Earth Science Informatics (2024) 17:3253–3269
https://doi.org/10.1007/s12145-024-01335-1

RESEARCH

YOLO-U: multi-task model for vehicle detection and road segmentation in UAV aerial imagery

Zhihong Zhao[1,2] · **Peng He**[1]

Received: 9 February 2024 / Accepted: 15 May 2024 / Published online: 4 June 2024

Abstract

Due to the constrained performance of embedded chips in devices such as drones, real-time processing of simultaneous vehicle detection and road segmentation networks becomes challenging, leading to a lack of associative feature learning. To tackle these issues, we introduce a novel multi-task model for vehicle detection and road segmentation in unmanned aerial vehicle(UAV) Aerial Imagery. Our approach introduces a lightweight Ghost-Dilated convolution, combining the large receptive field of dilated convolution with the efficiency of Ghost convolution, resulting in fewer parameters and reduced computational load. Building upon this, we propose the Ghost-Atrous Spatial Pyramid Pooling (G-ASPP) module, a multi-scale feature extraction module that enhances the model's multi-scale characteristics while minimizing the increase in network parameters and computational requirements associated with Atrous Spatial Pyramid Pooling(ASPP) modules. The constructed multi-task UAV aerial vehicle detection and road segmentation network incorporates a carefully designed backbone, neck, detection head, and segmentation head. By refining existing lightweight backbone networks, our model achieves superior real-time performance and accuracy, demonstrating enhanced detection and segmentation accuracy with lower parameters and computational overhead. Experimental validation on a self-constructed multi-task dataset highlights the proposed model's improved segmentation and detection performance, particularly for small targets and narrow roads, confirming its effectiveness. This research contributes valuable insights to the study of multi-task networks in the realm of UAV vision.

Keywords UAV · Deep learning · YOLO · Multi-task network · Vehicle detection · Road segmentation

Introduction

Currently, Unmanned Aerial Vehicle (UAV) find applications in nearly a hundred fields of civil use, spanning agriculture, forestry, power, environmental protection, land, ocean, water conservancy, and transportation, among others (Chao et al. 2022). In the field of transportation, UAV, with their characteristics of compact size, high maneuverability, and flexible deployment, offer significant advantages in areas such as illegal evidence collection, traffic guidance, and routine inspections (Ling et al. 2022). The cameras mounted on UAV can transmit real-time footage, allowing operators on the ground to view the captured scenes. The flexibility of UAV enables them to adapt to various complex conditions, and aerial images provide more information compared to ground perspectives. Through artificial intelligence technology, valuable information can be extracted from aerial images, such as tracking the object detected in the images (Xue et al. 2023a, 2023b; Sun et al. 2024), so as to improve the work efficiency of data analysts.

In traffic patrols, it is essential for management personnel to pay more attention to vehicles on the road. Therefore, the issue of how to avoid detecting vehicles outside the designated road area becomes crucial. One solution to this problem is to extract road information and retain detection targets that intersect with the road. However, this approach relies on accurately and efficiently extracting road information. Object detection and image segmentation, as one of the application directions of artificial intelligence technology, leverage various manually

Communicated by: Hassan Babaie

✉ Peng He
hep0168@foxmail.com

[1] School of Information Science and Technology, Shijiazhuang TieDao University, Shijiazhuang, China

[2] State Key Laboratory of Structural Mechanics Behavior and System Safety of Traffic Engineering, Shijiazhuang, China

curated datasets to train neural networks to detect and segment various objects in images or videos, meeting the requirements for vehicle detection and road information extraction. Object detection methods can be classified into one-stage and two-stage methods. One-stage methods (such as YOLO (Redmon et al. 2016) and SSD (Liu et al. 2016)) directly predict the position and category of objects in a single network, suitable for applications with high real-time requirements. Two-stage methods (such as Faster R-CNN (Ren et al. 2015)) first extract candidate regions and then classify and locate these regions, providing higher accuracy. Image segmentation algorithms based on Convolutional Neural Networks (CNN) have also made significant progress. Classical methods like Fully Convolutional Network (FCN) (Long et al. 2015), U-Net (Ronneberger et al. 2015), SegNet (Badrinarayanan et al. 2017), and deep learning architectures like DeepLab (Chen et al. 1412), PSPNet (Zhao et al. 2017), continuously emerge, effectively improving segmentation performance by introducing techniques such as dilated convolutions and pyramid pooling.

Existing vehicle detection and road segmentation networks are mostly single-task networks. Drones, when identifying vehicles and segmenting roads, require the simultaneous operation of two networks, demanding high performance from the UAV and potentially wasting computational resources. Additionally, they may fail to extract correlated features between tasks.

Multi-task networks can save computational and storage resources by sharing network parameters. In resource-constrained environments or on mobile devices, such resource sharing is crucial for practical deployment. Moreover, multi-task networks allow neural networks to share learned representations across different tasks. By sharing the underlying feature extraction layer, the model can learn universal representations, enhancing the understanding of correlations between tasks.

The current multi-task algorithms lack optimization for the perspective of UAV, which leads to poor detection performance when directly applied in the field of UAV aerial photography, and it is difficult to meet real-time requirements. Due to the large scale differences and complexity of targets in the UAV perspective, it is necessary for multi-task networks to have the ability to extract multi-scale features. Furthermore, considering the power consumption of onboard chips on UAV, further lightweight improvements are needed for multi-task networks to reduce runtime power consumption.

In addressing the aforementioned issues, this paper makes the following contributions:

- In response to the high parameter and computational complexity issues of current multi-task models, as well as their unsuitability for small target detection from the perspective of UAV, a new multi-task framework YOLO-U has been constructed by improving lightweight backbone networks, employing highly coupled backbone and neck networks, and incorporating lightweight Ghost-Dilated convolutions and G-ASPP modules. This framework is more suitable for vehicle detection and road segmentation from the perspective of UAV compared to other multi-task models. The network employs an improved lightweight backbone, neck, detection head, and segmentation head, allowing simultaneous detection and segmentation of vehicles and roads in UAV-captured images. As both head networks share a common backbone and neck network, computational and storage resources are conserved, and the model's understanding of correlations between tasks is improved, enhancing detection performance.
- In response to the issue of increased parameter and computational complexity caused by adding multi-scale feature extraction modules, a lightweight dilated convolution called Ghost-Dilated convolution is proposed. Ghost-Dilated Convolution combines the characteristics of Ghost convolution and dilated convolution, using a two-stage feature extraction approach. It achieves a large receptive field while having fewer parameters and computational requirements.
- The ASPP module is composed of multiple dilated convolutions, which has a high number of parameters and computational complexity. In order to reduce the increase in model parameters and computational complexity, a lightweight Ghost-Atrous Spatial Pyramid Pooling (G-ASPP) module based on Ghost-Dilated convolution is proposed. The G-ASPP module has a structure similar to the Atrous Spatial Pyramid Pooling (ASPP) module but uses Ghost-Dilated Convolution instead of dilated convolution. Therefore, compared to the ASPP module, the network using the G-ASPP module has nearly equivalent multi-scale feature extraction capability but with lower parameters and computational requirements.

This article will be primarily divided into five chapters. In the first chapter, the research significance of multitasking networks on UAV is discussed, and the main contributions of this paper are introduced. The second chapter covers current related work, providing an overview of the current research status of UAV aerial target detection and segmentation, as well as multitasking networks. The third chapter details the methodology, providing a comprehensive introduction to the overall network framework and the principles behind the proposed modules. The fourth chapter focuses on the experimental section, conducting detailed experiments to validate the contributions made and prove their effectiveness. The fifth chapter serves as the conclusion, summarizing the contributions, experimental findings, and conclusions drawn in this paper.

Related work

UAV Aerial object detection

The task of object detection from the UAV perspective has received extensive research attention due to challenges such as varying target sizes, uneven target distribution, and complex shooting environments in UAV-captured scenes. Liu et al. (2023) proposed a novel dual backbone network detection method (DB-YOLOv5), which enhances the feature extraction capability for small targets by utilizing multiple backbone networks, achieving high detection performance for small targets. However, the use of multiple backbone networks significantly increases the model's parameters and computational requirements, difficult to deploy directly on UAV. Huang et al. (2023) proposed a lightweight object detection network for UAV platforms, based on the YOLOv5 network. By adding a small target detection head, improving the IOU metric, and introducing FasterNet, they enhanced small target detection while reducing model parameters and improving real-time performance. Khan et al. (2022) proposed a multi-scale and multi-class unified framework for detecting objects in high-resolution satellite images. The framework addresses the multi-scale problem by utilizing multiple Region Proposal Networks (RPNs), each with its own scale range, and leveraging the independent level of the pyramid to generate scale-specific object proposals. Li et al. (2023) proposed a lightweight infrared target detection method, named Edge-YOLO, by improving the YOLOv5m backbone network to further enhance the running speed of the network. It has a parameter size of 5.2 million and a computational workload of 14.2 gigaflops, achieving a running speed of 31.9 frames per second on the RK3588 chip.

UAV Aerial image segmentation

Image segmentation from the UAV perspective faces challenges such as complex scenes, difficult to achieve real-time performance and significant differences in target sizes. Li et al. (2023) proposed the Dual-Stream Feature Fusion Network (DSFA-Net), which utilizes two branches to extract shallow and deep information separately. This network balances shallow and deep feature extraction, improving feature fusion for stronger segmentation capabilities, especially for targets with large size differences. Xu et al. (2023) introduced an automated segmentation method for insulator images based on DeepLab V3 + . This method demonstrated effective segmentation of insulator images captured from UAV. Shi et al. (2023) presented a UAV image city scene segmentation network based on the Transformer. By designing a backbone network with a deformable multi-head self-attention transformer block featuring an aggregation window, introducing a position attention module, and a V-shaped encoder network, they improved the accuracy of city scene segmentation. The above-mentioned algorithm lacks optimization for embedded devices, which increases the performance requirements of the model and makes it difficult to achieve real-time requirements on embedded devices.

Multi-task neural networks

Multi-task networks are widely applied in the field of autonomous driving. Wu et al. (2022) proposed a panoramic driving perception network, YOLOP, capable of simultaneously performing vehicle detection, drivable area segmentation, and lane line segmentation. It achieves high detection accuracy while maintaining real-time performance. However, due to the abundance of small targets in the drone's field of view and the significant difference from the vehicle's perspective, there is a lack of effective solutions for handling scale variations in aerial drone footage, making it difficult to apply to target detection tasks from a drone's perspective. He et al. (2017) introduced a model, Mask R-CNN, capable of simultaneously performing instance segmentation and object detection. They added a branch for predicting object masks based on the existing bounding box recognition branch, laying the foundation for multi-task networks. However, due to the lack of corresponding lightweight design and the low real-time performance of the R-CNN framework, it is difficult to achieve real-time UAV target detection and segmentation tasks. Zhang et al. (2020) proposed a novel Multi-Scale and Occlusion Aware Network (MSOA-Net) for UAV-based vehicle segmentation. The issue of scale change is addressed through the use of a multi-scale feature adaptive fusion network. However, the network can only detect and segment vehicles, and cannot perform separate detection and segmentation of different targets. Additionally, due to the lack of lightweight processing in the backbone network, it is difficult to guarantee real-time performance on embedded devices. Balamuralidhar et al. (2021) proposed a multitask Mult EYE object detection, which utilizes the characteristics of multitasking during model training to simultaneously train the road segmentation head and vehicle detection head. During inference, the road segmentation head is frozen while sharing the underlying feature extraction layer to improve the accuracy of vehicle detection. However, in order to improve real-time performance of vehicle detection, the model chooses to freeze the road segmentation head, resulting in the inability to output road segmentation results simultaneously during inference and thus losing the characteristics of multitasking.

Methodology

Ghost-dilated convolution

In object detection and image segmentation tasks, objects to be detected and segmented in the input images vary in scale. Therefore, the network needs the capability to capture features at different scales. Researchers have used ordinary convolution with a large kernel to expand the network's receptive field, but this leads to an increase in network parameters. Addressing this issue, YU et al. (2015) proposed a convolution known as dilated convolution with a large receptive field, controlling the size of the dilated convolution's receptive field through dilation factors. Compared to ordinary convolution with the same receptive field, this approach significantly reduces the number of parameters and computational requirements.

HAN et al. (2020) analyzed the feature maps extracted by ordinary convolution and found that the feature maps of some channels extracted by ordinary convolution were similar to the feature maps of other channels, which indicated that the feature maps of other channels could be transformed into these feature maps by some linear transformation. Based on this analysis, they proposed a lightweight convolution called Ghost convolution. Ghost convolution adopts a two-stage feature extraction approach, as shown in Fig. 1. In the first stage, intrinsic feature maps of the images are extracted using ordinary convolution, with the channel number set to a smaller value. In the second stage, group convolution is employed to further process (linear transformation) the feature maps extracted in the first stage, and the results from both stages are concatenated for output. Through this process, Ghost convolution exhibits nearly the same feature extraction capability as ordinary convolution but with lower parameters and computational requirements.

Dilated convolution follows the same process as ordinary convolution in feature extraction. Therefore, similar optimization techniques used for ordinary convolution can be applied to dilated convolution to further reduce its parameters and computational requirements. Combining the two-stage characteristic of Ghost convolution with dilated convolution, a lightweight Ghost-Dilated Convolution with a large receptive field is proposed. The Ghost-Dilated Convolution is illustrated in Fig. 2.

In the first stage, intrinsic feature maps are obtained by applying dilated convolution with a smaller channel number to the input feature map, as defined in Eq. (1).

$$Y^{'} = \chi \times f^{'} \tag{1}$$

where $\chi \in \mathrm{R}^{\mathrm{H\times W\times C}}$ is the input feature map, $f^{'} \in \mathrm{R}^{\mathrm{c\times k\times k\times m\times d}}$ denotes the dilated convolution operation, and $Y^{'} \in R^{H_2\times W_2\times m}$ represents the intrinsic feature map obtained.

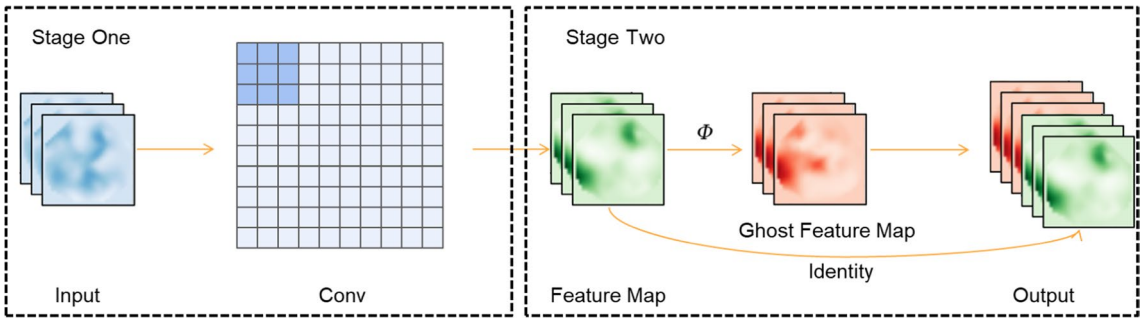

Fig. 1 Ghost convolution

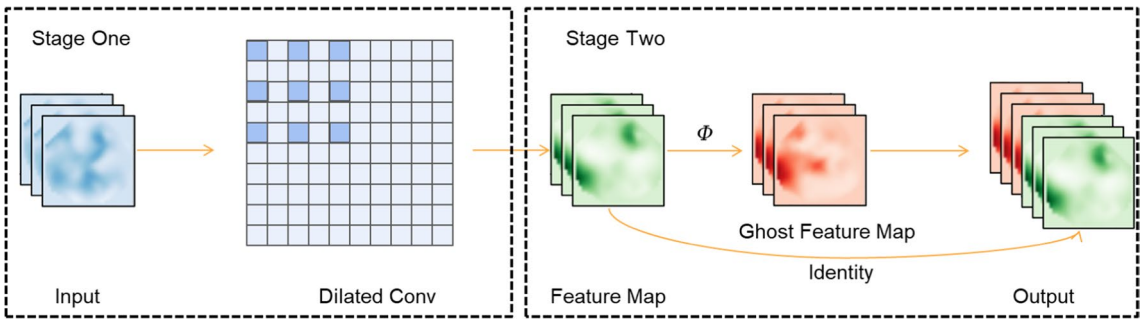

Fig. 2 Ghost-dilated convolution

In the second stage, linear transformations are applied to the intrinsic feature maps using 3 × 3 group convolution, and the computation process is expressed in Eq. (2).

$$y_{ij} = \Phi_{i,j}\left(y_i^{'}\right),\ \forall i = 1, ..., m,\ j = 1, ..., s \tag{2}$$

where $y_i^{'}$ represents the $Y^{'}$-th linear transformation process generating the i-th Ghost feature map y_{ij}, and $\Phi_{i,j}$ is the convolution kernel for the linear transformation.

Finally, the intrinsic feature maps and the linearly transformed feature maps are concatenated for the final output, as shown in Eq. (3).

$$Y = Cat\left(Y^{'},\ y_{i,j}\right),\ \forall i = 1, ..., m, j = 1, ..., s \tag{3}$$

where $Y \in R^{H_2 \times W_2 \times 2m}$ represents the final feature map generated by Ghost-Dilated Convolution.

Based on the Ghost-Dilated Convolution process described above, it combines the characteristics of dilated convolution with a large receptive field and the lightweight nature of Ghost convolution. This combination further reduces the parameters of networks employing dilated convolution.

G-ASPP module

Due to the relative flexibility of drones, the shooting perspective and altitude are not fixed. When a drone is at a high altitude, ground targets appear relatively small in the captured image, while they appear larger when the drone is at a lower altitude. Therefore, the network must have a multi-scale characteristic to accommodate the scale variations of targets captured by the drone. Basalamah et al. (2019) proposed a Scale-Driven Convolutional Neural Network (SD-CNN) model, which generates scale-aware object proposals by creating a scale map. This model effectively addresses the challenges of complex backgrounds, scale variations, nonuniform distributions, and occlusions in object detection tasks. He et al. (2015) proposed the use of Spatial Pyramid Pooling (SPP) modules in the network to enhance the capability of extracting scale features. YOLOv5 further improved the SPP module, introducing the SPPF module to enhance the network's recognition ability for multi-scale targets.

The ASPP module, based on the SPP module, uses dilated convolutions with parallel different dilation rates instead of max-pooling to extract features at different scales, and then combines these multi-scale features. Compared to SPP modules and SPPF modules that use simple max-pooling to increase the image's receptive field, the ASPP module enlarges the receptive field through dilated convolutions with different dilation rates. While using dilated convolutions to extract features from images can capture more multi-scale features than max-pooling operations, it also increases the network's parameters and computational requirements.

To avoid the increase in parameters resulting from adding dilated convolutions, a lightweight multi-scale feature extraction module called the G-ASPP module is proposed, based on Ghost-Dilated Convolution. The G-ASPP module replaces the original convolutions in the ASPP module with Ghost-Dilated Convolution, further reducing the module's parameters and computational requirements. The G-ASPP module is illustrated in Fig. 3. The G-ASPP module adopts a parallel structure, passing through Ghost-Dilated Convolution with dilation rates of 6, 12, and 18, and then concatenating the multi-scale features.

Compared to the ASPP module, the G-ASPP module demonstrates similar multi-scale feature extraction capabilities while further reducing parameters and computational requirements. Placing the G-ASPP module in the backbone effectively enhances the network's detection performance for various scale targets.

Overall structure of the multi-task network

When performing both vehicle detection and road segmentation tasks, a drone needs to run two networks simultaneously. Due to the lower performance of drone chips and the high cost of storage, running two neural networks concurrently is challenging and cannot guarantee real-time processing.

The emergence of multi-task networks effectively addresses the aforementioned issues. Currently, multi-task networks are predominantly designed with a parallel multi-task network structure, as illustrated in Fig. 4. The parallel multi-task network structure reduces redundancy by sharing convolutional layers. Moreover, the shared convolutional layers endow the network with the ability to extract features that are relevant to both tasks, thereby enhancing the detection and segmentation performance of the network.

Through the analysis of vehicle detection and road segmentation tasks, it is observed that these tasks exhibit a certain level of correlation. Vehicles are usually on the road, aligning with the characteristics of multi-task networks. Therefore, a multi-task aerial drone network, named YOLO-U, is proposed to perform vehicle detection and road segmentation tasks simultaneously. The overall network structure includes a backbone, neck, and head.

The multi-task network simultaneously accomplishes both vehicle detection and road segmentation tasks. Hence, the network is designed with separate head networks for vehicle detection and road segmentation, while sharing a common backbone and neck network. The YOLO-U network structure is illustrated in Fig. 5.

Constrained by the performance of drone devices, lightweight networks are chosen for the design of the multi-task network as the backbone network. Currently, there are various lightweight network models designed for mobile devices, such as ShuffleNet (2018), MobileNet (1704), GhostNet (1511), etc.

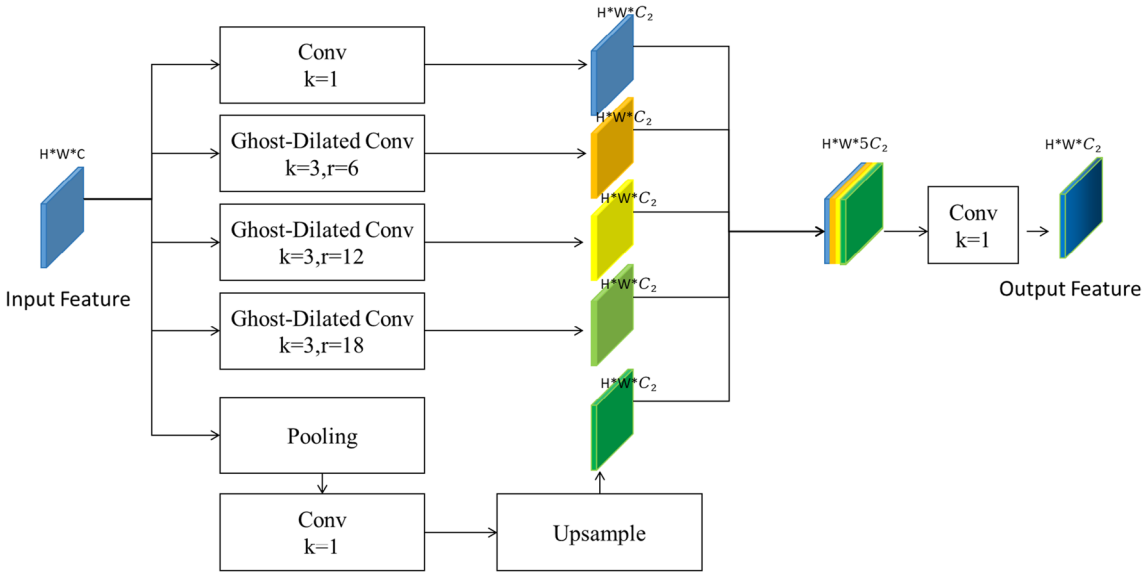

Fig. 3 G-ASPP module

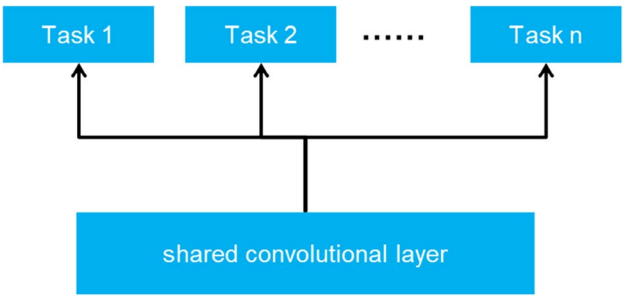

Fig. 4 Parallel multi-task network architecture

ShuffleNet introduces pointwise group convolution and channel shuffling mechanisms, reducing the parameters and computational requirements of convolutions. It also enhances the interaction between features, thereby improving the expressive capability of feature maps. MobileNet adopts depthwise separable convolutions instead of ordinary convolution, further reducing the parameters and computational requirements of convolution operations compared to pointwise group convolutions. The MobileNetV3 version introduces a lightweight Squeeze-and-Excitation (SE) attention mechanism, improving the focus on crucial features. GhostNet analyzes the feature maps generated by convolutions and proposes a Ghost convolution with lower parameters, avoiding redundant feature mappings and enhancing network efficiency.

Similar to MobileNetV3, GhostNet also integrates the SE attention mechanism. The SE attention mechanism primarily focuses on the relationships between channels, allowing the network to concentrate more on the feature channels that are crucial for the task. In comparison, the Efficient Channel Attention (ECA) mechanism replaces the fully connected

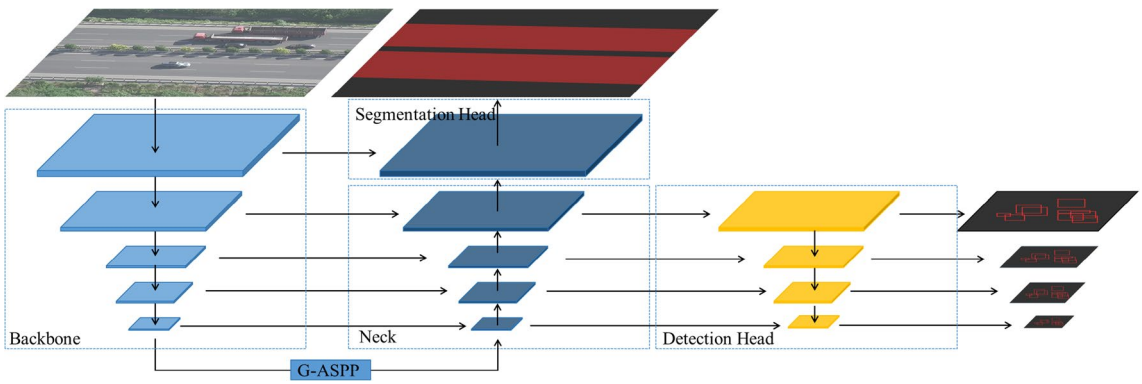

Fig. 5 YOLO-U network structure

layer in the SE attention mechanism with one-dimensional convolution. This not only reduces the computational and parameter requirements of the network but also enhances communication between channels.

ECA Attention Mechanism is illustrated in Fig. 6. Firstly, the input feature map undergoes global average pooling. Subsequently, a one-dimensional convolution operation with a kernel size of K is applied to the one-dimensional vector. The Sigmoid function is then employed to compute the convolution result, obtaining weights for each channel, as shown in Eq. (4). Finally, the original feature map is multiplied by the obtained weights, yielding a feature map that incorporates attention information.

$$\omega = \sigma(C1D_k(y)) \tag{4}$$

In the above formula, σ () denotes the Sigmoid function, $C1D_k$ represents the one-dimensional convolution computed through an adaptive convolution kernel, y signifies the channel after global average pooling, and ω represents the weights for each channel.

The SE attention mechanism in GhostNet was replaced with the ECA attention mechanism, further enhancing the efficiency of the backbone network. The structure of the backbone network is presented in Table 1. The input image first undergoes the Focus module, which divides the image into several smaller blocks at a certain ratio, enhancing the detection performance for small targets. Subsequently, GhostNet further extracts abstract features, utilizing convolution with a stride of 2 for downsampling the feature map, thereby reducing information loss caused by downsampling operations. Finally, the G-ASPP module is employed to extract and fuse multi-scale features, further strengthening the backbone network's capability to extract features from multi-scale targets.

The vehicle detection head network and neck network form a PAN structure. After feature extraction by the backbone network, a G-ASPP module is applied, and the neck network performs upsampling on the features, concatenating them with the features of the same scale from the backbone network. The vehicle detection head network undergoes multiple downsampling operations and concatenates with the neck network's features of the same scale. The network outputs four-scale feature vectors. The adoption of the PAN structure promotes the fusion and propagation of multi-scale features, improving the detection performance for multi-scale targets. Unlike YOLO's three detection heads, the network employs four detection heads, enhancing the detection performance for small targets.

The road segmentation head network, neck network, and backbone network together form a structure similar to the UNet network. The backbone network performs multiple downsampling operations on the image, and the neck network and road segmentation head network perform upsampling on the features extracted by the backbone network. The features of the same scale from the backbone network are concatenated, and the network outputs a segmentation result with a size of 640×640. The image segmentation network, adopting the UNet structure, fully utilizes shallow features, improving feature propagation and achieving better segmentation results compared to other image segmentation networks.

Both the neck network and the head network use the CSP module as the basic feature extraction module, as shown in Fig. 7. When the feature map passes through the CSP module, it goes through two branches before concatenation. The CSP module has a powerful feature extraction capability, lower computational and parameter overhead, and can save memory access. To avoid the increase in computational and parameter overhead caused by transpose convolution, linear interpolation is employed to perform upsampling on the feature map.

The network, overall, benefits from the high sharing of the backbone and neck networks between the two head networks, promoting network coupling. This enables the network to learn the correlation between tasks. The correlation between road segmentation and vehicle detection tasks mainly manifests when vehicles are on the road. The visualizations of vehicle detection labels and road segmentation labels are shown in Fig. 8. In the figure, the red area represents the road, and the green box represents the vehicles. It

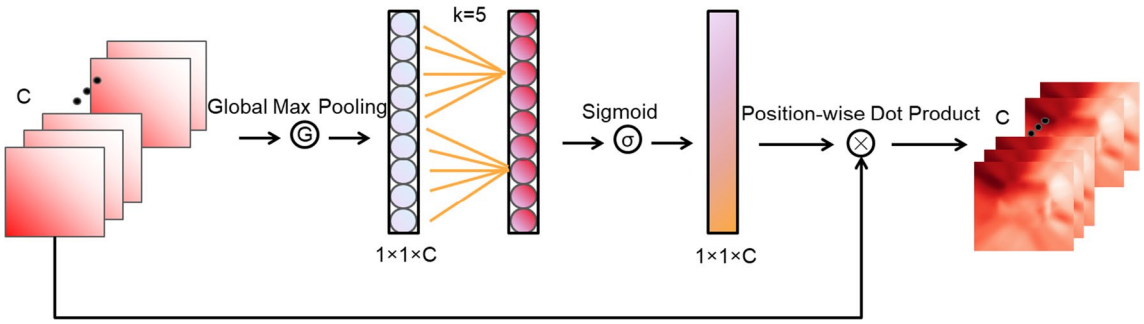

Fig. 6 ECA attention mechanism

can be observed from the figure that there is a high degree of overlap between the object detection labels and road segmentation labels, indicating an inherent correlation. When the network learns this correlation, it can avoid focusing on areas outside of the road, thereby reducing false detection rates for vehicles.

The model adopts GhostNet as the backbone network, which combines the ECA attention mechanism, making the model lightweight. The use of G-ASPP module for multi-scale feature extraction not only has lower performance overhead but also further enhances the model's ability to extract features at multiple scales. The detection head share the backbone network and neck network, further reducing performance overhead. Based on these characteristics, the time complexity of the network is lower compared to other multi-task networks, thus meeting the real-time requirements of UAV devices for algorithms and having a stronger learning ability for relevant features.

Table 1 Lightweight backbone network architecture

Number	Input Size	Module Name	Input Channels	Up Channels	Output Channels	ECA Module	Stride
1	640×640	Focus	3	/	16		/
2	320×320	GhostBottleNeck	16	16	16		1
3	320×320	GhostBottleNeck	16	48	24		2
4	160×160	GhostBottleNeck	24	72	24		1
5	160×160	GhostBottleNeck	24	72	40	√	2
6	80×80	GhostBottleNeck	40	120	40	√	1
7	80×80	GhostBottleNeck	40	240	80		2
8	40×40	GhostBottleNeck	80	200	80		1
9	40×40	GhostBottleNeck	80	184	80		1
10	40×40	GhostBottleNeck	80	184	80		1
11	40×40	GhostBottleNeck	80	480	112	√	1
12	40×40	GhostBottleNeck	112	672	112	√	1
13	40×40	GhostBottleNeck	112	672	160	√	2
14	20×20	GhostBottleNeck	160	960	160		1
15	20×20	GhostBottleNeck	160	960	160	√	1
16	20×20	GhostBottleNeck	160	960	160		1
17	20×20	GhostBottleNeck	160	960	160	√	1
18	20×20	G-ASPP	160	/	512		/

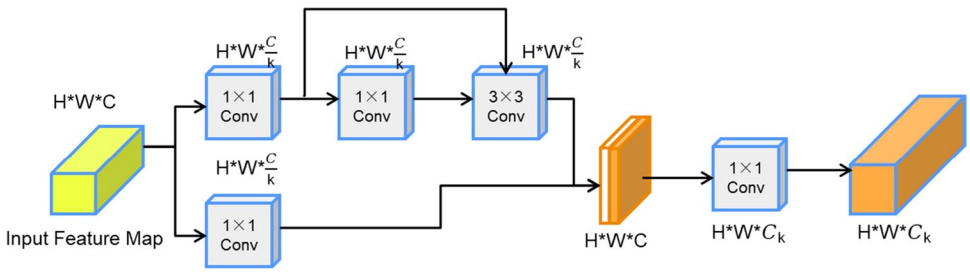

Fig. 7 CSP module structure

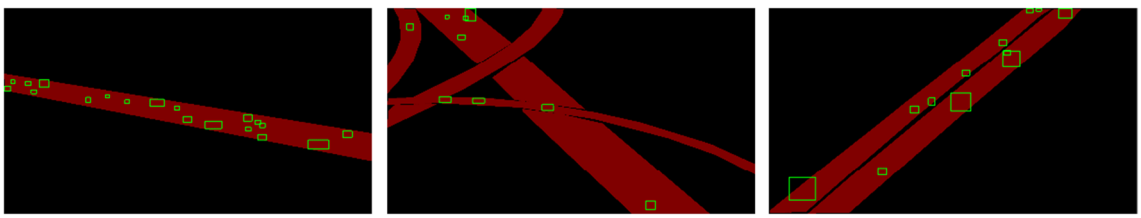

Fig. 8 Visualization of vehicle detection and road segmentation labels

Experiments

Experimental setup

Dataset

Currently, there is a relative lack of multi-task datasets from the perspective of UAV. To address this issue, we constructed a multi-task Dataset for UAV aerial object detection and road segmentation. Initially, UAV were deployed to capture ground-level footage, resulting in video data. Frames were extracted from the video data at fixed intervals of 10 s, yielding a total of 395 images containing vehicles and roads. The vehicles and roads were manually annotated using the LabelImg software, as illustrated in Fig. 9. The object detection task focuses on single-class object detection, and the distribution of anchor box sizes for object detection is depicted in Fig. 10, revealing a concentration of medium and small targets.

After augmenting the dataset through data augmentation techniques, we divided the dataset into training, validation, and test sets in an 8:1:1 ratio. Since multi-task dataset annotation is relatively difficult, it is relatively small. To enhance the model's generalization, pre-training was performed using publicly available datasets like Visdrone2019 (2019) and CHN6-CUG Road (2021a), which are closely related to the tasks.

Experimental environment and parameter settings

The experiments in this paper were conducted on a Dell Precision T7920 tower graphics workstation with an Intel Xeon Silver 4100@2.10 GHz×16 CPU. The GPU used was the Nvidia Quadro P5000 with 16 GB of memory. The system had 64 GB of RAM, and the storage configuration included a 512 GB SSD along with an 8 TB hard drive. The parameters related to model training are shown in Table 2.

Loss functions

Since the multi-task network has different output results, it requires joint loss functions corresponding to each task. The loss function for the vehicle detection task is defined as follows in Eqs. (5)-(8).

$$L_{conf} = \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} (C_i - C_i^{'})^2 \tag{5}$$

$$L_{cls} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \sum (p_i(c) - p_i^{'}(c))^2 \tag{6}$$

$$L_{iou} = IoU(B, B_{gt}) - \frac{\rho^2(B, B_{gt})}{c^2} - \alpha v \tag{7}$$

$$L_{det} = \alpha_1 L_{cls} + \alpha_2 L_{iou} + \alpha_3 L_{conf} \tag{8}$$

where L_{conf} represents the confidence loss, L_{cls} represents the classification loss, L_{iou} represents the IOU loss, and $\alpha_1, \alpha_2, \alpha_3$ are weight parameters.

The loss function for the road segmentation task uses the Cross-Entropy (CE) Loss Function, as shown in Eq. (9):

$$L_{seg} = -\sum_{i=1}^{N} [y_i (log\ y_i^{'}) + (1 - y_i) log(1 - y_i^{'})] \tag{9}$$

In the above equations, y_i represents the ground truth for the i-th sample, and $y_i^{'}$ represents the predicted value for the i-th sample.

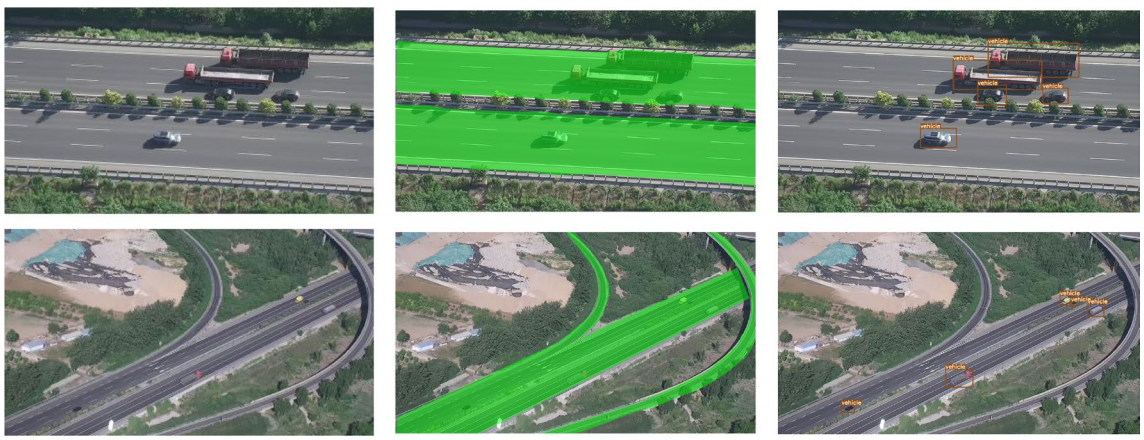

Fig. 9 Multi-task dataset for vehicle detection and road segmentation

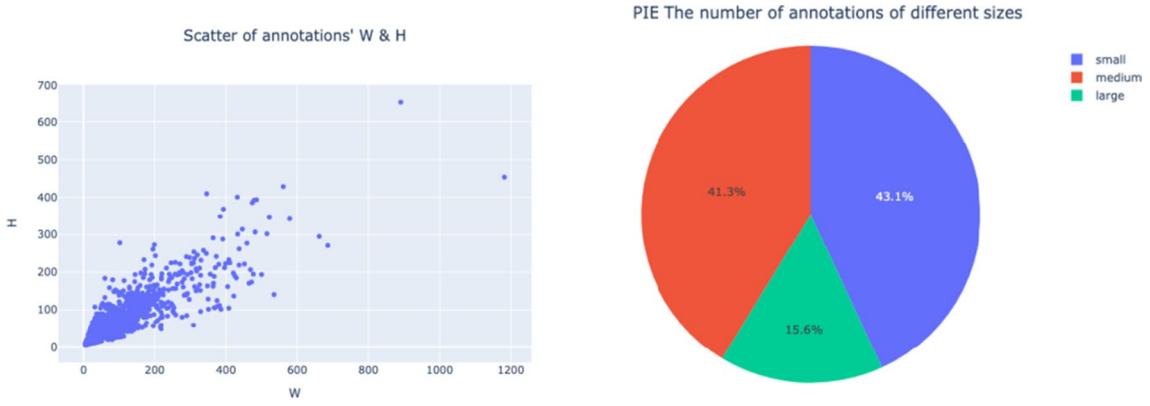

Fig. 10 Distribution of vehicle detection annotations sizes

Combining both tasks, the joint loss function is given by Eq. (10).

$$L_{all} = \beta_1 L_{det} + \beta_2 L_{seg} \tag{10}$$

In the above equation, L_{all} represents the joint training loss, and β_1, β_2 represents weight parameters.

Evaluation metrics

In this article, we evaluate the model using precision (P), recall (R), mean average precision (mAP), intersection over union (IOU), mean IOU (mIOU), as well as metrics related to model complexity such as parameter count and computational complexity. Specifically, P, R, and mAP are used as evaluation metrics for vehicle detection. The calculation formulas for P, R, and mAP are given by Eqs. (11), (12), and (13), respectively.

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

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

Table 2 Training-related parameters

Training-related parameters	Value
Optimizer	Adam
Learning rate	1e-2
epoch	480
BatchSize	3
ImgSize	640×640
Momentum	0.937
$\alpha_1, \alpha_2, \alpha_3$	[1,1,1]
β_1, β_2	[1,1]

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{13}$$

where TP represents the count of correctly detected bounding boxes. FP represents the count of bounding boxes mistakenly classified. FN represents the count of bounding boxes wrongly classified as background. AP_i denotes the model's average precision for the i-th class. In the context of a single-class detection task, mAP is numerically equivalent to AP.

Using IOU and mIOU as evaluation metrics for road segmentation, the calculation formulas for IOU and mIOU are as shown in Eqs. (14) and (15).

$$IOU = \frac{TP}{TP + FN + FP} \tag{14}$$

$$mIOU = \frac{1}{N+1} \sum_{i=1}^{N} \frac{TP}{FN + FP + TP} \tag{15}$$

where TP represents the number of correctly predicted road pixels, FP represents the number of background pixels incorrectly predicted as road, FN represents the number of road pixels incorrectly predicted as background, and N represents the total number of classes, specifically referring to the road and background classes in this context.

Model training

Due to the relatively small size of the annotated dataset, the training of the network suffers from the issue of weak generalization performance. To address this, pretraining is conducted on a dataset closely related to the task. Initially, the road segmentation head of the network is frozen, and a multitask network comprising backbone, neck, and detection networks is trained using the VisDrone2019 dataset,

with the epoch set 240. Subsequently, the backbone, neck, and detection networks are frozen, and the road segmentation head is trained using the CHN6-CUG Road dataset, with the epoch set 240. Finally, the entire network is trained using a multitask dataset for UAV target detection and road segmentation, with the epoch set 240. The hyperparameters for each training session remain unchanged. The pretraining algorithm is outlined in Algorithm 1.

Algorithm 1. Training of multi-task neural network

Algorithm 1 Training of Multi-Task Neural Network

Input:
  Multi-task neural network F parameters: $\Theta = \{\Theta_{backbone}, \Theta_{Neck}, \Theta_{Det}, \Theta_{Seg}\}$
  Training datasets: $T_{Vis}, T_{CHN}, T_{MultiTask}$
  Maximum training epochs: $epochs$
  Batch size for training: $BatchSize$
  Training loss function: L_{all}
Output:
  Trained network model $F(x;\Theta)$
procedure TRAIN(F, T)
  for $epoch \leftarrow 1, 2, 3, \ldots, epochs$ **do**
    $(x_s, y_s) \leftarrow Sample(T, BatchSize)$ ▷ Randomly sample batches from training set T
    $l \leftarrow L_{all}(F(x_s;\Theta), y_s)$ ▷ Compute loss
    $\Theta \leftarrow argmin_{\Theta} l$ ▷ Update Parameters
  end for
end procedure
$\Theta \leftarrow \Theta \setminus \{\Theta_{Seg}\}$ ▷ Freeze the road segmentation head network
TRAIN(F, T_{Vis})
$\Theta \leftarrow \Theta \cup \{\Theta_{Seg}\} \setminus \{\Theta_{backbone}, \Theta_{Neck}, \Theta_{Det}\}$ ▷ Unfreeze the road segmentation head network, freeze the backbone, neck, and detection networks
TRAIN(F, T_{CHN})
$\Theta \leftarrow \Theta \cup \{\Theta_{backbone}, \Theta_{Neck}, \Theta_{Det}\}$ ▷ Unfreeze the backbone, neck, and detection networks
TRAIN$(F, T_{MultiTask})$

Experimental results

Exploration experiment on task relevance

To further validate the correlation between road and vehicle positions, this section calculates the percentage of intersection between vehicle detection labels and road labels using algorithms. It also compares the percentage of intersection between detection boxes and road labels output by single-task network and multi-task network. The experimental results are shown in Table 3. The data in the table indicates that 98.5% of vehicle labels are located within the road in proportion to the ground truth. The percentage of detection boxes output by the multi-task network is 4.1% higher than that of the single-task network, indicating that the multi-task network has learned the correlation between road and vehicle positions, focusing more on road areas.

Table 3 Distribution of output labels in different networks

	Rate/%
True Label	98.5%
YOLO-UD	92.1%
YOLO-U	96.2%

In addition, it's noted that YOLO-UD refers to the model with the segmentation head removed, and YOLO-US refers to the model with the detection head removed.

In addition to the data comparison, visualizing the feature maps extracted from the backbone network, as shown in Fig. 11, reveals that the network focuses its attention primarily on the road area. This indicates that the network has

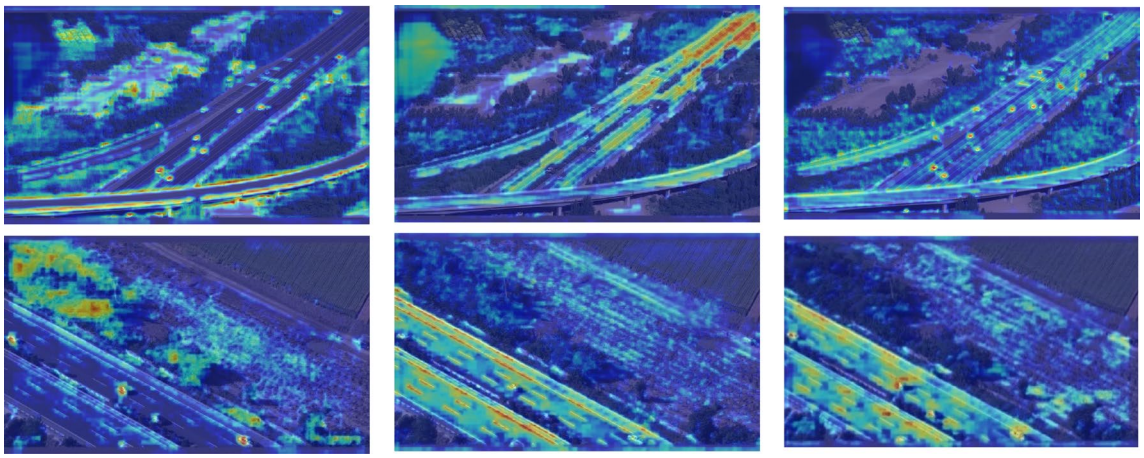

Fig. 11 Differences in attention distribution between multi-task and single-task networks

Table 4 Comparison results of lightweight backbone networks

	P(%)	R(%)	mAP(%)	IOU(%)	mIOU(%)	Parameter(M)	Flops(G)
MobileNetV3	79.9	75.4	73.9	88.7	92.2	8.49	12.33
ShuffleNetV2	79.4	75.7	72.6	86.1	90.4	14.01	14.65
FastNet-T0	72.2	83.6	79.0	89.4	92.7	13.07	15.49
GhostNet	79.0	80.0	77.3	87.8	91.6	10.80	11.77
GhostNet ECA	78.7	81.3	78.9	88.9	92.3	9.29	11.77

learned the correlation between tasks, thereby improving the accuracy of network detection and segmentation.

Comparison experiment of lightweight backbone networks

To further investigate the impact of different lightweight backbone networks on the multitask network, this section conducts experiments using popular lightweight networks as the backbone. The experimental results are summarized in Table 4.

MobileNetV3 has the fewest parameters compared to other networks, but its computational cost is higher than GhostNet. In terms of object detection accuracy comparison, the GhostNet ECA model demonstrates the best performance. It outperforms MobileNetV3, ShuffleNetV2 by 6.3%, 5.0% respectively, while achieving similar performance to FastNet-T0 but with lower computational and parameter requirements. Compared to GhostNet, it shows a 1.6% improvement. As for road segmentation accuracy comparison, the differences in performance among the networks are relatively small, with ShuffleNetV2 exhibiting the lowest performance.

The GhostNet network with the ECA attention mechanism significantly reduces the number of parameters, decreasing by 1.51 million. Considering the overall data, GhostNet ECA exhibits superior performance in detection accuracy, segmentation accuracy, and computational cost. This validates the effectiveness of the improvements made to GhostNet and underscores the efficiency of using GhostNet as the backbone network.

G-ASPP module ablation experiment

To validate the effectiveness of the proposed G-ASPP module and Ghost-Dilated convolution module, ablation experiments are designed. SPP module, ASPP module, and G-ASPP module are compared, and the experimental results are presented in Table 5.

Compared to the SPP module, the ASPP module improves the mAP metric by 4.2% and the mIOU metric by 0.5%. However, it comes with a substantial increase in parameters and computational cost, with a rise of 3.52 million parameters and 1.37 GFlops in computational cost.

By refining the ASPP module, the G-ASPP module slightly reduces the mAP and mIOU metrics compared to the ASPP module, showing a 1.2% decrease in mAP and a 0.1% decrease in mIOU. Nevertheless, there is a significant reduction in parameters and computational cost, with a decrease of 1.1 million parameters and 0.44 GFlops in computational cost.

Table 5 Comparison results of G-ASPP module

	P(%)	R(%)	mAP(%)	IOU(%)	mIOU(%)	Parameter(M)	Flops(G)
SPP	76.3	77.0	74.7	88.1	91.8	5.77	10.40
ASPP	78.7	81.3	78.9	88.9	92.3	9.29	11.77
G-ASPP	78.3	80.7	77.7	88.3	91.9	8.19	11.33

Table 6 Comparison results between YOLO-U and other networks

	P(%)	R(%)	mAP(%)	IOU(%)	mIOU(%)	Parameter(M)	Flops(G)
SSDLite	76.2	69.2	64.0	/	/	3.1	2.77
YOLOv5s	85.3	64.7	73.4	/	/	7.2	16.5
YOLOv8s	90.9	65.3	74.7	/	/	11.2	28.6
TPH -YOLOv5	74.2	80.3	77.1	/	/	45.3	260.8
YOLO-UD	77.1	79.3	75.9	/	/	5.73	8.78
UNet	/	/	/	80.4	81.2	31	167.65
MobileUNet	/	/	/	82.9	83.2	13.11	37.83
DeepLabV3-MobileNet	/	/	/	87.7	91.5	5.81	41.3
YOLO-US	/	/	/	90.6	93.5	3.13	5.16
YOLO-P	78.7	77.4	75.3	88.0	91.7	7.60	12.05
YOLO-U	78.3	80.7	77.7	88.3	91.9	8.19	11.33

The experimental data above demonstrates the effectiveness of the lightweight improvement made to the ASPP module through the G-ASPP module.

Comparison with other networks

To validate the effectiveness of the proposed method, comparisons are made with mainstream object detection algorithms,image segmentation algorithms and Multitasking Algorithms.The experimental results are presented in Table 6.

In the comparison of object detection models, the YOLO-UD model proposed in this paper outperforms YOLOv5s and YOLOv8s, showing a 2.5% and 1.2% improvement in mAP, respectively. Additionally, it reduces the number of parameters by 1.47 million and 5.47 million, and decreases computational cost by 7.72 GFlops and 19.82 GFlops. Compared with the TPH -YOLOv5 (Zhu et al. 2021b) model, which also performs well in drone aerial target detection, although there is a similar results in detection accuracy, the parameter and computational complexity of our model are much lower than those of TPH -YOLOv5. While the SSDLite (Sandler et al. 2018) with a lightweight backbone network has lower parameters and computational complexity, its detection accuracy is poor and cannot meet the requirements of the detection task.

In the comparison of image segmentation models, the YOLO-US model proposed in this paper significantly outperforms UNet and MobileUNet models in various metrics. It achieves a 12.3% and 10.3% improvement in mIOU compared to UNet and MobileUNet, respectively, while maintaining relatively smaller parameter and computational costs. Compared with DeepLabV3-MobileNet (Chen et al. 1706), the IOU and mIOU metrics have improved by 2.9% and 2.0% respectively, while the model's parameter count and computational workload have decreased by 2.68 M and 36.14Gflops respectively. It is worth noting that the multi-task model YOLO-U has shown a slight decrease in segmentation accuracy compared to YOLO-US, which may be attributed to the imbalance of multitasking caused by highly shared lower-level networks, leading the network to be more inclined towards vehicle detection tasks.

To further validate the superiority of the multitask model proposed in this paper under the perspective of UAV, a comparison was made with the YOLO-P (Wu et al. 2022) model, which has lane detection head removed. In terms of vehicle detection task, YOLO-U achieved a 2.5% higher mAP than YOLO-P. As for road segmentation task, both models performed similarly. In terms of parameters and computational cost related to real-time performance, although the parameter quantity of YOLO-U increased by 0.59 M, its computational cost decreased by 0.72 Gflops compared to YOLO-P. Overall, these data indicate that compared to the YOLO-P model, YOLO-U is more suitable for vehicle detection and road segmentation tasks from the perspective of UAV.

Visualization comparison experiment

To further validate the effectiveness of the proposed method, visual results are compared with mainstream object detection and image segmentation algorithms. The vehicle detection results are shown in Fig. 12.

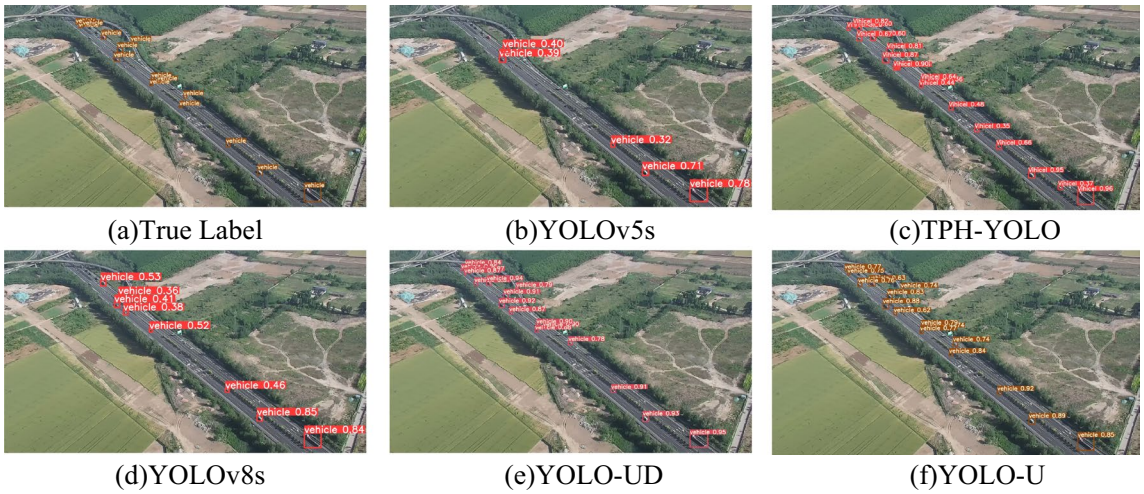

Fig. 12 Visualization results of vehicle detection models

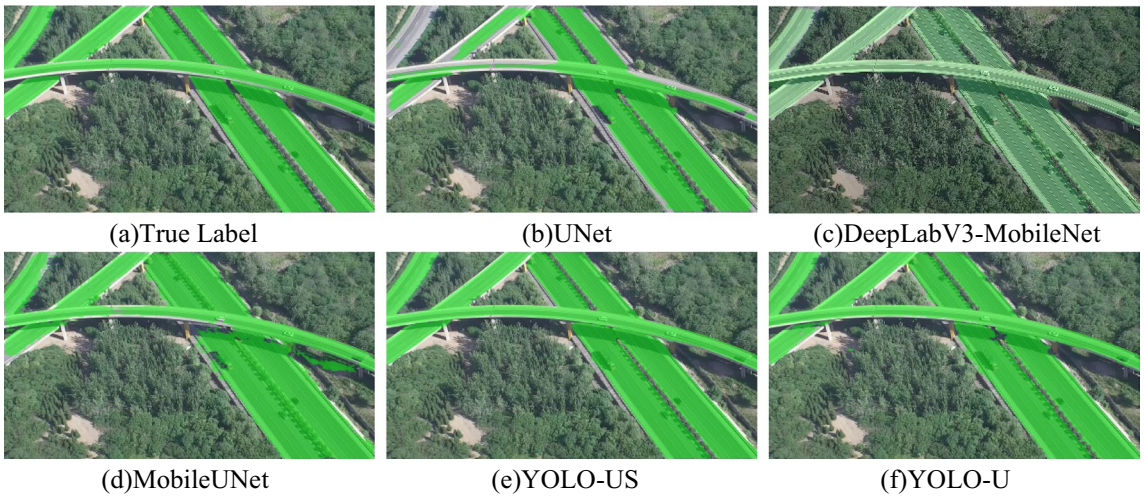

Fig. 13 Visualization comparison results of road segmentation models

In the task of vehicle detection, YOLOv5s exhibits the poorest detection performance, missing almost all small objects in the distance. YOLOv8s, while slightly better than YOLOv5s, still shows cases of missed detections with relatively low detection confidence. Thanks to the use of small object detection heads in the model, both YOLO-UD and YOLO-U models have good detection performance for distant small targets. However, YOLO-UD also has a few missed detections, while YOLO-TPH has a few false alarms. Overall, YOLO-U outperforms other models in terms of comprehensive performance.

Simultaneously, a visual comparison of road segmentation is conducted, as shown in Fig. 13. In the road segmentation task, UNet exhibits cases where some roads are not segmented, and the overall segmentation completeness is the poorest. Although MobileUNet segments all roads, it falsely detects green belts in the center of roads as part of the road. Both the DeepLabV3-MobileNet, YOLO-US, and YOLO-U models demonstrate similar overall performance. They accurately and completely segment the roads without encountering the issues observed in UNet and MobileUNet.

Visual comparison with the YOLO-P model, which is also a multitask model, as shown in Fig. 14. Due to the lack of optimization for small object detection and multi-scale feature extraction capabilities in the YOLO-P algorithm, there are missed detection issues when performing target detection from the perspective of unmanned aerial vehicles.

During the experiment, it was found that the model also has some deficiencies. As shown in Fig. 15, when there are a large number of vehicles in the area outside the road in the image, on one hand, it will interfere with the result of road segmentation

(a)True Label (b)YOLO-P (c)YOLO-U

Fig. 14 Visualization comparison results of road multitask models

Fig. 15 Failed case

(a)True Label (b)YOLO-U

Table 7 Comparative results of melting experiment

GhostNet	Small Object Head	ECA	G-ASPP	pretrain	P(%)	R(%)	mAP(%)	IOU(%)	mIOU(%)	Parameter(M)	Flops(G)
√					77.8	78.0	76.0	88.8	92.2	10.79	11.71
√	√				79.0	80.0	77.3	87.8	91.6	10.80	11.77
√	√	√			78.7	81.3	78.9	88.9	92.3	9.29	11.77
√	√	√	√		78.3	80.7	77.7	88.3	91.9	8.19	11.33
√	√	√	√	√	84.3	82.6	80.2	89.5	92.7	8.19	11.33

task, causing areas with more vehicles to be incorrectly segmented as roads. On the other hand, for vehicle detection outside the road, there are missed detections in the model.

Ablation experiment

To better understand the impact of each module on the network, ablation experiments were designed. As shown in Table 7, after adding the small object detection head, the network's mAP indicator increased by 1.3%. This indicates that the small object detection head further improves the network's ability to detect small object. The backbone network with ECA attention mechanism achieved improved detection and segmentation accuracy, with a 1.6% increase in mAP and a 0.7% increase in mIOU. Replacing the ASPP module with the G-ASPP module resulted in a slight decrease in detection and segmentation accuracy by 1.2% and 0.4%, respectively, but reduced the parameter count by 1.1 M and computational cost by 0.44GFlops. Furthermore, pre-training on multiple datasets further enhanced model generalization, leading to a 2.5% increase in mAP and a 0.8% increase in mIOU metrics. The results of these ablation experiments fully validate the effectiveness of each proposed module.

Conclusions

This paper addreses the performance limitations of current UAV running simultaneous object detection and road segmentation networks, unable to extract correlated features between tasks. The proposed multitask model for vehicle detection and road segmentation, named YOLO-U, leads to the following conclusions:

(1) The paper introduces a lightweight Ghost-Dilated convolution that combines the advantages of Ghost convolution and dilated convolution, maintaining a large receptive field with a lower parameter count. By addressing the parameter and computational cost

increase issues associated with the ASPP module, a lightweight multiscale feature extraction module, G-ASPP, is proposed, effectively reducing the model's parameter count and computational cost.

(2) GhostNet is chosen as the backbone network due to its effective feature extraction capabilities. An improved version, GhostNet ECA, is introduced by integrating the ECA module, resulting in a further reduction of parameters and increased detection accuracy. Leveraging these improvements, the YOLO-U model is proposed for multitask UAV aerial vehicle detection and road segmentation, sharing the backbone and neck networks between tasks to enhance feature correlation learning, leading to improved detection and segmentation results. Pretraining using self-built aerial vehicle detection and road segmentation datasets, combined with similar single-task datasets, further enhances model detection accuracy on the test set.

(3) Experimental results demonstrate that GhostNet ECA, as the backbone network, outperforms GhostNet by a 1.6% improvement in vehicle detection accuracy with a lower parameter count. The proposed G-ASPP module outperforms SPP and ASPP modules, improving detection accuracy while reducing parameter and computational costs by 1.1 million parameters and 0.44 GFlops, respectively. Comparisons with other single-task network models, both numerically and visually, show that the proposed YOLO-U model achieves superior accuracy and completeness in vehicle detection and road segmentation tasks. This validates the advantages of the proposed model.

Currently, our model also has some shortcomings, such as task imbalance and missed detection of targets outside the road. Additionally, there are issues with subsequent embedded porting that require further research in order to enhance the practical value of the model. This paper focuses on Deep Learning network models from the perspective of UAV and provides a direction for multitask networks carried by UAV.

Author contribution Zhihong ZHAO and Peng He participated in the design of this study, and they both performed the statistical analysis. HZ carried out the study and collected important background information. Peng He drafted the manuscript. All authors read and approved the final manuscript. Zhihong ZHAO and Peng He carried out the concepts, design, definition of intellectual content, literature search, data acquisition, data analysis and manuscript preparation. Zhihong ZHAO provided assistance for data acquisition, data analysis and statistical analysis. Peng HE carried out literature search, data acquisition and manuscript editing. Zhihong ZHAO performed manuscript review. All authors have read and approved the content of the manuscript.

Funding This research is supported by National Natural Science Foundation of China (12393780,11972236,12172234), and Shijiazhuang Tiedao University Graduate Innovation Funding Project (YC202451).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495

Balamuralidhar N, Tilon S, Nex F (2021) MultEYE: Monitoring system for real-time vehicle detection, tracking and speed estimation from UAV imagery on edge-computing platforms[J]. Remote Sensing 13(4):573

Basalamah S, Khan SD, Ullah H (2019) Scale driven convolutional neural network model for people counting and localization in crowd scenes[J]. IEEE Access 7:71576–71584

Chao Y, Lianghui T, Yuhao W et al (2022) Application of unmanned aerial vehicle in civil field in China[J]. Flight Dynamics 40(03):1–6+12. https://doi.org/10.13645/j.cnki.f.d.20220412.006

Chen LC, Papandreou G, Kokkinos I et al (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062

Chen LC, Papandreou G, Schroff F et al (2017) Rethinking atrous convolution for semantic image segmentation[J]. arxiv preprint arxiv:1706.05587

Dandan H, Han G, Zhi L et al (2023) Lightweight target detection network for UAV platforms[J]. Optics and Precision Engineering 31(20):3021–3033

Du D, Zhu P, Wen L et al (2019) VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 0–0

Han K, Wang Y, Tian Q et al (2020) Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 1580–1589

He K, Gkioxari G, Dollár P, et al (2017) Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2961–2969

He K, Zhang X, Ren S et al (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916

Howard AG, Zhu M, Chen B et al (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861

Khan SD, Alarabi L, Basalamah S (2022) A unified deep learning framework of multi-scale detectors for geo-spatial object detection in high-resolution satellite images[J]. Arab J Sci Eng 47(8):9489–9504

Li J, Ye J (2023) Edge-YOLO: Lightweight infrared object detection method deployed on edge devices[J]. Appl Sci 13(7):4402

Ling W, Peng Y, Jindong X et al (2022) Application of Unmanned Aerial Vehicle System in National Defense Traffic Field[J]. Journal of Military Transportation 1(12):37–41

Liu W, Anguelov D, Erhan D et al (2016) Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 21-37

Liu Y, Li W, Tan L et al (2023) DB-YOLOv5: A UAV Object Detection Model Based on Dual Backbone Network for Security Surveillance[J]. Electronics 12(15):3296

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 3431–3440

Redmon J, Divvala S, Girshick R et al (2016) You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779–788

Ren S, He K, Girshick R et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems 28

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 234-241

Runzeng Li, Zaifeng S, Fanning K et al (2023) Dual-Stream Feature Aggregation Network for Unmanned Aerial Vehicle Aerial Images Semantic Segmentation[J]. Laser & Optoelectronics Progress 60(24):291–299

Sandler M, Howard A, Zhu M, et al (2018) Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 4510–4520

Shi Y, Xi L, Junjie L et al (2023) UAVformer: A Composite Transformer Network for Urban Scene Segmentation of UAV Images[J]. Pattern Recognition, 133

Sun N, Zhao J, Shi Q, et al (2024) Moving Target Tracking by Unmanned Aerial Vehicle: A Survey and Taxonomy[J]. IEEE Transactions on Industrial Informatics, 7056–7068

Wu D, Liao M, Zhang W et al (2022) YOLOP: You Only Look Once for Panoptic Driving Perception[J]. Machine Intelligence Research 19(06):550–562

Xue Y, Jin G, Shen T et al (2023a) SmallTrack: Wavelet Pooling and Graph Enhanced Classification for UAV Small Object Tracking[J]. IEEE Trans Geosci Remote Sens 61:1–15

Xue Y, Jin G, Shen T et al (2023b) Template-guided frequency attention and adaptive cross-entropy loss for UAV visual tracking[J]. Chin J Aeronaut 36(9):299–312

Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122

Yue X, Xin D, Zivi L et al (2023) Automatic Segmentation Method for UAV Aerial Images of Insulators Based on DeepLab V3+[J]. Insulators and Surge Arresters 02:180–188. https://doi.org/10.16188/j.isa.1003-8337.2023.02.024

Zhang X, Zhou X, Lin M et al (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 6848–6856

Zhang W, Liu C, Chang F et al (2020) Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images[J]. Remote Sensing 12(11):1760

Zhao H, Shi J, Qi X et al (2017) Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2881–2890

Zhu X, Lyu S, Wang X et al (2021) TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2778–2788

Zhu Q, Zhang Y, Li Z, Yan X, Guan Q, Zhong Y, Zhang L, Li D (2021a) Oil Spill Contextual and Boundary-Supervised Detection Network Based on Marine SAR Images. IEEE Trans Geosci Remote Sens. https://doi.org/10.1109/TGRS.2021.3115492