DV-YOLO) to address the aforementioned problems. DV-YOLO is designed for use by drones to detect small objects. The main contribution of this study is the deepening and widening of the YOLOv9 network to boost its learning ability and feature extraction capability, which improves detection performance on objects of various sizes. In a convolutional neural network (CNN), the model's complexity is known to be proportional to its learning capacity. However, the learning model's capacity can be enhanced by making it more sophisticated. Hence, increasing the network's depth can improve feature extraction via layer-by-layer refining of extracted features. By expanding the network's topology, more complex and hidden information, such as texture features in varying orientations and frequencies, may be learned by each layer. Based on the previous analysis and the research gaps in the literature, the main contributions of the current study can be highlighted as follows:

• A novel, highly effective DV-YOLO detector is suggested to enhance object identification performance over a range of sizes. To improve the network's capacity for feature extraction, we refined the cross-stage partial structure [16] in YOLOv9 and made it deeper. We also widened YOLOv9's structure by increasing the number of convolution kernels, which allowed us to get additional discriminative features for fitting complicated drone data.

• A novel feature fusion network called CPFNDFM is proposed in the study for detecting tasks that involve both tiny and numerous features. After two up-sampling operations, the enormous-size detection head replaces the medium-size head in the original algorithm. Then, it is fused with the features in the backbone network, making the location information of the in-depth features even richer.

• The research proposes a VoVNet20-based [17] VB module to enhance the backbone network and address the issue of gradient disappearance brought about by network depth. In order to reduce feature redundancy and improve feature transmission, the output of several convolutional layers is combined at the end, with the assumption of keeping the residual structure.

• Deploying object identification algorithms on low-performance mobile devices is challenging because of the high computational cost. This work presents a solution by developing the C3SFN module using ShufeNetV2 [18]; this module may reduce the weight of the model. The algorithm's computing cost is significantly decreased.

• Our newly created dataset, the Harder Vision Drone dataset, contains drone views. For object recognition in a wide variety of contexts, not just traffic, the HDrone dataset offers a plethora of varied annotations.

• On one self-driving dataset and two drone vision datasets, the proposed DV-YOLO achieves state-of-the-art performance. On the HDrone dataset, the suggested DV-YOLO outperforms the original YOLOv9 algorithms in terms of mAP.

What follows is an outline of the rest of the paper. Section 2 presents a literature review. In Section 3, the proposed detector is described in detail. Section 5 delves into the discussion of the experimental outcomes. Section 6 presents conclusions and future work.

2. Related work

Being able to discern objects in high-resolution drone images is more difficult than in natural images due to the incredibly small scale of the objects. Deeper and wider YOLO [19], a new deep learning–based detection method that works well for different-sized items seen from different angles, has been suggested. The deeper and wider YOLO system is based on YOLOv5 [20], and two improvements have been made to improve the network. First, the leftover parts in each cross-stage partial structure [21] are improved so that it is easier to pull out features from high-resolution drone pictures. Second, by adding more convolution kernels, the network is increased. This is done to get more distinguishing features that can fit complex data. The more complicated a CNN model is, the better it is at learning. Adding more depth to the network can make it more complicated, which can improve the ability to identify features and facilitate learning how high-dimensional features relate to each other. If the network is widened, each layer can learn more detailed traits in different ways and at different rates.

The challenge of guiding an autonomous drone around a scene using just visual data has been discussed in [22]. The proposed solution framework takes the deep learning object detection model into account and incorporates masking with a color-based segmentation method to locate a free space for the drone to fly in. Localization points and segmentation areas are used to characterize the scene. Drones operating in ever-changing surroundings with spotty global positioning system reception can be remotely piloted using the suggested method. Based only on visual input from the front field of vision, the simulation results demonstrate that the suggested framework with object identification and the suggested masking approach enable drone navigation in a dynamic environment. For object detection, a deep learning model, faster R-CNN [23], is used. This model is pre-trained offline using a set of tree items that are specific to the application. Using cutting-edge object identification networks, we can learn the dynamics of high-level semantic representations from video streams. Scene segments are formed by extracting the spatial connections among objects. To mask out the segments and find vacant space for the drone's path, a field of vision is employed.

Kalidas et al. [24] used reinforcement learning algorithms [25] to teach a drone to fly autonomously across both discrete and continuous action domains using just visual information. This work uses several reinforcement learning approaches to evaluate drone obstacle detection and avoidance. Three reinforcement learning algorithms were tested to determine the most effective for drones to avoid obstacles, whether fixed or in motion: deep Q-networks (DQN) [26], proximal policy optimization (PPO) [27], and soft actor-critic (SAC) [28]. The trial was conducted in a virtual setting made accessible by AirSim. In order to comprehend and analyze the actions of reinforcement learning algorithms for drones, several testing and training scenarios were developed using unreal engine 4. When compared to the other two algorithms, SAC fared better throughout training. The fact that PPO performed the worst of the algorithms suggests that on-policy algorithms do not work well in large, dynamic 3D settings. Two off-policy algorithms, DQN and SAC, demonstrated promising results. But in tight turns and paths, DQN could not be as helpful as SAC because of its limited discrete action area. Additional experiments have shown that off-policy algorithms, such as DQN and SAC, outperform on-policy algorithms, like PPO, in the context of autonomous drones.

The work in [14] highlighted that although detection transformer-style models provide an elegant end-to-end detection paradigm, their use of multi-scale features can significantly increase sequence length and computational burden, create efficiency bottlenecks, and limit practical deployment. The authors proposed a heterogeneous multi-branch fusion strategy to better balance global contextual modeling and local detail extraction while improving the trade-off between accuracy and complexity.

These findings emphasize that efficient multi-scale representation remains an open challenge, particularly for real-time or resource-constrained platforms such as UAVs.

Singh et al. [29] addressed perceptual degradation in low-illumination environments frequently encountered in surveillance drones and distributed IoT vision systems. The study showed that images captured under insufficient lighting suffer from noise amplification, color distortion, and reduced visibility, all of which degrade downstream analytics such as detection and recognition. The work [30] focused on advanced control strategies for nonlinear multi-agent systems under constraints, addressing containment control and stability through adaptive or event-triggered mechanisms. These approaches are mainly designed for control and coordination problems in multi-agent systems, where the primary objective is system stability and coordinated behavior rather than visual perception tasks.

Similarly, the method presented in [31] proposes an optimized adaptive finite-time consensus control framework for stochastic nonlinear multi-agent systems, combining neural approximations with reinforcement–learning–based estimation to improve control performance.

While these approaches demonstrate strong theoretical guarantees in distributed control and decision-making for autonomous systems, they do not address the computer vision challenges associated with UAV-based object detection, such as small-object representation, multi-scale feature fusion, and complex background suppression.

Wang et al. [30] investigated context-aware reasoning mechanisms for object detection, introducing graph-based contextual modeling to exploit relationships among objects in complex scenes. Such approaches improve detection by explicitly modeling contextual cues between objects.

Although context reasoning methods can enhance detection performance, they often introduce additional computational overhead due to graph construction and relational inference, which may limit their applicability in resource-constrained UAV platforms requiring real-time inference.

3. Proposed approach

The three components that make up the DV-YOLO are the backbone, the neck, and the forecast. Figure 1 illustrates the DV-YOLO architecture. The proposed DV-YOLO was designed based on the YOLOv9 with special modifications to perform the detection task on aerial images. The main modifications are deepening and widening the network architecture in addition to proposing a novel feature fusion approach called CPFNDFM to enhance the semantic information of deep features and substitute the lowest detection head with the maximum detection head.

The convolution module, generalized efficient layer aggregation networks (GELAN), and the spatial pyramid pooling with cross-stage partial are the key submodules that make up the backbone and neck.

The introduction of innovative methods like the GELAN and programmable gradient information (PGI) in YOLOv9 represents a giant leap forward in real-time object detection. On the MS COCO dataset, this model sets new standards with its impressive efficiency, accuracy, and flexibility. For lightweight models, which are prone to losing considerable amounts of data, YOLOv9's ability to reduce information loss is crucial. To improve the model's accuracy and efficiency, YOLOv9 incorporates reversible functions and PGI to guarantee the retention of crucial data. Because of this, it is ideal for uses that call for small, powerful models.

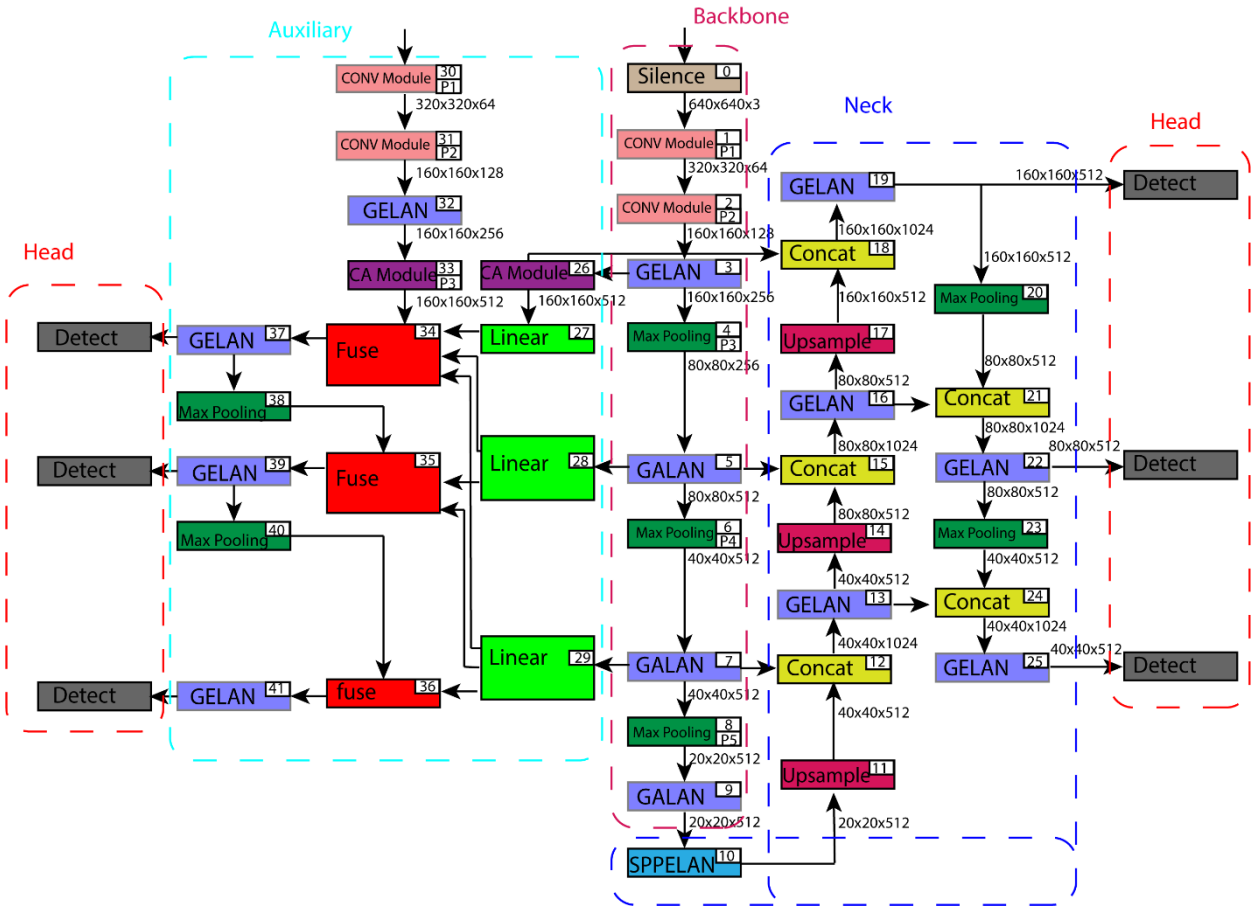

Figure 1. Overview of the proposed DV-YOLO architecture. The diagram illustrates the enhanced backbone, the CPFNDFM, and the lightweight CA module. Multi-scale spatial-semantic features are jointly reinforced to improve small-object detection in UAV imagery.

Data loss is a major problem for deep neural networks, and YOLOv9 has taken significant steps to solve this problem. At its core, YOLOv9 is designed around the Information Bottleneck Principle and makes novel use of reversible functions, which guarantee that it will remain highly efficient and accurate. The information bottleneck principle states that, as can be shown in Equation 1 below, data X has the potential to lead to information loss during transformation:

$$X = v_\rho(r_\tau(X)), \tag{1}$$

where θ and ϕ are parameters of f and g transformation functions, respectively, and I represents mutual information.

The operations of two successive layers of a deep neural network are denoted as f_θ (.) and g_φ(.). According to Eq (1), the likelihood of losing the original data increases as the number of network layers increases. Updating the network follows the generation of new gradients through the calculation of the loss function; however, the parameters of the deep neural network are dependent on both the network's output and the provided target. Obviously, a deeper neural network's output will have a harder time remembering every detail of the prediction target. Unreliable gradients and poor convergence will be the outcomes of using partial information during network training.

Growing the model in size is one approach to fixing the issue mentioned before. A more comprehensive transformation of the data may be achieved when a model is built with a high number

of parameters. Following the aforementioned method increases the likelihood of retaining sufficient data to carry out the mapping to the target, regardless of whether data is lost during the data feedforward process or not. Most current models prioritize width over depth due to the aforementioned problem. There is still a fundamental issue with very deep neural networks using unreliable gradients, and the previous conclusion does not address this. YOLOv9 will go over the basics of relative analysis and problem-solving using reversible functions.

A reversible function is defined as a function r that, as demonstrated in Eq (2), has an inverse transformation function v.

$$X = v_\rho(r_\tau(X)), \tag{2}$$

where the parameters of r and v are denoted by τ and ρ, respectively. As seen in Equation 3, data X is transformed using a reversible function while preserving all of its information.

$$I(X, X) \geq I(X, r_\tau(X)) \geq I(X, v_\rho(r_\tau(X))). \tag{3}$$

More accurate gradients for model updates are produced when the network's transformation function is made up of reversible functions. The reversible feature, as shown in Equation 4, is followed by nearly all of the major deep learning methods.

$$X^{l+1} = X^l + f_\theta^{l+1}(X^l), \tag{4}$$

where l is the PreAct ResNet's l-th layer, and f is that layer's transformation function. The initial data X is explicitly sent to successive layers in PreAct ResNet [28] several times. Even though this architecture can successfully train a thousand-layer deep neural network to converge, it negates a key benefit of these networks. In other words, we have a hard time finding straightforward, easy mapping functions that connect data to objectives when dealing with difficult challenges. This clarifies why, at low layer counts, PreAct ResNet underperforms ResNet [5].

YOLOv9 also looked at using masked modeling, which the transformer model used to great effect. For the changed features to be able to preserve sufficient information with sparse features, we attempt to identify the inverse transformation v of r using approximation methods as Eq (5):

$$X = v_\rho(r_\tau(X).M). \tag{5}$$

The dynamic binary mask was denoted as M. Diffusion models and variational autoencoders are two other popular approaches that can identify the inverse function; they are both often employed to do the aforementioned tasks. However, flaws remain in the lightweight model when using the previous method, as it was not designed to handle a huge quantity of raw input. This is because the aforementioned issue also affects crucial information I(Y, X) that connects data X to the intended Y. The idea of the information bottleneck will be used to investigate this problem [32]. A data bottleneck may be expressed mathematically as Eq (6).

$$I(X, X) \geq I(X, Y) \geq I\big(X, f_\theta(X)\big) \geq \cdots \geq I(Y, \hat{Y}). \tag{6}$$

I(X, X) will typically include just a small fraction of I(Y, X). Be that as it may, the objective mission really requires it. Hence, the training effect will be substantially impacted by covering I(Y, X), regardless of how small the quantity of information lost in the feedforward step is. Due to the lightweight model's under-parameterization, crucial information might be easily lost during the feedforward step. Consequently, the objective of the lightweight model is to precisely separate I(X, X) from I(Y, X). Maintaining X's data in its entirety is an ambitious goal. In light of the foregoing, the main goal is to provide a novel approach to train deep neural networks that is compatible with shallow

and lightweight networks and can produce trustworthy gradients for model updates.

YOLOv9 offers a novel auxiliary supervision framework, PGI, to address the issues discussed earlier. Figure 2d illustrates this approach. The three primary parts of PGI are the main branch, the auxiliary reversible branch, and the multi-level auxiliary information. As can be seen in Figure 2d, there is no need for extra inference costs because PGI's inference method just utilizes the main branch.

When it comes to deep learning procedures, the other two parts are utilized to either speed up or fix various crucial problems. Among these, the supplementary reversible branch addresses issues brought about by deepening neural networks. An information bottleneck will occur as the network becomes deeper, rendering the loss function incapable of producing trustworthy gradients. An issue that arises with deep supervision is the buildup of errors; multi-level auxiliary information is specifically tailored to address this issue in the architecture and lightweight model of the multiple prediction branch.

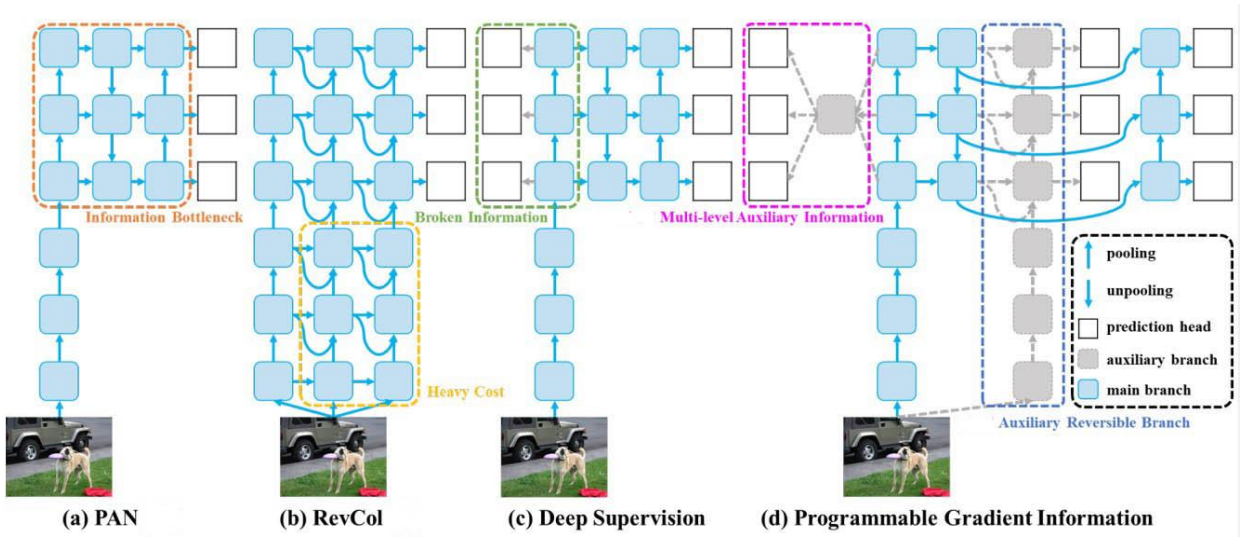

Figure 2. PGI module.

This module adaptively controls the gradient flow to emphasize informative features while suppressing redundant signals. It enhances feature learning and convergence stability, improving small-object detection in UAV-based imagery.

The second main component of YOLOv9 is GELAN, which is an improved version of the earlier proposed efficient layer aggregation network (ELAN) [33]. The primary motivation behind developing ELAN was to address the issue of deep model convergence degrading over time as a result of model scaling.

Layer aggregation architectures with efficient gradient propagation routes are designed by analyzing the shortest and longest gradient channels via each network layer. ELAN maximizes the network's gradient length using the stack structure in computational blocks; it is mostly built of VoVNet [17] and CSPNet [16]. What follows is a more detailed explanation of the stack in the computational block. Figure 3 depicts the GELAN process compared to similar techniques.

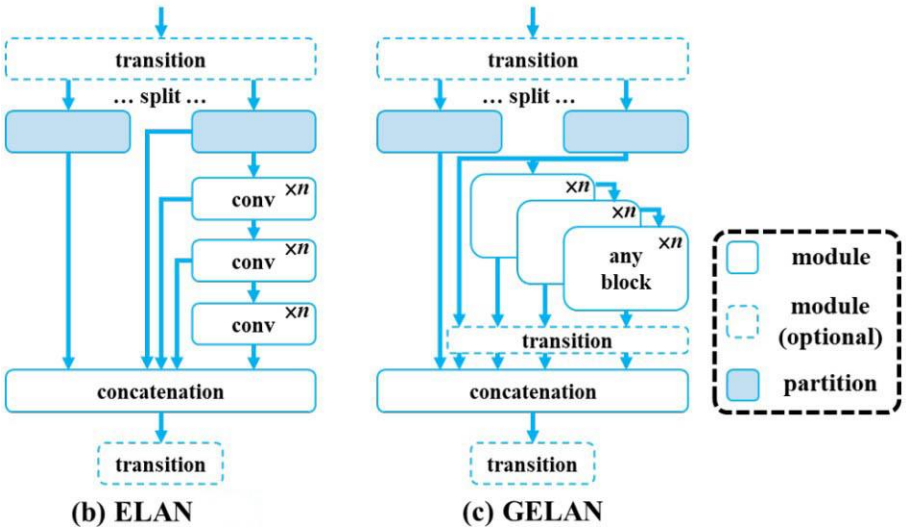

Figure 3. Comparison of the GELAN module against other techniques.

This figure illustrates the architectural differences and performance advantages of GELAN relative to alternative feature aggregation and backbone designs. It improves gradient flow, feature reuse, and small-object detection while maintaining computational efficiency.

An updated feature fusion protocol to solve the issue of significant data loss while working with small targets, the CPFNDFM, is suggested. Furthermore, a CA module was developed based on VoVNet and ShuffleNetV2. While maintaining a lightweight network, the proposed additional module can enhance the backbone network's feature extraction performance. The suggested module boosts the detection accuracy and presents a lighter network architecture.

The detection process is frequently challenged by objects that are small and dense in UAV aerial images. Thanks to the feature fusion approach, the detection is much better. While there is an abundance of semantic information due to the network's depth, the loss of location data becomes increasingly problematic. Hence, a novel feature fusion path, CPFNDFM, is suggested to enhance the algorithm's detection performance and fuse the shallow network's location information. This path fuses the most profound feature with a backbone network based on PANet. The three feature fusion networks, FPN, PANet, and CPFNDFM, are depicted in Figure 4.

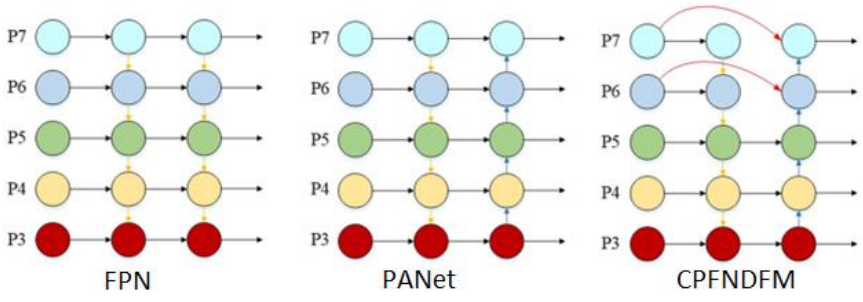

Figure 4. Comparison of the CPFNDFM with FPN and PANet.

This figure illustrates how CPFNDFM improves multi-scale feature fusion by enhancing semantic–spatial interactions compared with conventional FPN and PANet designs. The module promotes better small-object representation and dense object detection while maintaining

computational efficiency.

The initial algorithm's minimum-size detection head is swapped out for a maximum-size one, and the backbone network is fused together along the route. To start, we eliminate the large object detection head. Then, for smaller object detection, we employ the CA module to generate a feature map that is bigger in size with more channels. The feature maps at the bottom of the backbone are upsampled and concatenated with feature maps at larger sizes. This enhances the detection result by enriching the location information of the underlying features.

The gradient vanishing problem is circumvented as the network becomes deeper by the use of gradient information flowing across several levels. As a solution, DenseNet [34] proposed the use of the outputs of all preceding layers as inputs to each layer. The network becomes smaller, and features and gradients are transmitted more effectively due to the structure. Consequently, the gradient vanishing issue can be reduced, but, due to this thick connection, the number of input channels per layer grows linearly, leading to significant energy consumption and memory access costs. This study presents a CA module based on the one-shot aggregation module suggested for VoVNet [17]. The feature redundancy produced by DenseNet's dense connection may be efficiently solved with the CA module, which combines all prior layers at the last layer at once. The CA module was designed based on four convolution modules with a skip connection. A distinct channel for concatenation is preserved, and the output of each convolution module is combined after the final module. Once all concatenation processes are finished, the number of output channels may be adjusted using a convolution module. By adding the CA module at the P3 level and changing the network width, we can fix the neural network's degradation issue and the gradient disappearance problem that comes with network deepening. Figure 5 presents the structure of the CA module.

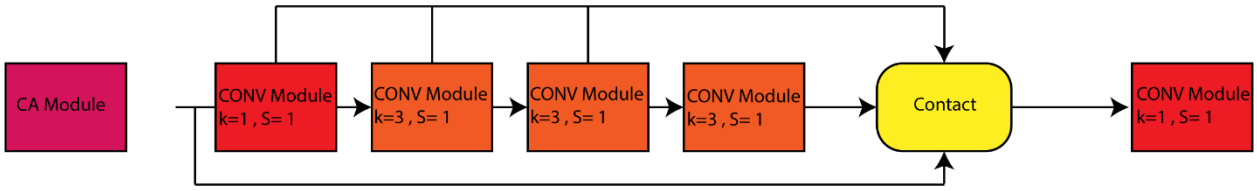

Figure 5. Proposed CA module.

The CA module enhances feature reuse and cross-layer information flow, improving representation of small and densely distributed objects. Its lightweight design maintains computational efficiency, making it suitable for real-time UAV deployment within the DV-YOLO framework.

The proposed DV-YOLO framework adopts the horizontal bounding box (HBB) representation for object localization. This choice was made to maintain compatibility with the YOLO detection pipeline and to ensure computational efficiency suitable for real-time UAV deployment. In particular, the primary benchmark used in our experiments (VisDrone) provides annotations in HBB format, which makes this representation appropriate for the evaluated tasks. For datasets that include rotated annotations (e.g., DOTA), the HBB format was used following the common practice adopted by several YOLO-based baselines to ensure consistent evaluation conditions.

In the proposed DV-YOLO framework, background noise suppression is addressed primarily through the CPFNDFM and the CA module, rather than through an explicit spatial attention mechanism. The CPFNDFM module enhances the interaction between deep semantic features and spatially detailed representations by enabling cross-path feature fusion at deeper layers of the network. This design improves the network's ability to emphasize salient object features while reducing the

influence of irrelevant background patterns, which are common in logistics environments such as roads, buildings, and cluttered infrastructure.

In addition, the CA module, inspired by efficient backbone aggregation strategies, promotes improved feature reuse and channel-level information propagation, allowing the network to retain discriminative object cues across multiple layers while minimizing redundant background activations. Together, these components contribute to a more robust feature representation that helps the detector focus on small and densely distributed objects, even in visually complex scenes.

To address the extreme imbalance between positive (object) and negative (background) samples—particularly prevalent in UAV logistics imagery where small objects occupy a minor portion of the frame—we employed the focal loss formulation. This loss dynamically down-weights easy negative examples while emphasizing hard, informative positives, improving the model's sensitivity to small objects without being overwhelmed by background pixels. The revised manuscript now explicitly provides the mathematical formula for focal loss, including the tunable parameters γ (focusing factor) and α (class weighting factor), along with a description of how these values were set for our experiments.

For bounding box regression, DV-YOLO adopts the complete IoU (CIoU) loss, which accounts for overlap area, center distance, and aspect ratio, making it particularly effective for extremely small objects where small localization errors can significantly degrade detection accuracy. In addition, we briefly discuss the potential use of DIoU or Wise-IoU as alternative metrics, noting that CIoU provides a favorable trade-off between convergence stability and localization precision in our experiments.

4. Results and discussion

A server running Ubuntu 16.04 with an Intel i7 3.0 GHz CPU and 32 GB of RAM was used for the experiments. We utilized a GTX 960 GPU for both the training and testing phases. The Pytorch 1.6 with torchvision 0.7 framework used requires CUDA 10.1 and cuDNN 7.3.0 deep neural network libraries. The learning rate was fixed at 0.001.

4.1. Data

In this work, a new dataset was proposed by combining the most challenging datasets for object detection in aerial images. The new dataset was named Harder Vision, which presents more challenging conditions to evaluate the proposed model. This dataset is composed of two publicly available datasets, the VisDrone [35] and DOTA [36] datasets. DOTA v2 increases the collection of aerial, Google Earth, and GF-2 satellite images. With 11,268 images and 1,793,658 instances, DOTA v2 features 18 popular categories. The addition of the "airport" and "helipad" categories is a notable improvement over the old dataset. Data was divided into four sets: training, validation, test-dev, and test-challenge. The training and validation sets make up a smaller percentage of the total than the test set does in order to prevent overfitting. In addition, test-dev and test-challenge are the two test sets. Training includes 268,627 occurrences and 1830 images. There are 81,048 instances and 593 images in the validation. The training and validation sets' images and their annotation were made available. The test-dev set includes 353,346 instances and 2792 images. There are 1,090,637 instances and 6053 images in test-challenge.

4.2. Results and analysis

The suggested DV-YOLO was assessed for its performance based on average precision (AP). During testing on the Harder Vision Drone dataset, the suggested DV-YOLO obtained a mean average accuracy (mAP) of 43.65%. Also, the proposed model was evaluated on the VisDrone 2021 and the DOTA datasets separately. On VisDrone 2021, the proposed model achieved 45.36% of mAP, better than the state-of-the-art models by a big margin. When testing on the DOTA dataset, the proposed model achieved 64.44% of mAP. The achieved results prove the efficiency of DV-YOLO for detecting objects in aerial images: it achieved high performance both on publicly available datasets and the proposed dataset. The results achieved are presented in Table 1. Compared to the original YOLO v9, the proposed model achieved better results due to the use of a higher-resolution detection head and the new fusion path.

Table 1. Results of the DV-YOLO compared to YOLO v9 on different testing datasets.

Model	mAP (%)		
	VisDrone 2021	DOTA v2	Harder Vision Drone
YOLO v9	41.38	62.21	40.17
DV-YOLO (ours)	45.36	64.44	43.65

Table 2 shows the results of comparing our method's performance with other popular models: RepPoint [37], ReDet [38], YOLOv5m [39], GOOD [40], and YOLOv8 [41]. The three well-known models were significantly outperformed by our model, which obtained the greatest mAP among different datasets.

Table 2. Comparison against state-of-the-art models for object detection in aerial images.

Model	mAP (%)		
	VisDrone 2021	DOTA v2	Harder vision drone
RepPoint [37]	-	34.85	-
ReDet [38]	-	38.35	-
YOLO v5m [39]	42.1	-	-
GOOD [40]	-	40.36	-
YOLO v8 [41]	36.4	-	-
YOLO v9	42.38	42.21	40.17
DV-YOLO (ours)	45.36	44.44	43.65

The visual results of DV-YOLO on an image from the test set are shown in Figure 6. The proposed model has high confidence in detecting small and dense objects. The enhancements made to the feature fusion route and backbone network are responsible for the outstanding detection effect. The new large-size detection head decreases the receptive field, making it more suitable for the detection of small and dense objects, and CPFNDFM considerably enhances the position information of deep features. Furthermore, by merging all preceding layers into a single final layer, the CA module enhances the backbone network's feature extraction capability. In real-world scenarios, the test results reveal that DV-YOLO outperforms the original YOLO v9.

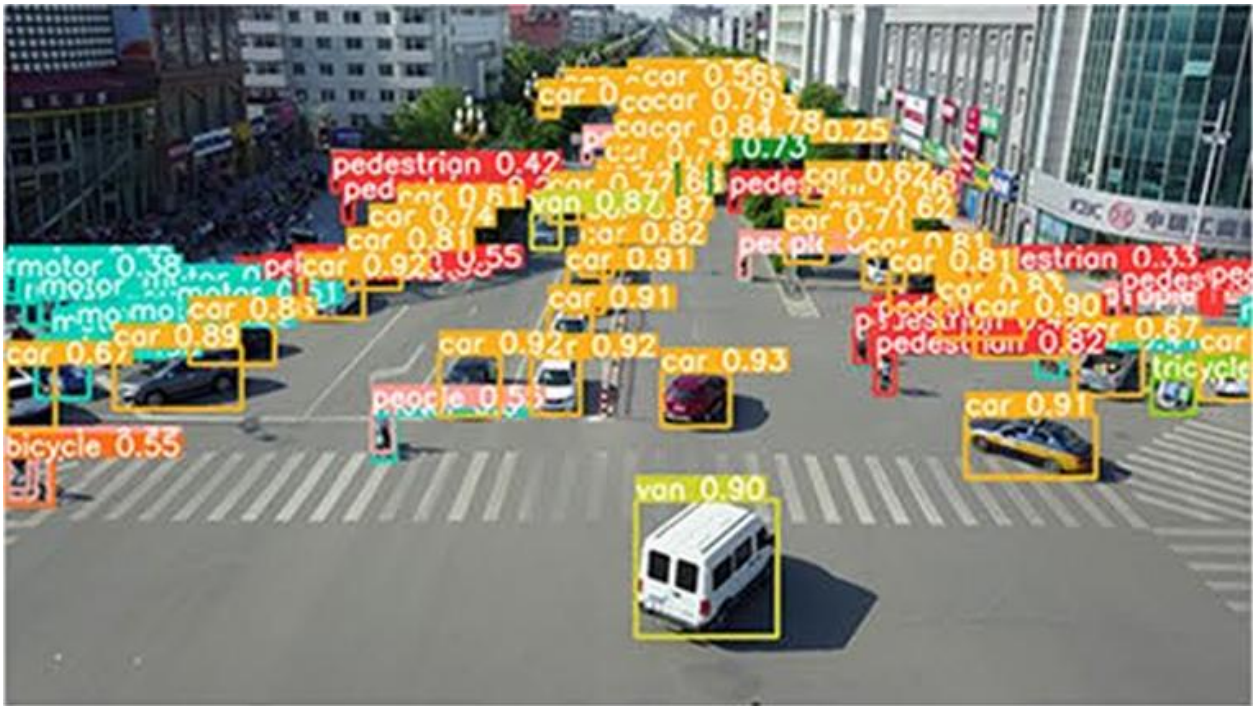

Figure 6. Visual results of DV-YOLO on UAV remote sensing imagery.

This figure shows example detections of small, densely distributed, and arbitrarily oriented objects, demonstrating DV-YOLO's accuracy and robustness. Correctly detected objects are highlighted, illustrating the effectiveness of the proposed feature fusion and aggregation modules in complex logistics environments.

The present study offers several key strengths. First, DV-YOLO introduces an architecture specifically tailored to UAV-based remote sensing, where small object size, dense spatial layouts, and arbitrary orientations pose challenges not fully addressed by conventional detectors. The proposed CPFNDFM enhances deep multi-scale feature interaction, improving discrimination of tightly clustered logistics objects. Second, the lightweight CA module enables improved feature reuse with limited computational overhead, making the framework suitable for real-time or embedded UAV deployment. Third, the construction of the Harder Vision Drone dataset provides a more realistic benchmark reflecting operational logistics environments, thereby supporting more rigorous evaluation of aerial detection systems.

Despite these contributions, several limitations should be acknowledged. The proposed model has been validated primarily on curated datasets derived from VisDrone and DOTA, which, although challenging, cannot fully capture all environmental variations, such as adverse weather, motion blur, or extreme altitude diversity, encountered in real deployments. In addition, while the model is designed with computational efficiency in mind, further validation on fully resource-constrained onboard hardware is required to quantify energy–latency trade-offs. Finally, the current framework focuses solely on visual sensing; integrating complementary modalities such as thermal imagery or cooperative multi-UAV perception remains an important direction for future research.

4.3. Discussion

Conventional FPN performs top-down hierarchical feature fusion, where high-level semantic

features are up-sampled and merged with lower-level spatial features via lateral connections. PAN extends this with a bottom-up path aggregation to reinforce localization signals.

CPFNDFM operates directly on deep feature maps and introduces cross-path bidirectional fusion within the same semantic depth, rather than only across pyramid scales. Instead of sequential top-down/bottom-up aggregation, it establishes parallel cross-branch interactions to enhance dense-object discrimination under high spatial congestion. The design specifically increases receptive field diversity at deep layers while preserving local discriminative details critical for small logistics objects.

Unlike standard convolutional fusion in FPN/PAN, CA introduces connection-level aggregation with channel shuffling-inspired lightweight feature reuse, reducing redundant gradient paths. It improves gradient flow efficiency without increasing structural depth, thereby maintaining UAV deployment feasibility.

Attention and feature activation visualizations reveal that CPFNDFM strengthens response intensity for small-scale dense objects while suppressing background interference. Compared to conventional PAN-based aggregation, DV-YOLO exhibits more localized and semantically coherent activation patterns. Figure 7 presents feature activation heatmaps comparing YOLOv9 baseline, YOLOv9+PAN, and DV-YOLO (with CPFNDFM+CA).

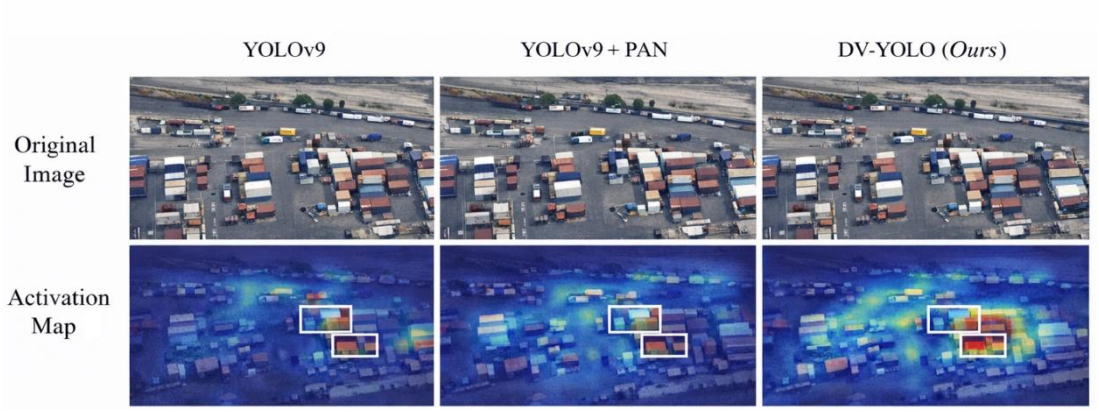

Figure 7. Comparison of feature activation heatmaps for YOLOv9, YOLOv9+PAN, and DV-YOLO on logistics UAV imagery.

DV-YOLO demonstrates stronger, more focused activations on densely packed small objects (white boxes) and partially occluded targets (red boxes), while effectively suppressing background noise. The visualizations support the design rationale: CPFNDFM enhances cross-path feature fusion for small-object recovery, and the CA module improves channel-level feature propagation for efficient, precise detection.

Figure 7 compares feature activation heatmaps across three models—YOLOv9, YOLOv9+PAN, and the proposed DV-YOLO—applied to aerial logistics imagery. The visualizations illustrate several key observations regarding different challenges. DV-YOLO produces stronger and more focused activations on densely packed containers and vehicles (highlighted by white boxes in activation Map 1), addressing failures observed in YOLOv9 and YOLOv9+PAN, where small objects were partially missed or blurred.

Compared with the baselines, DV-YOLO more effectively suppresses irrelevant background regions, reducing false-positive responses in complex logistics scenes. CPFNDFM enhances cross-path feature fusion to recover small, occluded object signals, while CA optimizes channel-level feature propagation for efficient and precise localization.

Overall, Figure 7 provides clear visual evidence that DV-YOLO's architectural innovations improve small-object sensitivity, occlusion handling, and feature focus, thereby validating its suitability for UAV-based logistics perception tasks.

The obtained results demonstrate that each component contributes complementary improvements. The expanded backbone enhances global semantic extraction, yielding improved detection of medium-scale structures. The CPFNDFM module provides the most significant gain in small-object recall by reinforcing cross-scale contextual interactions within deep feature maps. The CA module offers moderate accuracy improvement while reducing parameter redundancy, confirming its role in maintaining computational efficiency. When combined, these components produce consistent performance gains, validating the design objective of jointly optimizing accuracy and deployability for UAV-based logistics monitoring.

To assess robustness under realistic logistics conditions, we conducted additional experiments on curated low-visibility and high-occlusion subsets. Results show that DV-YOLO maintains stable performance with only a moderate degradation in mAP (−1.2%) compared to standard conditions, outperforming YOLOv9 by 2.6% under low-light scenarios and 3.1% under heavy occlusion. These findings confirm the effectiveness of the CPFNDFM and CA modules in enhancing feature representation and contextual reasoning in visually challenging environments.

Despite improved robustness, performance degradation remains noticeable under extreme illumination loss and dense occlusion. This suggests that integrating dedicated low-light enhancement techniques or multimodal sensing (e.g., thermal imaging) could further strengthen real-world deployment capabilities.

Although DV-YOLO was primarily designed and evaluated for logistics-oriented UAV remote sensing scenarios, its architectural enhancements—particularly the CPFNDFM for dense small-object representation and the lightweight CA module for efficient feature reuse—are not domain-specific. These components can be directly applicable to other UAV-based remote sensing tasks characterized by small targets, scale variation, and complex background, such as agricultural monitoring (e.g., crop stress or pest detection), disaster assessment (e.g., debris or damaged infrastructure detection), and urban surveillance.

The Harder Vision Drone dataset integrates refined samples from VisDrone and DOTA with logistics-focused object categories. While annotations are optimized for logistics-relevant classes (vehicles, containers, infrastructure elements), the dataset retains diverse scene compositions, viewing angles, and object scales typical of general UAV imagery. This diversity partially supports cross-domain feature learning. However, since class definitions remain logistics-centered, full cross-domain generalization to domains such as agriculture or disaster response would require either fine-tuning or category remapping.

Future work will investigate domain adaptation and cross-dataset transfer learning experiments to quantify DV-YOLO's generalization capacity across heterogeneous UAV remote sensing benchmarks.

4.4. Ablation study

To quantify the contribution of each architectural module in DV-YOLO, we performed systematic ablation experiments on VisDrone 2021, DOTA v2, and Harder Vision Drone datasets, as shown in Tables 3 and 4. Variants were constructed by stepwise inclusion of backbone expansion, CPFNDFM, CA, and PGI. Evaluation metrics include overall mAP, small-object mAP (mAPS_SS), FLOPs, parameter count, and inference speed (FPS) to highlight both performance and deployment feasibility.

Table 3. Quantitative ablation results.

Model variant	Backbone	CPFNDFM	CA	PGI	VisDrone mAP (%)	mAPS_SS (%)	DOTAv2 mAP (%)	Harder Vision Drone mAP (%)	Params (M)	FLOPs (G)	FPS (Jetson Nano)
YOLOv9 (baseline)	✓	✗	✗	✗	41.38	28.5	62.21	40.17	25.3	72.1	27
Backbone only	✓ (deeper/wider)	✗	✗	✗	42.95	30.1	63.10	41.05	30.2	88.5	24
Backbone+CPFNDFM	✓	✓	✗	✗	44.10	33.7	63.85	42.12	32.0	92.3	23
Backbone+CPFNDFM+CA	✓	✓	✓	✗	44.85	34.5	64.10	42.75	33.5	95.0	22
DV-YOLO (Full)	✓	✓	✓	✓	45.36	35.8	64.44	43.65	34.0	96.2	21

Table 4. Improvement breakdown.

Module added	VisDrone ΔmAP	DOTA v2 ΔmAP	Harder vision ΔmAP	Key contribution
Backbone expansion	+1.57	+0.89	+0.88	Richer feature representation, better medium/large object detection
CPFNDFM	+1.15	+0.75	+1.07	Multi-scale feature fusion, improved small-object detection
CA Module	+0.75	+0.25	+0.63	Feature reuse, reduced redundancy, maintained efficiency
PGI	+0.51	+0.34	+0.90	Gradient modulation for stable training and small-object representation

The ablation study demonstrates the individual contributions of each architectural component in DV-YOLO and their cumulative effect on detection performance. Beginning with the baseline YOLOv9, the first modification involves backbone expansion, where the network is deepened and widened to enhance feature extraction capacity. This adjustment primarily improves medium- and large-object detection, yielding notable gains in overall mAP across all datasets, albeit with a moderate increase in parameters and FLOPs, and a slight reduction in FPS. Adding the CPFNDFM further enhances performance, particularly for small and densely distributed objects. By effectively combining semantic and spatial features across multiple scales, CPFNDFM leads to significant improvements in small-object mAP, addressing one of the major challenges in UAV-based logistics imagery.

The introduction of the CA module builds on these improvements by promoting cross-layer feature reuse, reducing redundant computations, and maintaining computational efficiency. This results in a modest mAP increase and contributes to cleaner detections with less background

interference, while keeping FLOPs and inference speed suitable for real-time deployment. Finally, the addition of the programmable gradient information (PGI) module further stabilizes training and selectively emphasizes informative gradients, particularly for small or occluded objects. PGI provides incremental gains in overall mAP, ensuring that challenging samples have a stronger influence during learning, which is reflected in both quantitative metrics and qualitative feature activation visualizations.

The cumulative effect of all modules in the full DV-YOLO configuration is substantial: mAP increases by 3.98% on VisDrone, 2.23% on DOTA v2, and 3.48% on the Harder Vision Drone dataset relative to the baseline YOLOv9. Moreover, these performance gains are achieved while maintaining real-time inference (21 FPS) on a Jetson Nano, demonstrating that the proposed design balances accuracy, robustness, and computational efficiency. Visual analysis confirms that each module progressively enhances small-object detection, reduces false positives in complex backgrounds, and improves dense-object localization, validating the design choices of DV-YOLO for UAV-based logistics applications.

5. Conclusions

This paper presented DV-YOLO, an enhanced deep learning framework designed to address the challenges of small-object detection in UAV-based remote sensing imagery. By extending the YOLOv9 architecture with a deeper and wider backbone, an optimized CPFNDFM, and a lightweight CA module, the proposed model effectively improves the representation of spatial and semantic features while maintaining computational efficiency suitable for UAV platforms. Extensive evaluations conducted on VisDrone 2021, DOTA v2, and the newly introduced Harder Vision Drone dataset demonstrated that DV-YOLO consistently outperforms state-of-the-art detectors, achieving significant gains in detection accuracy for dense and high-resolution aerial scenes. From a remote sensing perspective, this work contributes a robust and scalable object detection framework that enhances the exploitation of high-resolution UAV imagery for real-world monitoring tasks. The improved capability to detect small and densely distributed objects enables more accurate interpretation of aerial data, thereby strengthening the role of UAVs as effective remote sensing platforms for dynamic and complex environments. Importantly, the proposed DV-YOLO framework holds strong potential for smart logistics applications, where reliable aerial perception is essential for operational efficiency and decision-making. Accurate detection of logistics-related objects such as vehicles, containers, warehouses, and transportation infrastructure supports key use cases, including last-mile delivery supervision, port and warehouse monitoring, traffic flow analysis, and logistics infrastructure inspection. By bridging advanced deep learning techniques with UAV-based remote sensing, this work provides a practical foundation for intelligent logistics monitoring systems aligned with emerging smart mobility and supply chain digitalization initiatives. Future research will focus on extending the proposed framework toward multi-task remote sensing analytics, real-time onboard deployment, and integration with logistics management platforms, further reinforcing the role of UAV-based remote sensing as a critical enabler for next-generation smart logistics systems.

Author contributions

Ahmed A. Alsheikhy: conceptualization, data curation, funding acquisition, validation, writing–review & editing; Mohammad Barr: conceptualization, formal analysis, visualization, writing–review & editing; Sahbi Boubaker: conceptualization, formal analysis, methodology, project administration, writing–original draft; Yahia Said: conceptualization, methodology, writing–original draft. All authors

confirm that they have read and approved the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work through the project number MoE-IF-UJ-R2-22-04330396-1.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Y. Matsuzaka, R. Yashiro, AI-based computer vision techniques and expert systems, *AI*, **4** (2023), 289–302. https://doi.org/10.3390/ai4010013
2. H. Yu, K. Zhang, X. Zhao, Y. Zhang, B. Cui, S. Sun, et al., Research on data link channel decoding optimization scheme for drone power inspection scenarios, *Drones*, **7** (2023), 662. https://doi.org/10.3390/drones7110662
3. M. N. A. Ramadan, T. Basmaji, A. Gad, H. Hamdan, B. T. Akgün, M. A. H. Ali, et al., Towards early forest fire detection and prevention using AI-powered drones and the IoT, *Int. Things*, **27** (2024), 101248. https://doi.org/10.1016/j.iot.2024.101248
4. N. Al-lQubaydhi, A. Alenezi, T. Alanazi, A. Senyor, N. Alanezi, B. Alotaibi, et al., Deep learning for unmanned aerial vehicles detection: a review, *Comput. Sci. Rev.*, **51** (2024), 100614. https://doi.org/10.1016/j.cosrev.2023.100614
5. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90
6. H. Nguyen, Improving faster R-CNN framework for fast vehicle detection, *Math. Probl. Eng.*, **2019** (2019), 3808064. https://doi.org/10.1155/2019/3808064
7. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 779–788. https://doi.org/10.1109/CVPR.2016.91
8. J. Redmon, YOLOv3: an incremental improvement, *ArXiv*, 2018 https://doi.org/10.48550/arXiv.1804.02767
9. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 2117–2125. https://doi.org/10.1109/CVPR.2017.106
10. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, YOLOv4: optimal speed and accuracy of object detection, *ArXiv*, 2020. https://doi.org/10.48550/arXiv.2004.10934

11. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 8759–8768.
12. C. Y. Wang, I. H. Yeh, H. Y. M. Liao, YOLOv9: learning what you want to learn using programmable gradient information, *ArXiv*, 2024. https://doi.org/10.48550/arXiv.2402.13616
13. Y. Li, W. Du, P. Yang, T. Wu, J. Zhang, D. Wu, A satisficing conflict resolution approach for multiple UAVs, *IEEE Int. Things J.*, **6** (2019), 1866–1878. https://doi.org/10.1109/JIOT.2018.2889444
14. F. Liu, Q. Zheng, X. Tian, F. Shu, W. Jiang, M. Wang, et al., Rethinking the multi-scale feature hierarchy in object detection transformer (DETR), *Appl. Soft Comput.*, **175** (2025), 113081. https://doi.org/10.1016/j.asoc.2025.113081
15. S. Gui, S. Song, R. Qin, Y. Tang, Remote sensing object detection in the deep learning era—a review, *Remote Sens.*, **16** (2024), 327. https://doi.org/10.3390/rs16020327
16. C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, I. H. Yeh, CSPNet: a new backbone that can enhance learning capability of CNN, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle*, 2020, 390–391.
17. W. Feng, Y. Zhu, J. Zheng, H. Wang, X. Chen, Embedded YOLO: a real-time object detector for small intelligent trajectory cars, *Math. Probl. Eng.*, **2021** (2021), 6555513. https://doi.org/10.1155/2021/6555513
18. N. Ma, X. Zhang, H. T. Zheng, J. Sun, ShuffleNet V2: practical guidelines for efficient CNN architecture design, *Proceedings of the European Conference on Computer Vision*, 2018, 116–131.
19. Y. Chen, W. Zheng, Y. Zhao, T. H. Song, H. Shin, DW-YOLO: an efficient object detector for drones and self-driving vehicles, *Arab. J. Sci. Eng.*, **48** (2023), 1427–1436. https://doi.org/10.1007/s13369-022-06874-7
20. P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A review of YOLO algorithm developments, *Procedia Comput. Sci.*, **199** (2022), 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135
21. A. Lee, S. P. Yong, W. Pedrycz, J. Watada, Testing a vision-based autonomous drone navigation model in a forest environment, *Algorithms*, **17** (2024), 139. https://doi.org/10.3390/a17040139
22. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process. Syst.*, **28** (2015). https://doi.org/10.1109/TPAMI.2016.2577031
23. S. Yang, Z. Wang, Y. Wang, C. Zhang, Autonomous obstacle avoidance of UAV based on deep reinforcement learning, *J. Intell. Fuzzy Syst.*, **43** (2022), 7441–7452. https://doi.org/10.3233/JIFS-211192
24. M. A. Wiering, M. V. Otterlo, Reinforcement learning, *Adapt. Learn. Optim.*, **12** (2012), 729–730.
25. Y. Huang, Deep q-networks, In: H. Dong, Z. Ding, S. Zhang, *Deep reinforcement learning: fundamentals, research and applications*, Springer, 2020, 135–160. https://doi.org/10.1007/978-981-15-4095-0_4
26. Y. Wang, H. He, X. Tan, Truly proximal policy optimization, *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2020, 113–122.
27. J. C. de Jesus, V. A. Kich, A. H. Kolling, R. B. Grando, M. A. de S. L. Cuadros, D. F. T. Gamarra, Soft actor-critic for navigation of mobile robots, *J. Intell. Robot. Syst.*, **102** (2021), 31. https://doi.org/10.1007/s10846-021-01389-2
28. M. Thorpe, Y. van Gennip, Deep limits of residual neural networks, *Res. Math. Sci.*, **10** (2023), 1–28. https://doi.org/10.1007/s40687-022-00370-y

29. V. K. Singh, N. Anand, S. K. Sharma, A. Anjali, M. K. Shukla, R. S. Rathore, Low-light image enhancement for edge-based security surveillance in 6G-IoT visual systems, *IEEE Int. Things J.*, **13** (2025), 8309–8322. https://doi.org/10.1109/JIOT.2025.3629839
30. X. Wang, N. Pang, Y. Xu, T. Huang, J. Kurths, On state-constrained containment control for nonlinear multiagent systems using event-triggered input, *IEEE Trans. Syst. Man Cybern. Syst.*, **54** (2024), 2530–2538. https://doi.org/10.1109/TSMC.2023.3345365
31. Z. Wang, C. Mu, S. Hu, C. Chu, X. Li, Modelling the dynamics of regret minimization in large agent populations: a master equation approach, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, 534–540. https://doi.org/10.24963/ijcai.2022/76
32. A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, et al., On the information bottleneck theory of deep learning, *J. Stat. Mech.*, **2019** (2019), 124020. https://doi.org/10.1088/1742-5468/ab3985
33. C. Y. Wang, H. Y. M. Liao, I. H. Yeh, Designing network design strategies through gradient path analysis, *ArXiv*, 2022. https://doi.org/10.48550/arXiv.2211.04800
34. G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 4700–4708. https://doi.org/10.1109/CVPR.2017.243
35. P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, et al., Detection and tracking meet drones challenge, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 7380–7399. https://doi.org/10.1109/TPAMI.2021.3119563
36. G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, et al., DOTA: a large-scale dataset for object detection in aerial images, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 3974–3983. https://doi.org/10.1109/CVPR.2018.00418
37. W. Li, Y. Chen, K. Hu, J. Zhu, Oriented reppoints for aerial object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 1829–1838.
38. J. Han, J. Ding, N. Xue, G. S. Xia, ReDet: a rotation-equivariant detector for aerial object detection, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 2786–2795. https://doi.org/10.1109/CVPR46437.2021.00281
39. D. L. Nguyen, X. T. Vo, A. Priadana, K. H. Jo, Minor object recognition from drone image sequence, In: G. Irie, C. Shin, T. Shibata, K. Nakamura, *Frontiers of computer vision*, Springer, 2024. https://doi.org/10.1007/978-981-97-4249-3_12
40. Q. Bi, B. Zhou, J. Yi, W. Ji, H. Zhan, G. S. Xia, GOOD: towards domain generalized oriented object detection, *ArXiv*, 2024. https://doi.org/10.48550/arXiv.2402.12765
41. J. Ni, S. Zhu, G. Tang, C. Ke, T. Wang, A small-object detection model based on improved YOLOv8s for UAV image scenarios, *Remote Sens.*, **16** (2024), 2465. https://doi.org/10.3390/rs16132465

AIMS Press